Vers une utilisation synaptique de composants mémoires innovants pour l'électronique neuro-inspirée

Adrien F. Vincent

NNT : 2017SACLS034

Thèse de doctorat de l'Université Paris-Saclay préparée à l'Université Paris-Sud

École doctorale n° 575
Physique et ingénierie : électrons, photons, sciences du vivant
Spécialité de doctorat : électronique et optoélectronique, nano et microtechnologies

par

M. Adrien F. Vincent

Vers une utilisation synaptique de composants mémoires innovants pour l'électronique neuro-inspirée

Thèse présentée et soutenue à Orsay, le 3 février 2017.

Composition du Jury :

M.	Ian O'Connor	Professeur INL, Lyon	(Président du jury et rapporteur)
M.	Jean-Michel Portal	Professeur IM2NP, Marseille	(Rapporteur)
M.	Guillaume Prenat	Chercheur SPINTEC, Grenoble	(Examinateur)
M.	Sylvain Saïghi	Maître de conférences IMS, Bordeaux	(Examinateur)
Mme	Sylvie Galdin-Retailleau	Professeure C2N, Orsay	(Directrice de thèse)
M.	Damien Querlioz	Chargé de recherche C2N, Orsay	(Co-encadrant de thèse)

Thèse effectuée au sein du **Centre de Nanosciences et de Nanotechnologies**
de l'Université Paris-Sud
Bâtiment 220, rue André AMPÈRE
91405 ORSAY cedex
FRANCE

Remerciements

En premier lieu, je tiens à remercier chaleureusement MM. O'Connor et Portal pour leur étude attentive du présent manuscrit et le temps passé à rédiger leur rapport de pré-soutenance. Je suis également reconnaissant aux examinateurs du jury qui ont consacré du temps à la lecture de ce document en vue de la soutenance. J'ai une pensée particulière pour Sylvain, qui m'a apporté conseils et encouragements depuis la première fois que je l'ai rencontré, lorsque j'étais alors étudiant de master.

À Sylvie qui a accepté d'être ma directrice de thèse, malgré un emploi du temps qui aurait pu l'en dissuader dès le début de l'aventure et qui n'est pas allé en s'allégeant : un grand merci pour le temps passé à suivre ce projet, tout comme pour l'accompagnement dans les enseignements !

Je ne saurais être suffisamment reconnaissant à Damien pour le temps et l'énergie qu'il m'a consacré au cours de ces années. J'ai été très fortuné d'avoir eu un encadrant aussi disponible et à l'écoute, qu'il se soit agi de recherche ou non. Ces travaux de thèse doivent également beaucoup au groupe de recherche, fort de personnalités riches et éclectiques, que Damien a su fédérer autour de lui. Merci tout d'abord à Alice, pour avoir permis à Damien de faire l'expérience fructueuse de diriger simultanément deux doctorants aux méthodes de travail et aux personnalités très différentes, mais aussi pour avoir su animer cette équipe, à la fois en dedans et en dehors du laboratoire. Merci à Damir pour ces raisonnements toujours hors du cadre et pour ne jamais s'être énervé lorsque je faisais preuve de mauvaise foi envers le C++… Merci à Nicolas, pour son altruisme et ses conseils avisés, ainsi que pour les rappels mutuels de nos années à Cachan. Merci à Christopher, pour avoir constitué une source de motivation supplémentaire pour essayer de souffrir moins longtemps sur 42,195 km. Merci à Joseph, qui a notamment pris à cœur de nous faire découvrir certains us et coutumes d'outre-Atlantique. Merci également à Laurie pour l'intérêt qu'elle a porté à mes travaux (et pour m'avoir aider à me sentir moins seul à coder en Python). Merci à Jacques-Olivier pour le temps consacré à cette équipe Nanoarchi en dehors de ses enseignements à l'IUT et à Weisheng pour l'infrastructure de recherche qu'il a mise en place dans l'équipe. Enfin, je n'oublie pas Qifan, dont le passage en stage nous a permis d'avancer Nicolas et moi sur certains défis posés par nos recherches.

Ayant eu la chance d'être à cheval sur deux équipes de recherche, je tiens également à remercier les membres de l'équipe COMICS. Ainsi, merci à Arnaud, dont les cours sont une des raisons de mon arrivée au laboratoire en tant que doctorant. À Philippe, ne serait-ce que pour avoir si bien formé Damien (j'espère découvrir le caractère transitif de la chose). À (grand) Jérôme, directeur efficace de cette équipe. Et merci à Christophe pour le soutien logistique et informatique, élément crucial pour une équipe de simulation. Je remercie aussi les doctorants et post-doctorants avec lesquels j'ai pu interagir au cours des années, qu'il s'agisse des anciens comme Salim, (petit) Jérôme, Truong, Su Li ou encore Mai Chung, ou bien de ceux qui sont encore au milieu du gué, tels Ivan, Brice ou Jean.

Parce qu'il serait réducteur d'attribuer mon épanouissement au sein du laboratoire unique-

ment aux membre des équipes de recherche auxquelles j'appartiens, je remercie vivement Liza, Aloyse, Hugues, Valérie, Éric ou encore Tom, que ce soit pour m'avoir aidé à entretenir une certaine dépendance à la caféine ou [I] simplement pour avoir pu discuté de tout et de rien. Je remercie aussi les membres, anciens comme actuels, de l'équipe NOMADE avec qui j'ai pu interagir : Joo-Von, Jean-Paul, Thibaut, Sylvain, Dafiné, et leurs stagiaires, doctorants et post-doctorants.

J'avais été prévenu avant d'arriver en thèse que rien ne se déroule jamais entièrement comme prévu et je peux désormais le confirmer. J'ai grandement apprécié travailler avec Selina et Fabien, sur ce qui semblait au départ un petit projet annexe de quelques semaines et s'est finalement transformé en quelque chose de bien plus conséquent. Je tiens donc à les remercier pour leurs qualités humaines et scientifiques, qui m'ont offert l'opportunité de sortir de ma zone de confort et de considérer mes travaux de thèse sous un angle différent. Et pour avoir pu constater qu'effectivement, dans les Hauts-de-France, les gens ont dans le cœur le soleil qu'ils n'ont pas dehors.

J'adresse également de vifs remerciements à Julie, dont les étudiants, doctorants et post-doctorants ont beaucoup de chance de l'avoir, et avec qui j'ai eu du plaisir à interagir au cours des nombreuses occasions qu'ont été les groupes de lecture, les réunions du projet BAMBI ou encore les événements organisés ou soutenus par le groupement de recherche BioComp. Merci également l'influence positive indirecte sur la vie du bureau 113, en ayant participé à l'épanouissement d'Alice dans le cadre d'un co-encadrement très fructueux avec Damien. Enfin, que cela ait été fait de manière volontaire ou non, merci pour avoir entraîné Damien à viser grand en ce qui concerne les demandes de financement.

À Sandip, qui est venu plusieurs fois nous rendre visite depuis l'autre côté de l'Atlantique, je souhaite exprimer ma gratitude pour les nombreuses discussions que nous avons eues, m'amenant le plus souvent à douter, à mon grand désarroi mais aussi pour mon plus grand bien, de choses que je considérais jusqu'alors évidentes. Par ailleurs, je lui suis reconnaissant de ne s'être jamais plaint des conditions d'accueil, quand bien même celles-ci ont parfois été spartiates suite aux caprices de l'Yvette…

Merci à Jean-François, avec qui j'ai eu la chance d'être représentant des doctorants et post-doctorants du laboratoire. Malgré sa propre thèse à mener et ses (trop) nombreuses autres fonctions administratives, il a su trouver le temps d'être un élément moteur de la vie du laboratoire, participant à l'organisation d'événements festifs comme scientifiques, fortement appréciés.

Je les ai rencontrés il y a maintenant plus de sept ans, lors de mon arrivée Cachan et mon parcours aurait certainement été très différent s'ils n'avaient pas été là durant mon passage à l'École normale supérieure. Pour m'avoir supporté toutes ses années, je dois déjà un grand merci à Charles et à Rémy. J'ai été très heureux de les avoir pour colocataires durant la thèse et les remercie vivement d'avoir franchis le pas de cette aventure humaine. Grâce à eux, j'ai par ailleurs eu l'occasion de vivre par procuration deux thèses supplémentaires, pour le meilleur et pour le pire. Pour cela, et pour tous les autres bons moments, trop nombreux pour que je me

I. Au cas où, je précise le caractère inclusif de ce « ou ».

souvienne de tous, je m'estime avoir été excessivement chanceux de les avoir à mes côtés.

Enfin, je ne saurais terminer sans exprimer ma profonde reconnaissance à ma famille. Je n'aurais pu être là où j'en suis aujourd'hui si je n'avais pas été continûment poussé à faire ce dont je rêvais et à me donner les moyens d'y parvenir.

Adrien

P.-S. Auprès de ceux que j'aurais malheureusement omis et qui se sentiraient lésés, par avance je m'excuse. Le cas échéant, je vous invite à me contacter pour que je puisse vous faire parvenir une version officieuse de ce manuscrit qui répare cet oubli.

Table des matières

Table des figures

Liste des tableaux

Introduction

« On fait la science avec des faits, comme on fait une maison avec des pierres : mais une accumulation de faits n'est pas plus une science qu'un tas de pierres n'est une maison. »

Henri POINCARÉ

“CES QUELQUES PAGES *abordent brièvement les motivations qui sous-tendent les recherches présentées dans la suite du présent document. Les usages informatiques évoluant vers le traitement d'informations toujours plus abondantes, il est donc pertinent de chercher à développer des puces d'accélération matérielle spécialisées pour surmonter les limites rencontrées par les technologies actuelles. Le fonctionnement des cerveaux biologiques offre notamment une source d'inspiration prometteuse pour concevoir des systèmes artificiels énergétiquement plus efficaces pour réaliser des tâches que nous résolvons quotidiennement sans difficulté majeure, telles qu'identifier les émotions d'un interlocuteur, mais qui demeurent complexes pour les architectures de calcul conventionnelles. Les nouvelles technologies de mémoire non volatile constituent un levier de développement potentiel important de tels systèmes neuro-inspirés, en offrant de manière intrinsèque des comportements physiques qui peuvent être exploités pour permettre à ces derniers d'*apprendre *la solution de la tâche à résoudre. ”*

Une avalanche de données

Un constat général dans le domaine des technologies de l'information et de la communication (TIC) est l'augmentation au fil du temps de la quantité annuelle des données produites et manipulées dans le monde[1]. Il est attendu que cette tendance se poursuive dans les années à venir, avec par exemple un doublement du trafic Internet d'ici à trois ans[2]. La naissance d'un « Internet des objets » dont la taille croît d'année en année[3] participe également à l'explosion du volume de données numériques en circulation. D'un point de vue énergétique, environ 5 % de la consommation électrique mondiale est consacrée au traitement informatique de ces données numériques[4] et si la croissance de cette proportion s'est révélée plus faible que les prévisions[5], elle se poursuit néanmoins.

Cette avalanche des données disponibles a favorisé l'émergence de nouveaux usages informatiques[6], tels que la classification d'images et de données vidéo, la reconnaissance vocale et du langage naturel ou encore l'analyse des tendances. Une caractéristique générale des ces usages est de requérir des systèmes pouvant tolérer des données d'entrée incomplètes, bruitées voire partiellement erronées. Ces tâches sont généralement qualifiées de « cognitives » en raison de leur similarité avec celles que les entités vivantes dotées d'un cerveau réalisent dans leur vie quotidienne sans difficulté particulière : reconnaître un prédateur, communiquer avec des congénères, anticiper une trajectoire, etc.

Les réseaux de neurones artificiels comptent parmi les techniques d'apprentissage automatique qui sont particulièrement adaptées pour traiter les tâches cognitives. Le principe de ces systèmes consiste à *apprendre* la solution au problème qui leur est soumis, grâce à la présentation de nombreux exemples d'entrée[7]. Il est à noter que les techniques utilisées à l'époque actuelle nécessitent pour la plupart de disposer d'un volume conséquent de données étiquetées, c'est-à-dire d'exemples d'entrée pour lesquels la réponse attendue est également fournie au système. Malheureusement dans la pratique, les données disponibles ne sont le plus souvent pas étiquetées de la sorte. Le recours à une intervention humaine pour construire un ensemble de données d'entrée adéquates à partir d'un sous-ensemble des données à traiter est alors généralement la solution retenue, ce qui est un processus coûteux en termes de temps comme de ressources humaines.

1. M. Hilbert et P. López, Science, 2011.
2. Visual Networking Index Cisco, Cisco white paper, 2013.
3. J. Gubbi et coll., Future Generation Computer Systems, 2013.
4. B. Lannoo et coll., Network of Excellence in Internet Science, FP7, 2013.
5. Académie des technologies, EDP Sciences, 2015.
6. P. Dubey, Technology Intel Magazine, 2005.
7. A. Krizhevsky, I. Sutskever et G. E. Hinton, Advances in neural information processing systems, 2012.

La loi de MOORE peut-elle encore nous sauver ?

Du fait d'excellentes propriétés de passage à l'échelle, la technologie des transistors métal-oxyde-semiconducteur complémentaires (CMOS[I]), est une pierre angulaire de la « loi de MOORE » qui vise au doublement biennal du nombre transistors sur une puce électronique[8]. Historiquement, l'association de cette croissance exponentielle à celle de la fréquence d'horloge a permis à l'industrie de répondre à l'augmentation de la demande en capacités de traitement informatique. La mise en œuvre des réseaux de neurones artificiels s'est ainsi développée majoritairement sous une forme logicielle adaptée à une exécution sur des processeurs généralistes.

Malheureusement, les dimensions toujours plus réduites des transistors les rendent plus sensibles à la variabilité des procédés de fabrication ou au bruit électronique, exigeant des investissements toujours plus conséquents[II] et des développements toujours plus complexes pour passer au nœud technologique suivant[III]. Par ailleurs, l'augmentation de la fréquence d'horloge a quant à elle accru la puissance surfacique dissipée par les processeurs jusqu'à atteindre la limite pouvant être évacuée par un refroidissement actif à air (de l'ordre de $100\,\mathrm{W{\cdot}cm^{-2}}$). Pour cette raison, la fréquence d'horloge maximale des processeurs s'est stabilisée depuis une dizaine d'années à quelques gigahertz.

L'étude d'approches différentes de celle se basant sur la « loi de MOORE » pour augmenter les capacités de traitement des processeurs connait ainsi depuis quelques années un nouveau souffle[9,10]. Certaines approches proposent par exemple de se tourner vers la conception d'architecture probabilistes, qui *exploitent* les nanocomposants non fiables plutôt que de chercher à en rendre déterministe le comportement[11].

Une solution possible : l'approche neuromorphique

Comme l'illustre la figure 1, l'évolution à la hausse de la fréquence d'horloge à travers les différentes générations de processeurs a tendu à éloigner ces derniers de l'approche observée dans les cerveaux biologiques. Ceux-ci associent en effet une fréquence environ cent millions de fois plus faible que celles des processeurs actuels à un fonctionnement massivement plus parallèle, l'un des avantages de cette approche étant d'ordre énergétique[12]. Il semble donc raisonnable de chercher à s'inspirer du fonctionnement des cerveaux biologiques pour réaliser des puces accélératrices basse consommation spécialisées dans les tâches cognitives. De telles puces sont alors qualifiées de « neuromorphiques ».

Les réseaux de neurones biologiques sont par ailleurs capables de fonctionner malgré le

I. *Complementary Metal-Oxide-Semiconductor* en langue anglaise.
8. G. MOORE, Electronics, 1965.
II. http://semiengineering.com/how-much-will-that-chip-cost/
III. http://semiengineering.com/10nm-versus-7nm/
9. W. ARDEN et coll., ITRS, 2010.
10. K. ROY et coll., IEEE Design Test, 2016.
11. K. PALEM et A. LINGAMNENI, Proceedings of the Annual Design Automation Conference, 2012.
12. D. ATTWELL et S. B. LAUGHLIN, Journal of Cerebral Blood Flow & Metabolism, 2001.

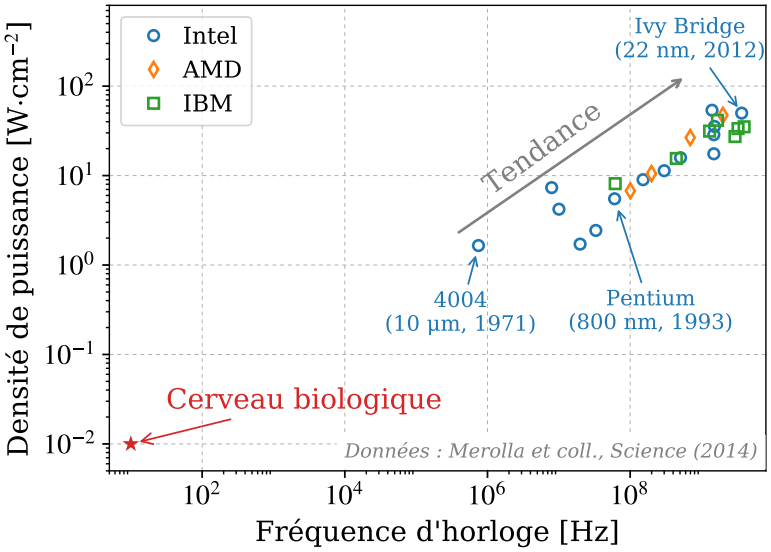


FIGURE 1 – Densité de puissance dissipée, en fonction de la fréquence d'horloge utilisée, par des processeurs généralistes conçus par trois des principaux acteurs industriels du domaine (Intel® ○, AMD® ◇ et IBM® □). La tendance historique est à l'éloignement avec la situation d'un cerveau biologique. *Source* : figure adaptée de la référence [13].

recours à des composants peu fiables, pouvant même disparaître (mort cellulaire), notamment parce que l'apprentissage permet de tenir compte des défauts et variations des composants disponibles. Nous pouvons donc espérer que la « neuro-inspiration » permette naturellement d'atténuer les problèmes posés par l'influence croissante des fluctuations sur les nanocomposants de dimensions toujours plus réduites.

La dépendance exponentielle du courant de drain à la tension de grille d'un transistor à effet de champ en régime de faible inversion peut être utilisée pour simuler de manière analogique des comportements biochimiques similaires fréquemment rencontrés dans les neurones biologiques. Énoncée à la fin des années 1980 par Carver MEAD[14], cette idée a encouragé le développement de neurones artificiels analogiques en technologie métal-oxyde-semiconducteur, basse consommation puisqu'ils exploitent essentiellement les *fuites* des transistors qui les composent[15–17].

Néanmoins, des défis demeurent avant d'espérer connaître l'adoption massive de systèmes neuromorphiques. L'une des caractéristiques principales des réseaux neuronaux est qu'ils intriquent de façon importante les neurones, tournés vers le calcul, et les synapses, chargées d'apprendre et de mémoriser les connexions entre ces derniers. Cette topologie particulière est mal adaptée aux architectures de calcul conventionnelles, où la mémoire est généralement

14. C. MEAD, Proceedings of the IEEE, 1990.
15. C. MEAD, Analog VLSI and Neural Systems, 1989.
16. S. SAÏGHI et coll., IEEE Transactions on Biomedical Circuits and Systems, 2011.
17. G. INDIVERI et coll., Frontiers in Neuroscience, 2011.

éloignée des unités de calcul et de contrôle[18]. Pour pouvoir réaliser des systèmes artificiels efficaces, il apparaît donc nécessaire de pouvoir disposer à proximité des circuits neuronaux d'une quantité de mémoire significative et de préférence capable de mettre en place l'apprentissage du système.

À la manière des circuits neuronaux susmentionnés qui exploitent la physique des transistors à effet de champ pour en faire davantage que de simples interrupteurs, *il semble pertinent de chercher à exploiter la physique de composants mettant en œuvre les fonctions synaptiques pour aller au-delà du rôle d'une simple banque de mémoire co-intégrée*. À ce titre, les technologies mémoires innovantes développées actuellement ouvrent des pistes de recherches intéressantes. En effet, ces nouveaux nanocomposants offrent des propriétés particulièrement pertinentes pour un usage synaptique, telle qu'un caractère non volatil, une densité d'intégration élevée ainsi que pour certaines technologies, *des comportements physiques intrinsèques pouvant être exploités pour mettre en œuvre l'apprentissage désiré*. Toutefois, les candidats sont encore nombreux, à différents états de maturité technologique et présentent chacun leurs avantages et leurs inconvénients.

Dans ce travail de thèse, nous évaluons le potentiel d'application synaptique de deux familles de technologies mémoires innovantes : les jonctions tunnel magnétiques à transfert de spin et les cellules électrochimiques métalliques de sulfure d'argent (Ag_2S). L'un des fils directeurs des travaux présentés est l'exploitation à notre avantage de la richesse des comportements intrinsèques offerts par la physique de ces nanocomposants, pour faire de ces derniers des synapses artificielles capables de mettre directement en œuvre tout ou partie de l'apprentissage d'un système neuromorphique. Nous verrons notamment que les familles de composants retenues sont naturellement adaptées à des apprentissages non supervisés, qui peuvent être utilisés sans données étiquetées et sont de ce fait pressentis pour être particulièrement appropriés aux usages à venir[19]. Les études présentées explorent de multiples échelles : celle du nanocomposant unique, celle d'un système neuro-inspiré entier ou encore celle des circuits électroniques assurant le fonctionnement d'un tel système. Les simulations informatiques réalisées pour cela sont basées sur des modèles analytiques des composants synaptiques et des outils logiciels développés durant la thèse.

Structure du document

À l'issue d'une présentation des grandes tendances dans le domaine des architectures neuromorphiques (chapitre 1), nous étudions le potentiel d'application synaptique d'une première famille de candidats synaptiques : les jonctions tunnel magnétiques à transfert de spin. Ces composants d'électronique de spin[I], qui n'exploitent pas uniquement la charge des électrons mais également leur moment magnétique intrinsèque (spin), sont l'objet de développements

18. G. INDIVERI et S.-C. LIU, Proceedings of the IEEE, 2015.
19. Y. LECUN, Y. BENGIO et G. HINTON, Nature, 2015.

I. Ce domaine est également connu sous le nom de « spintronique ».

industriels importants pour concevoir de nouvelles puces mémoires, capables de remplacer les mémoires informatiques conventionnelles. Après un premier travail de modélisation du comportement naturellement stochastique de ces nanocomposants (chapitre 2), nous montrerons qu'il est possible de réaliser un système qui les utilise comme synapses binaires et demeure capable d'apprendre une tâche cognitive par le biais d'une stratégie d'apprentissage probabiliste exploitant leur comportement physique intrinsèque (chapitre 3). Les seconds candidats synaptiques étudiés sont des cellules électrochimiques métalliques Ag_2S, fabriquées par nos collaborateurs de l'Institut d'Électronique, de Microélectronique et de Nanotechnologie de l'Université de Lille I. Après une présentation du modèle analytique développé pour rendre compte des multiples comportements qu'offre la physique de ces composants, nous présenterons une preuve de concept qui illustre à l'échelle d'un système la possibilité d'exploiter l'interaction de ces comportements mémoires pour réaliser l'apprentissage d'une tâche cognitive dynamique (chapitre 4). Des obstacles se posent néanmoins à l'échelle des circuits électroniques permettant de réaliser de tels systèmes neuromorphiques. Nous poserons des pistes de réflexions quant aux principaux de ces défis dans le cas de l'utilisation de jonctions tunnel magnétiques à transfert de spin comme synapses artificielles binaires (chapitre 5).

Un point sur les unités

Sauf mention contraire explicite, les équations présentées dans ce manuscrit sont exprimées dans le système international (SI) d'unités.

Afin d'éviter les imprécisions, les préfixes *binaires* sont privilégiés pour désigner les quantités de mémoire. À la différence des préfixes décimaux du système international d'unités reposant sur des puissances de 10, ils permettent d'exprimer naturellement les capacités des mémoires informatiques, à l'aide de puissances de 2. Un rappel des principaux préfixes de chaque famille est donné dans la table 1.

Préfixe binaire	Nom	Facteur	Préfixe décimal	Nom	Facteur
Ki	kibi	2^{10}	k	kilo	10^3
Mi	mébi	2^{20}	M	méga	10^6
Gi	gibi	2^{30}	G	giga	10^9
Ti	tébi	2^{40}	T	téra	10^{12}
Pi	pébi	2^{50}	P	péta	10^{15}
Ei	exbi	2^{60}	E	exa	10^{18}

TABLE 1 – Noms et significations des principaux préfixes binaires et décimaux.

Chapitre 1

Solutions neuro-inspirées : des origines aux tendances matérielles actuelles

> "We can only see a short distance ahead, but we can see plenty there that needs to be done."
>
> Alan TURING

"CE CHAPITRE *dresse un panorama des systèmes de calcul neuro-inspirés. Nous verrons comment les premiers travaux de formalisation des réseaux neuronaux artificiels ont évolué pour donner naissance aux systèmes que nous connaissons aujourd'hui, tels que les réseaux profonds ou impulsionnels. Devant l'intérêt croissant pour ce type de systèmes, se pose légitimement la question de l'amélioration de leur efficience, à laquelle nous répondrons par une revue des principales pistes de recherches actuelles. Une attention particulière sera portée aux architectures matérielles développées récemment par la communauté des systèmes neuromorphiques. Enfin, nous verrons dans quelle mesure les nouveaux nanocomposants mémoires innovants ont le potentiel d'améliorer ces puces neuro-inspirées, en permettant de réaliser des synapses artificielles capables d'effectuer un apprentissage directement sur puce tout en maintenant une densité d'intégration élevée.* "

Introduction

Les réseaux de neurones artificiels sont un ensemble de techniques du domaine de l'apprentissage automatique, dont les principes *s'inspirent* des réseaux de neurones biologiques. Comme nous l'avons mentionné dans l'introduction générale, ces systèmes sont particulièrement adaptés aux tâches cognitives, telles que la classification d'images et de données vidéo, l'interaction en langage naturel ou encore la prédiction de tendances. L'augmentation croissante de ces usages pose néanmoins, de manière plus aigüe que par le passé, les questions de la mise en œuvre efficace des réseaux de neurones artificiels et de leur passage à l'échelle.

Dans ce chapitre, nous procédons à une présentation générale de ces systèmes, de façon à en identifier les traits caractéristiques et les pistes d'amélioration possibles. Un intérêt particulier est porté sur les projets actuels d'architectures matérielles optimisées pour l'exécution de réseaux de neurones artificiels. Nous verrons notamment que le recours à des technologies mémoires innovantes pourrait offrir la possibilité de réaliser des systèmes capables d'apprendre *directement* sur puce.

1.1 Évolution des réseaux de neurones artificiels

Depuis leur apparition au cours du XX^{e} siècle, les réseaux de neurones artificiels ont connu plusieurs évolutions, les dernières en date étant les réseaux profonds ainsi que les réseaux impulsionnels. Jusqu'à très récemment, la mise en œuvre en production de réseaux de neurones artificiels était quasiment exclusivement réalisée de manière logicielle, tandis que la recherche sur les approches matérielles demeurait un domaine de niche. Ces dernières sont toutefois désormais l'objet de plus en plus de projets de recherche, portées par les avancées dans le domaine de l'électronique neuromorphique, qui s'inspire du fonctionnement d'un cerveau biologique, et l'émergence de nanocomposants mémoires innovants.

1.1.1 Formalisation des réseaux de neurones

Les briques de base en biologie. Dans un cerveau biologique, les neurones constituent les unités de traitement de l'information. L'information d'un neurone (présynaptique) est transmise sous la forme de « potentiels d'action », des impulsions électriques qui se propagent le long d'un de ses axones vers une dendrite d'un autre neurone (postsynaptique). La figure 1.1 illustre ce processus de façon (très) schématique. La jonction chimique reliant l'axone et la dendrite est une synapse, dont le « poids » détermine la quantité de courant que reçoit le neurone postsynaptique en provenance du neurone présynaptique.

Lors de l'apprentissage, le poids des synapses est modulé, selon une « règle d'apprentissage ». Si le poids augmente ou diminue, il est respectivement question de « potentialisation » ou de « dépression » synaptique. Cette « plasticité synaptique » peut se manifester à différentes échelles de temps.

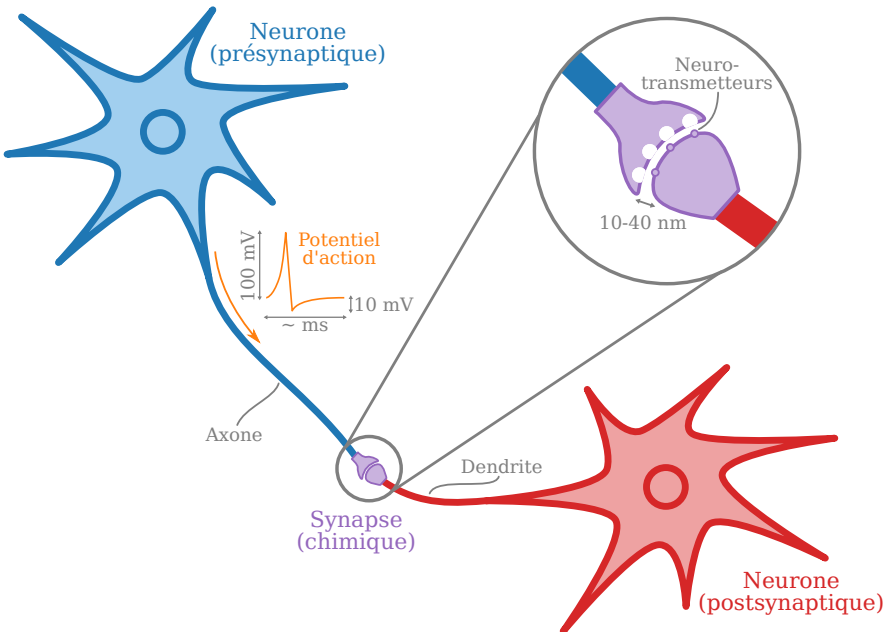


FIGURE 1.1 – Schéma de principe d'un couple de neurones reliés par une synapse dans un cerveau biologique.

Les cerveaux biologiques sont des systèmes remarquables de par leur extrême parallélisme. Si un cerveau humain contient par exemple aux alentours de cent milliards (10^{11}) de neurones[20], le nombre de synapses qui leur est associé est quant à lui estimé à un billiard[21] (10^{15}). La sortance moyenne des neurones biologique est ainsi de l'ordre de plusieurs milliers. Cette situation est très différente de celle des circuits électroniques numériques modernes, dans lesquels la sortance moyenne est de quelques fois l'unité seulement.

Vers une représentation mathématique : le neurone formel. Au début des années 1940, les chercheurs en neurologie et psychologie cognitive MCCULLOCH et PITTS proposent une modélisation mathématique de la fonction neuronale[22]. Ce neurone formel intègre un nombre N de valeurs d'entrées binaires x_i (provenant d'autres neurones), préalablement pondérées par la valeur de leur synapse respective w_i. La sortie binaire y_j dépend alors d'une « activation », selon que la somme précédente excède ou non un seuil ϑ_j. Mathématiquement, ce comportement s'exprime par

$$y_j = \mathrm{h}\left(-\vartheta_j + \sum_{i=0}^{N-1} w_i x_i\right) \tag{1.1}$$

20. F. A. C. AZEVEDO et coll., Journal of Comparative Neurology, 2009.
21. P. LENNIE, Current Biology, 2003.
22. W. S. MCCULLOCH et W. PITTS, The bulletin of mathematical biophysics, 1943.

où h est une fonction de HEAVISIDE, par exemple

$$\mathrm{h}(u) = \begin{cases} 1 & \text{si } u \geq 0 \\ 0 & \text{sinon} \end{cases}, \tag{1.2}$$

et est résumé graphiquement sur la figure 1.2.

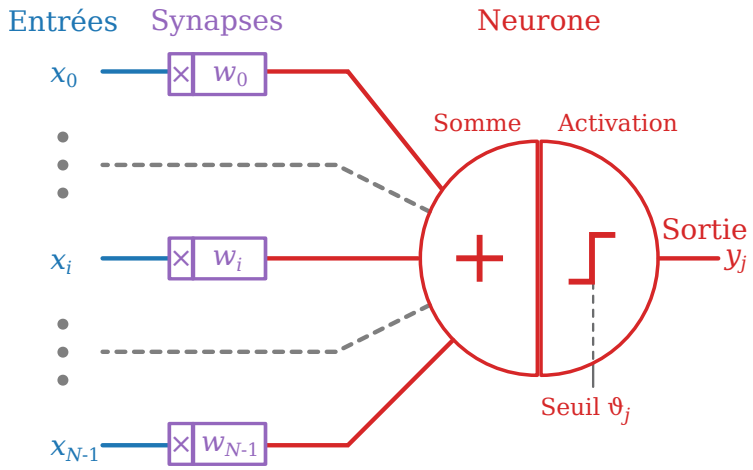


FIGURE 1.2 – Schéma du neurone formel de MCCULLOCH et PITTS. Les valeurs (binaires) x_i présentées en entrée sont multipliées par celle w_i de leur synapse respective avant d'être sommées par le neurone, dont la valeur (binaire) de sortie y_j dépend d'une activation consistant à comparer la somme des entrées pondérées au seuil ϑ_j.

Cette définition formelle de la fonction neuronale, basée sur une somme de pondérations synaptiques suivie d'un mécanisme d'activation, est encore aujourd'hui une source d'inspiration de nombreux modèles de neurones artificiels plus récents .

Modulation des poids synaptiques. Comme mentionné précédemment, la plasticité synaptique est au cœur du processus d'apprentissage des réseaux de neurones. Cette modulation des poids synaptiques est régie par un algorithme, qualifié de « règle d'apprentissage », simple en regard de la tâche à remplir et qui fait évoluer les poids synaptiques en fonction notamment des exemples présentés à l'entrée du système et des prédictions de ce dernier. Il s'agit-là de la grande force des réseaux de neurones artificiels : nul besoin de prévoir à l'avance l'ensemble des situations possibles et des actions correspondantes. Au contraire, un réseau de neurones construit petit à petit une représentation *interne* de la tâche à effectuer. Ceci lui permet ensuite de *généraliser* ses prédictions à des situations non rencontrées durant l'apprentissage, ainsi que de pouvoir continuer à *fonctionner avec des données floues, incomplètes ou bruitées*, ainsi que de mieux *tolérer la variabilité* à laquelle peuvent être soumis ces différents composants[23,24].

23. D. QUERLIOZ et coll., IEEE Transactions on Nanotechnology, 2013.
24. P. MEROLLA et coll., arXiv preprint arXiv :1606.01981, 2016.

Un exemple connu de stratégie d'apprentissage est la règle de HEBB[25], parfois appelée « théorie des assemblées de neurones », que ce dernier formulait par :

> *« Let us assume that the persistence or repetition of a reverberatory activity (or "trace") tends to induce lasting cellular changes that add to its stability... When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased. »*

Cette hypothèse fut initialement formulée en neurosciences pour décrire le cas d'une potentialisation synaptique lorsqu'un neurone présynaptique stimule de façon répétée et persistante un neurone postsynaptique. Il s'agit également d'une source d'inspiration de nombreuses règles d'apprentissage dans le domaine des réseaux de neurones artificiels, notamment celles utilisées dans les travaux présentés dans les chapitres suivants. La règle de HEBB est souvent résumée comme l'idée que des neurones qui s'activent en même temps se lient ensemble (« *neurons wire together if they fire together* »[26]).

Trois modalités d'apprentissage sont généralement distinguées dans le domaine de l'apprentissage automatique (et schématisées sur la figure 1.3 page suivante) :

- l'*apprentissage supervisé*, où une correction explicite des poids synaptiques est calculée et appliquée suite aux prédictions faites sur un ensemble d'entrées étiquetées (c'est-à-dire dont la prédiction correcte attendue est connue) ;
- l'*apprentissage par renforcement*, une autre méthode supervisée, moins explicite, dans laquelle le système adapte son comportement de façon à maximiser une récompense provenant de son environnement extérieur ;
- l'*apprentissage non supervisé* qui consiste à inférer *sans nécessiter d'étiquetage* une structure sous-jacente aux données qui lui sont présentées.

1.1.2 Du perceptron à l'apprentissage profond

Le perceptron. À la fin des années 1950, ROSENBLATT présente le perceptron[27], un système qui combine le neurone formel de MCCULLOCH et PITTS avec une règle d'apprentissage. Il s'agit d'un classifieur linéaire à une couche et une sortie (binaire également), permettant de séparer en deux classes les éléments présentés en entrée, au moyen d'un hyperplan dont l'orientation est définie par les poids synaptiques. Ces poids sont ajustés lors de l'apprentissage en cas d'erreur de prédiction. Les deux phases de cet apprentissage supervisé sont ainsi :

1. une phase d'*inférence*, où la classe de l'exemple présenté est prédite selon l'équation (1.1), schématisée à la figure 1.2 ;
2. une phase d'*apprentissage*, durant laquelle les poids synaptiques w_i sont mis à jour en

25. D. O. HEBB, The organization of behavior : A neuropsychological approach, 1949.
26. S. LÖWEL et W. SINGER, Science, 1992.
27. F. ROSENBLATT, Psychological review, 1958.

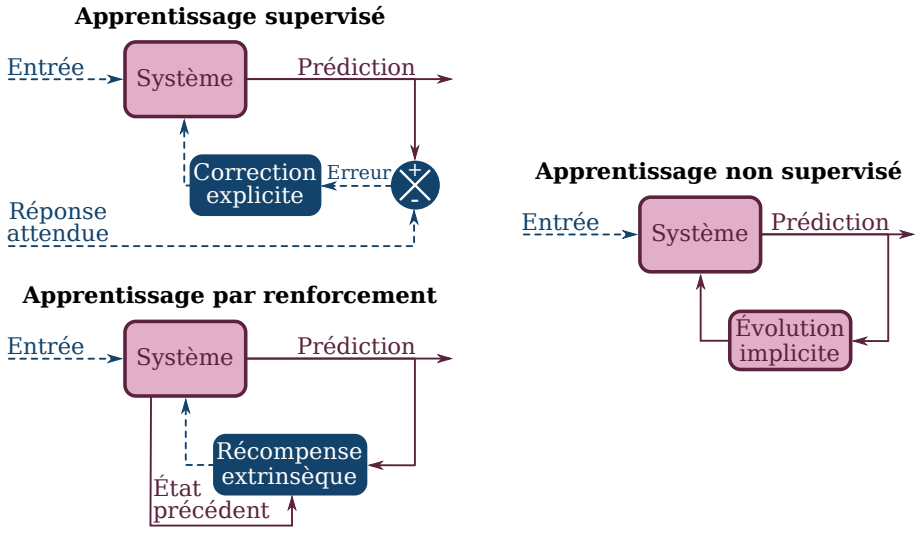


FIGURE 1.3 – Les grandes familles d'algorithmes d'apprentissage. À gauche les méthodes supervisées, soit de manière explicite, soit par le biais d'une récompense à maximiser ; à droite l'approche non supervisée.

cas de contradiction entre une prédiction y_j et la réponse attendue t_j, selon le schéma

$$w_i \leftarrow w_i + \rho \cdot x_i \cdot \underbrace{(t_j - y_j)}_{\text{Erreur}} \tag{1.3}$$

où x_i est la valeur de l'entrée reliée au poids synaptique w_i, tandis que ρ est un paramètre réel régulant la vitesse d'apprentissage. L'idée principale de cette phase est la modification *incrémentale* des poids synaptiques, en fonction de l'erreur, de façon à obtenir les prédictions attendues.

Vers des réseaux à plusieurs couches. Si les perceptrons sont capables d'apprendre à classifier des problèmes linéairement séparables (figure 1.4a), cette approche montre ses limites dans le cas des problèmes non linéairement séparables (figure 1.4b, ou plus simplement l'exemple d'un « ou exclusif »). Une solution possible à cette limitation est d'associer des couches successives, à la manière de la figure 1.5, ce qui permet de délimiter l'espace des classes de façon plus complexe qu'avec un unique hyperplan. Le cercle en pointillé de l'exemple (b) de la figure 1.4 pourrait ainsi être approximé par une combinaison de droites.

Remarque

Dans le présent document, nous nous focaliserons sur les réseaux de neurones artificiels dits « acycliques » (*feed-forward* en anglais), qui ne présentent pas de connexions récurrentes entre leurs différentes couches.

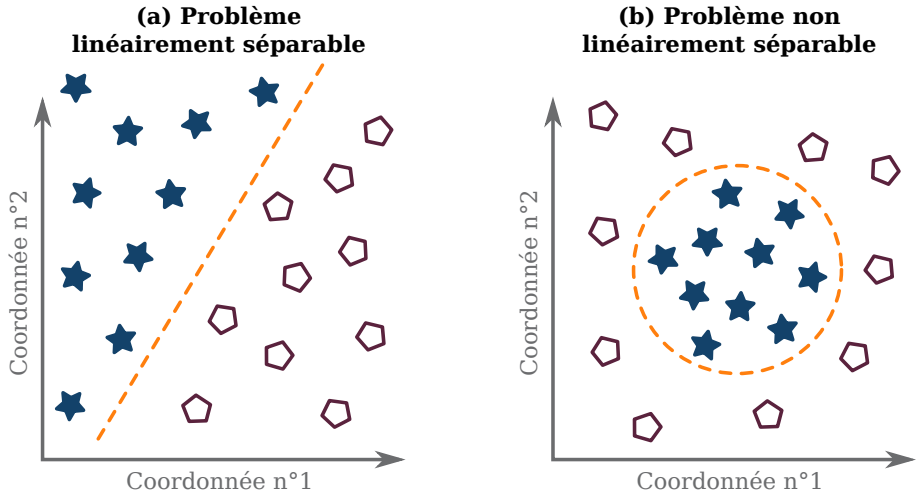


FIGURE 1.4 – Problèmes de classification à deux dimensions des classes ⬠ et ★. *(a)* Exemple linéairement séparable (par exemple selon la droite en pointillé). *(b)* Exemple non linéairement séparable.

En parallèle de l'association de plusieurs couches, quelques évolutions supplémentaires ont fait leur apparition, telles que l'utilisation d'autres fonctions d'activation (sigmoïde, tangente hyperbolique ou encore rampe pour ne citer que les plus employées) ainsi que le recours à des entrées et sorties analogiques ce qui permet par exemple d'obtenir des classificateurs à plus de deux classes.

L'apprentissage par des systèmes présentant des couches « cachées » est plus complexe que dans le cas de réseaux de neurones plus simples. Originellement, l'approche généralement adoptée consistait à associer successivement des extracteurs de caractéristiques (par exemple des détecteurs de fronts dans un problème de traitement d'images) minutieusement conçus et ajustés manuellement, suivis d'une dernière couche constituée d'un classificateur où le réel apprentissage automatique avait lieu. Le grand défaut de cette façon de faire était de nécessiter beaucoup d'interventions humaines, notamment lors de la conception des premières couches.

Le défi de l'apprentissage à plusieurs couches. Les années 1970 et 1980 voient la mise au point par plusieurs groupes en parallèle[28–31] d'une règle d'apprentissage qui permet de réellement *entraîner* les multiples couches d'un réseau de neurones artificiels à couches cachées. Il s'agit de l'algorithme de « rétropropagation du gradient » (*backpropagation* en anglais), qui rend possible un apprentissage supervisé, en s'appuyant sur la règle de dérivation en chaîne pour propager successivement vers l'arrière la contribution des différentes couches à l'erreur de prédiction et permettre la modulation de leurs poids synaptiques respectifs. Il est à noter

28. P. WERBOS, Ph.D. thesis, 1974.
29. D. B. PARKER, Learning-logic : Casting the cortex of the human brain in silicon, 1985.
30. Y. LECUN, Proceedings of Cognitiva, 1985.
31. D. E. RUMELHART, G. E. HINTON et R. J. WILLIAMS, DTIC Document, 1985.

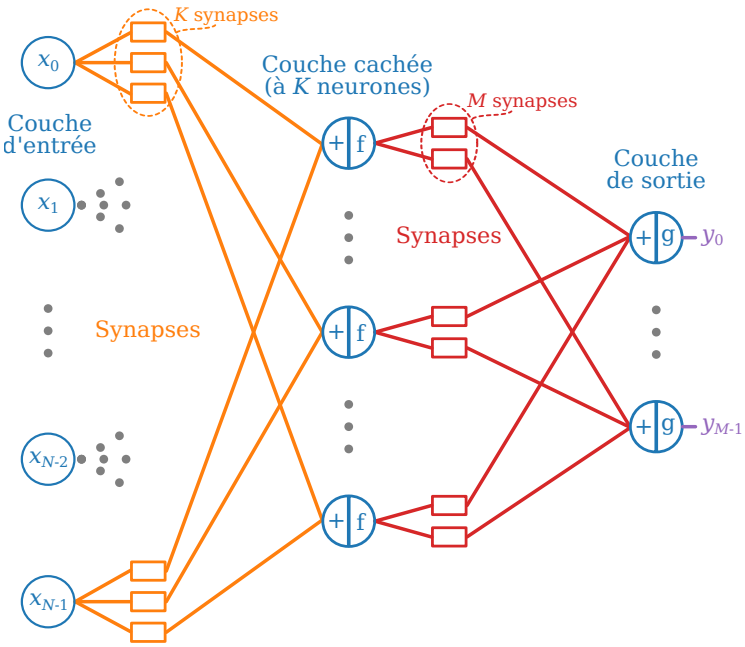


FIGURE 1.5 – Exemple d'un réseau de neurones artificiels à N entrées et M sorties comportant une couche cachée de K neurones, en sus des couches d'entrée et de sortie. Dans cet exemple précis, tous les neurones d'une couche sont connectés à l'ensemble des neurones de la couche suivante. Les fonctions d'activation employées sont désignées par f et g.

que cette règle d'apprentissage requiert l'utilisation des fonctions d'activation *dérivables*.

L'intérêt d'un algorithme tel que celui de rétropopagation du gradient est de permettre aux différentes couches de neurones *d'apprendre* des associations non linéaires d'entrées et de sorties successives, qui leur permettent d'extraire au fur et à mesure des caractéristiques de plus en plus haut niveau des signaux présentés. La grande force de cet algorithme est de fortement limiter les interventions humaines (en comparaison des méthodes antérieures), si ce n'est lors de la recherche des paramètres régulant l'apprentissage [I] ainsi que lors de l'étiquetage des données servant à réaliser ce dernier.

L'algorithme de rétropropagation du gradient souffre néanmoins d'un inconvénient intrinsèque, connu sous le nom de « problème de l'affaiblissement du gradient » (*vanishing gradient problem* en anglais)[32]. L'affaiblissement de la valeur numérique du gradient lors de la diffusion rétrograde de ce dernier tend en effet à rendre particulièrement lent l'apprentissage des couches les plus éloignées de la sortie, c'est-à-dire les premières. Ceci est d'autant plus problématique que les couches initiales sont à priori celles qui extraient les caractéristiques (simples) de plus bas niveaux, sur lesquelles reposent les motifs de plus haut niveaux assemblés par les couches ultérieures. Jusqu'à il y a désormais une dizaine d'années, ce phénomène a empêché l'apparition de réseaux de neurones entièrement connectés comportant davantage que

I. Le terme ρ de l'équation (1.3) de l'algorithme d'apprentissage du perceptron en est un exemple.
32. S. HOCHREITER et coll., A field guide to dynamical recurrent neural networks, 2001.

quelques couches.

Apprentissage initial non supervisé : vers des réseaux réellement profonds. Le problème de l'affaiblissement du gradient ne faisant que ralentir la convergence de l'apprentissage par rétropropagation de ce dernier, une solution naturelle est d'augmenter le nombre d'exemples étiquetés utilisés. La présentation d'une grande quantité d'exemples est également utile pour limiter le surapprentissage[I] et accroître les capacités de généralisation du réseau entraîné. Malheureusement l'étiquetage des données est un processus coûteux puisqu'il requiert généralement une intervention humaine. Afin de pallier le nombre limité d'exemples étiquetés disponibles, certaines rustines sont parfois employées, telles que la construction de versions légèrement différentes de ces exemples en les bruitant ou en les distordant légèrement (il est alors question d'« augmentation artificielle des données »).

Des démonstrations d'apprentissage sur des réseaux de neurones artificiels véritablement profonds (c'est-à-dire comportant un grand nombre de couches intermédiaires) sont présentées à partir de 2006 dans la littérature[33,34]. L'innovation est le recours à un apprentissage *initial* non supervisé et réalisé couche par couche, permettant de placer le réseau dans une position de départ plus favorable à l'apprentissage supervisé conventionnel qui vient ensuite (par exemple par rétropropagation du gradient). L'intérêt pour les réseaux profonds s'en trouve relancé, l'accélération de l'apprentissage supervisé atténuant par ailleurs la nécessité de disposer d'une grande base d'exemples étiquetés, coûteuse à établir.

L'émergence spontanée de cartes de fonctions à l'issue de l'étape de « pré-entraînement » n'est pas sans rappeler celles du cortex visuel du chat[35] ou du singe[36] rapportées en biologie, qui apparaissent suite à l'observation répétée de certaines formes et orientations après la naissance.

Quand le paradigme d'architecture classique des ordinateurs ne suffit plus. L'un des verrous à la mise en œuvre efficace des réseaux de neurones artificiels sur une architecture matérielle est illustré à la figure 1.6 page suivante. Les réseaux de neurones artificiels sont structurés en couches dont les unités de calcul (les neurones) sont très fortement interconnectées (figure1.6a). Leur fonctionnement est hautement *parallèle* et nécessite finalement peu de calcul en regard du nombre d'accès mémoires « simultanés ». À l'inverse, les ordinateurs actuels présentent une architecture de VON NEUMANN, où un processeur généraliste (regroupant des unités de calcul et la logique de contrôle associée) accède de façon *séquentielle* à une banque de mémoire centrale au travers d'un bus de communication. Cette inadéquation entre l'architecture matérielle des ordinateurs actuels et la topologie des réseaux de neurones artificiels

I. Il s'agit de la spécialisation excessive du système aux exemples utilisés durant l'apprentissage, qui entraîne une baisse de ces performances sur des exemples inconnus.

33. G. E. HINTON, S. OSINDERO et Y.-W. TEH, Neural computation, 2006.
34. Y. BENGIO et coll., Advances in neural information processing systems, 2007.
35. M. C. CRAIR, D. C. GILLESPIE et M. P. STRYKER, Science, 1998.
36. G. G. BLASDEL, The Journal of Neuroscience, 1992.

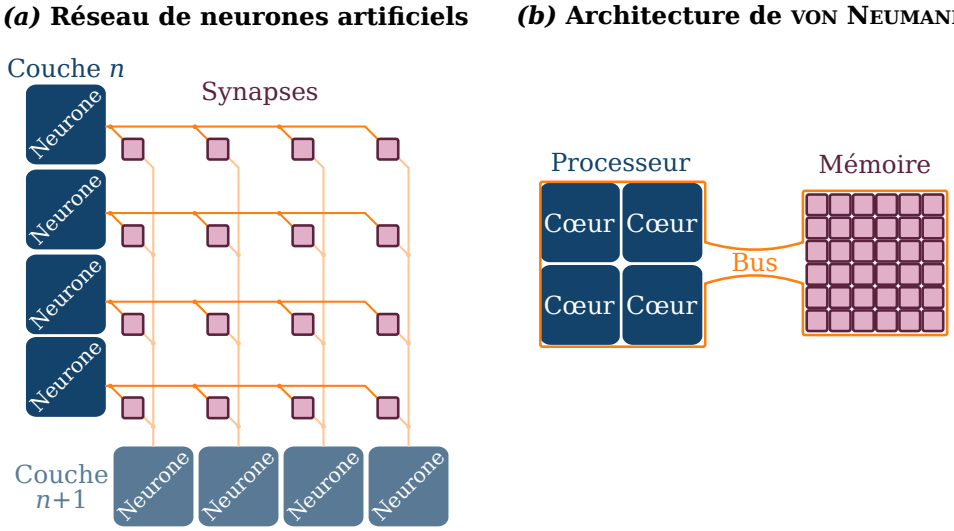


FIGURE 1.6 – *(a)* Architecture d'une couche de type « entièrement connectée » d'un réseau de neurones artificiels : chaque neurone de la couche d'entrée (n) est connecté à chaque neurone de la couche suivante ($n+1$) par le biais d'une synapse.
(b) Architecture de VON NEUMANN : le processeur (qui regroupe des unités arithmétiques et logiques munie d'une unité de contrôle des instructions) est physiquement éloigné de la mémoire centrale, avec laquelle il communique à travers un bus.

constitue le « goulot d'étranglement de VON NEUMANN »[37]. La lecture et la modification des paramètres synaptiques, stockés dans la mémoire centrale et non en cache du fait de leur nombre trop important, nécessite ainsi de fréquents accès mémoires, coûteux en énergie. La nature séquentielle de l'architecture matérielle limite également la rapidité d'exécution, en raison du nombre limité de cœurs de calcul.

Le recours aux processeurs graphiques (GPU[I]), disposant d'unité de calculs certes plus simples mais en nombre supérieur et associées à des banques mémoires physiquement plus proches, a participé au regain d'intérêt pour les réseaux de neurones artificiels profonds au début des années 2010, en réduisant d'un ordre de grandeur la durée d'apprentissage[38]. Ce type de processeurs est par exemple à l'origine en 2012 d'un bond très significatif des performances lors d'une compétition majeure de traitement automatique d'images[7] sur la base de données ImageNet[II]. Sur cette compétition en particulier, le taux d'erreur a été réduit de moitié en entraînant durant cinq à six jours par rétropropagation du gradient un réseau de 1,7 million de neurones et 60 millions de paramètres sur un (simple) couple de cartes graphiques *grand public*[III]. Quelques limitations ressortent néanmoins de cette étude :

37. J. BACKUS, Commun. ACM, 1978.
I. De l'anglais *Graphical Processing Unit*.
38. R. RAINA, A. MADHAVAN et A. Y. NG, Proceedings of the 26th annual international conference on machine learning, 2009.
7. A. KRIZHEVSKY, I. SUTSKEVER et G. E. HINTON, Advances in neural information processing systems, 2012.
II. http://image-net.org/
III. Deux cartes Nvidia® GTX 580 3 Gio.

1. les données étiquetées demeurent une denrée précieuse, puisque si les auteurs n'utilisent pas d'étape de « pré-entraînement » non supervisé, ils ont recours à une augmentation artificielle des données ;

2. la taille du réseau est limitée par la quantité de mémoire disponible sur les cartes graphiques ;

3. les auteurs suggèrent qu'une fois l'apprentissage réalisé, mesurer les distances par des valeurs numériques de faible dynamique pourrait être plus efficace que d'employer des valeurs réelles pleine précision.

Le système AlphaGo de Google DeepMind® (entraîné pour jouer au jeu de Go) est un exemple encore plus avancé d'utilisation de processeurs spécialisés pour exécuter un réseau de neurones artificiels. L'apprentissage a ainsi réalisé à l'aide de plus d'un millier de processeurs généralistes et d'une centaine de cartes graphiques[39]. Par ailleurs, l'entreprise a été jusqu'à concevoir un modèle de circuit intégré développé pour un client (ASIC[I]), baptisé *Tensor Processor Unit*, spécialement optimisé pour la bibliothèque logicielle d'apprentissage automatique TensorFlow[II]. Le système AlphaGo était ainsi exécuté sur un ensemble de ces puces durant son match (victorieux) contre le joueur professionnel Lee SEDOL[III]. Ces circuits exploitent une représentation numérique de précision limitée, ce qui leur permet d'être énergétiquement plus efficaces dans le cadre des tâches cognitives tout en améliorant le potentiel d'intégration.

Il est possible de commencer à tirer certaines conclusions préliminaires de ces exemples récents de mise en œuvre de réseaux de neurones artificiels :

1. une quantité importante de mémoire est requise, notamment pour décrire les paramètres synaptiques ;

2. une solution de plus en plus envisagée pour gagner en performance ou en consommation énergétique est de se tourner vers des circuits spécialisés, dont l'architecture est davantage en accord avec la topologie et les principes de base des réseaux de neurones artificiels (plus de parallélisme, mémoire plus proche des unités de calculs, etc.) ;

3. la question est posée quant à la pertinence d'utiliser un codage numérique de grande précision pour représenter les variables de ces systèmes, là où la biologie semble se satisfaire de composants peu fiables et peu précis.

Par ailleurs, sur le long terme, l'apprentissage *non supervisé* est pressenti comme primordial, autant pour son caractère biologiquement plus plausible qu'en raison de l'explosion des données non étiquetées disponibles (et le coût d'étiqueter ces dernières)[19].

39. D. SILVER et coll., Nature, 2016.
I. De l'anglais *Application-Specific Integrated Circuit* .
II. https://www.tensorflow.org/
III. https://cloudplatform.googleblog.com/2016/05/Google-supercharges-machine-learning-tasks-with-custom-chip.html
19. Y. LECUN, Y. BENGIO et G. HINTON, Nature, 2015.

Un type particulier de réseaux profonds : les réseaux convolutifs. Les réseaux de neurones artificiels convolutifs[40,41] (CNN[I]) constituent une famille de réseaux profonds dont l'approche est particulièrement adaptée aux données qui se présentent sous la forme de plusieurs tableaux associés, telles que les images matricielles, généralement représentées par trois tableaux de composantes (par exemple rouge-vert-bleu). Les résultats en traitement d'images sont ainsi probants, qu'il s'agisse de détection de panneaux de circulation[42], de localisation d'objets[43] ou encore de reconnaissance de visages[44].

Les points forts des réseaux convolutifs sont :

1. le fait de paver les entrées pour n'exploiter qu'un ensemble de connexions adjacentes, à la manière des champs récepteurs en biologie, ce qui permet de réduire fortement le nombre de poids synaptiques par rapport aux réseaux profonds précédents ;
2. le fait de partager les poids synaptiques associés aux neurones pour une même « carte de fonction » (*feature map* en anglais), ce qui limite une nouvelle fois la quantité de synapses à représenter, tout en assurant par ailleurs une invariance par translation dans l'extraction des motifs associés aux cartes de fonction ;
3. le fait d'intercaler des couches de « mise en commun » (*pooling layers* en anglais), permettant de réduire les dimensions des cartes de fonction au fur et à mesures des couches.

Les deux premiers points sont à l'origine du nom des cette architecture, puisqu'ils correspondent au calcul d'une convolution discrète, tandis que le dernier point peut être interprété comme une régularisation par sous-échantillonnage, réduisant le risque de surapprentissage. Par ailleurs, il convient de noter que d'une façon similaire aux précédent réseaux profonds, la dernière couche de ce type de structures est généralement un classique classificateur de type « entièrement connecté ».

Si le partage des poids synaptiques est assez irréaliste du point de vue biologique, cette architecture est néanmoins généralement considérée comme neuro-inspirée, du fait de sa similitude avec l'organisation en couches interconnectées du cortex visuel des mammifères[45,46].

1.1.3 Émergence des réseaux impulsionnels

En biologie, les neurones transmettent l'information sous la forme d'impulsions électriques, appelées « potentiels d'action » (figure 1.1). Si les valeurs analogiques « continues » employées dans le cadre des réseaux de neurones artificiels profonds sont généralement interprétées comme l'équivalent des fréquences d'impulsions des neurones biologiques, une autre famille

40. R. VAILLANT, C. MONROCQ et Y. LECUN, IEEE Proceedings of Vision Image and Signal Processing, 1994.
41. S. LAWRENCE et coll., IEEE Transactions on Neural Networks, 1997.
I. De l'anglais *Convolutional Neural Network*.
42. D. CIREŞAN et coll., Neural Networks, 2012.
43. J. TOMPSON et coll., Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015.
44. Y. TAIGMAN et coll., Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014.
45. D. H. HUBEL et T. N. WIESEL, The Journal of physiology, 1962.
46. D. J. FELLEMAN et D. C. VAN ESSEN, Cerebral cortex, 1991.

de réseaux de neurones artificiels, dits *impulsionnels*, s'est également développée en parallèle, davantage tournée vers le bio-réalisme ainsi que la basse consommation. Il est parfois question de 3^{e} génération de réseaux de neurones[47], après le perceptron originel et les réseaux à plusieurs couches.

Le double intérêt de l'approche impulsionnelle. Les réseaux de neurones artificiels impulsionnels présentent plusieurs avantages :

1. il est *énergétiquement plus efficace* d'utiliser des impulsions à la place de signaux électriques continus (ce qui constitue un argument non négligeable dans le cas de la biologie pour un organe pouvant consommer 20 % du budget énergétique pour seulement 2 % de la masse corporelle[12]) ;
2. un tel système est *plus robuste aux défaillances*[48] :
   - alors qu'en représentation binaire, une défaillance du bit de poids fort est bien plus problématique qu'une défaillance du bit de poids faible, les conséquences d'une défaillance d'un neurone dans le cadre impulsionnel dépendent peu du neurone en question ;
   - s'il est difficile de distinguer d'un signal non valide un signal émis par une porte logique conventionnelle bloquée dans un état donné, la défaillance d'un neurone impulsionnel est quant à elle silencieuse.

Quelques modèles de neurones impulsionnels. Il existe différents modèles pour décrire un comportement impulsionnel des neurones. Un exemple connu de modèle biologiquement plausible est le neurone de HODGKIN et HUXLEY[49], capable de reproduire un certain nombre de comportements rencontrés dans la nature. Ce modèle provient de la modélisation électrique de la membrane plasmique d'un neurone, qui résulte en un système de quatre équations différentielles couplées. La figure 1.7 présente un exemple de comportement de ce type de neurone.

D'autres modèles, plus simples à mettre en œuvre, sont généralement préférés dans les domaines des neurosciences computationnelles ou des systèmes neuromorphiques. Notamment le neurone dit « intègre-et-tire », qui s'inspire d'une idée ancienne[51] et connaît de nombreuses variantes, telles que le modèle « intègre-et-tire avec fuites[52] », dont un circuit électrique équivalent est représenté à la figure 1.8. La figure 1.9 présente un exemple de comportement d'un neurone « intègre-et-tire avec fuites »(LIF[I]). La fréquence des impulsions qu'émet un tel neurone dépend de l'amplitude du courant reçu en entrée, de sa constante de temps (représentée

47. W. MAASS, Neural networks, 1997.
12. D. ATTWELL et S. B. LAUGHLIN, Journal of Cerebral Blood Flow & Metabolism, 2001.
48. S. B. FURBER, IET Computers & Digital Techniques, 2016.
49. A. L HODGKIN et A. F. HUXLEY, The Journal of physiology, 1952.
51. L. LAPICQUE, J. Physiol. Pathol. Gen, 1907.
52. R. B. STEIN, Biophysical Journal, 1965.
I. De l'anglais *Leaky Integrate-and-Fire.*

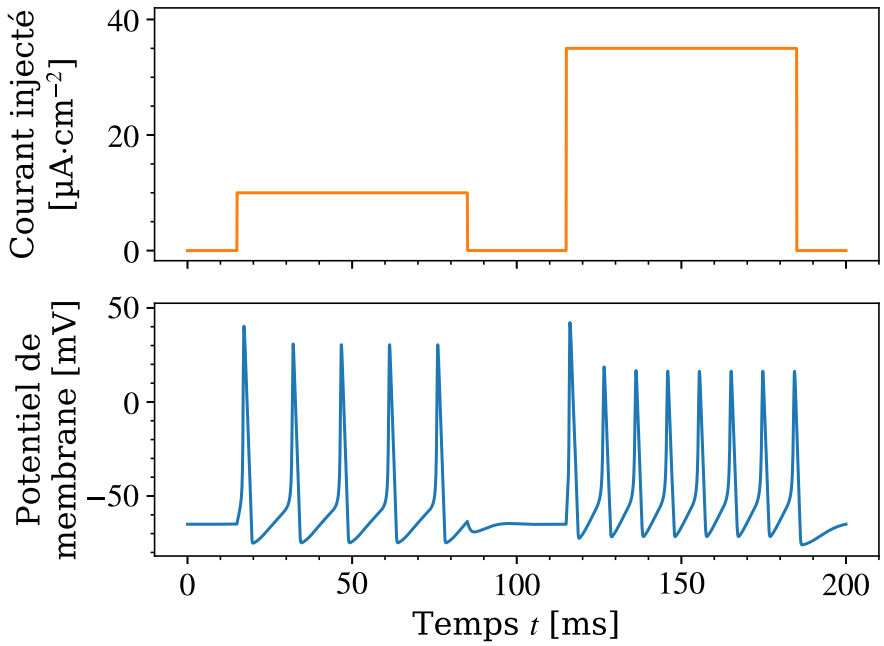


FIGURE 1.7 – *(b)* Évolution du potentiel de membrane d'un neurone de HODGKIN et HUXLEY lorsque ce dernier reçoit le courant *(a)*. La fréquence des impulsions dépend notamment de l'amplitude du courant injecté. *N.B.* : les valeurs numériques des paramètres de simulation sont tirées de la référence [50].

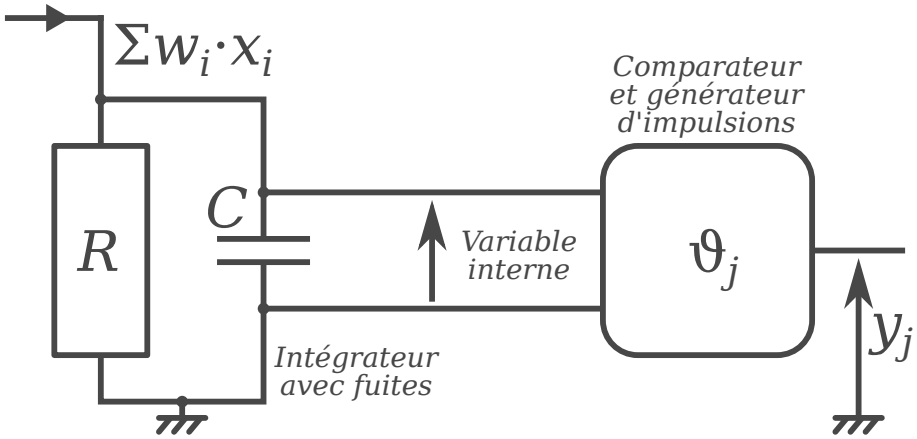


FIGURE 1.8 – Circuit électrique équivalent d'un neurone « intègre-et-tire avec fuites ». Le condensateur C assure le rôle d'intégrateur du courant image de la somme des entrées x_i pondérées par leur poids synaptique respectif w_i, tandis que la résistance R est à l'origine des fuites. La tension aux bornes de ces deux composants constitue la variable interne du neurone. Elle est utilisée comme entrée d'un comparateur de seuil ϑ_j et d'un circuit de génération d'impulsions qui impose la tension de sortie y_j.

par le produit RC dans la figure 1.8) et de sa période réfractaire, ainsi que de la valeur du seuil ϑ_j.

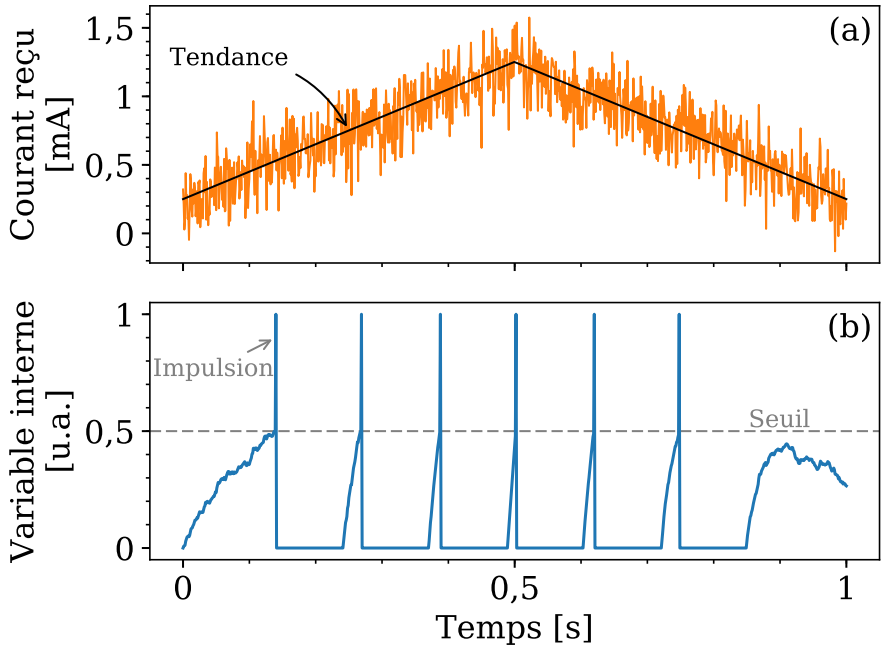


FIGURE 1.9 – *(b)* Évolution de la variable interne d'un neurone « intègre-et-tire avec fuites » (LIF) lorsque ce dernier reçoit le courant *(a)*, orienté algébriquement de façon identique à celle de la figure 1.8. Une impulsion est déclenchée lorsque la variable interne atteint le seuil ϑ_j (ici égal à 0,5) et le neurone entre alors dans un état réfractaire, durant lequel sa variable interne est maintenue à zéro. La durée nécessaire pour atteindre le seuil est d'autant plus courte que le courant est important. Par ailleurs, à la fin de la trace, le courant reçu ne permet plus de contrebalancer les fuites : même en présence d'un courant incident positif, la valeur de la variable interne continue de décroître.

Au-delà du codage en fréquence. Certaines expériences et théories récentes dans le domaine des réseaux de neurones impulsionnels suggèrent qu'il existe des stratégies de codage l'information ne reposant pas uniquement sur la fréquence des impulsions[53,54]. Ou tout du moins cette vision des choses ne permet-elle pas d'expliquer tous les phénomènes pertinents rencontrés en biologie. Le temps de réaction visuelle de l'homme à une scène complexe a par exemple été évalué à moins de 150 ms par S. THORPE et coll.[55]. Durant ce laps de temps, un neurone ayant une fréquence d'impulsions autour de la dizaine de hertz n'a guère le temps de se déclencher plus d'une fois : un codage de l'information dans la fréquence des impulsions semble peu plausible dans cette situation. Les propositions alternatives de codage de l'information

53. W. GERSTNER, R. RITZ et J. L. VAN HEMMEN, Biological cybernetics, 1993.
54. R. BRETTE, Frontiers in systems neuroscience, 2015.
55. S. THORPE, D. FIZE, C. MARLOT et coll., Nature, 1996.

dans le cadre des réseaux de neurones impulsionnels sont multiples, par exemple en utilisant l'ordre d'arrivée des impulsions[56], ou encore en exploitant le délai entre certaines impulsions[57].

Une règle d'apprentissage basée sur la distance temporelle entre impulsions. Suggérée par des expériences de biologie *in vitro* sur des neurones de rat[57], dont certains résultats sont reproduits sur la figure 1.10, la *Spike-Timing Dependent Plasticity* (STDP) est une famille de règles d'apprentissage qui exploite la *distance temporelle* entre une impulsion présynaptique (PRÉ) et une impulsion postsynaptique (POST). À la manière de la règle de Hebb évoquée précédemment, cette stratégie d'apprentissage vise à moduler les poids synaptiques lorsque qu'une impulsion postsynaptique est temporellement proche d'une impulsion présynaptique. Ainsi, dans le cas présenté à la figure 1.10, une connexion synaptique participant :

- à une suite causale d'impulsions (l'impulsion présynaptique précède l'impulsion postsynaptique) est potentialisée (sa conductance augmente) ;
- à une suite anti-causale d'impulsions (l'impulsion *postsynaptique* précède l'impulsion présynaptique) est déprimée (sa conductance diminue).

Par ailleurs, les changements relatifs de conductance associés sont d'autant plus conséquents que la distance temporelle est faible.

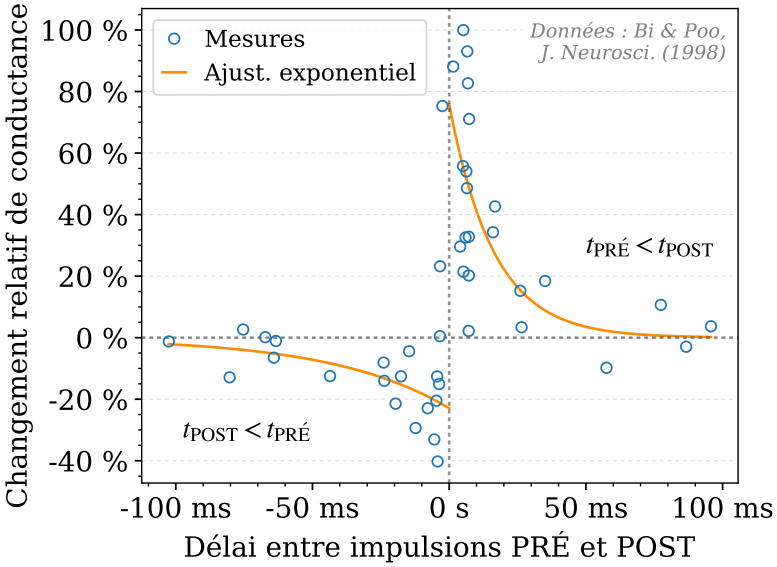


FIGURE 1.10 – Observation de la *Spike-Timing Dependent Plasticity* en biologie : changement relatif de la conductance synaptique selon le délai $t_{POST} - t_{PRÉ}$ entre un potentiel d'action présynaptique et un potentiel d'action postsynaptique. Les symboles sont des mesures expérimentales tirées de la référence [57], tandis que les lignes en trait plein sont des fonctions exponentielles ajustées sur ces données.

56. S. THORPE, A. DELORME et R. VAN RULLEN, Neural networks, 2001.
57. G.-Q. BI et M.-M. POO, The Journal of neuroscience, 1998.

Les règles d'apprentissage de la famille *Spike-Timing Dependent Plasticity* sont de type non supervisé, ce qui les rend intéressantes pour développer des systèmes capables de classifier des données non étiquetées. Certains travaux suggèrent que ces règles d'apprentissage pourraient expliquer l'émergence de cartes de fonction dans un système visuel biologique, par simple exposition à un flux d'images naturelles[58].

Des capteurs bio-inspirés comme sources d'impulsions. L'émergence de capteurs bio-inspirés apporte un intérêt supplémentaire aux réseaux de neurones impulsionnels. En effet, la plupart de ces capteurs encodent directement l'information sous la forme de trains d'impulsions asynchrones[59], ce qui en fait des sources naturelles de signaux d'entrée pour les réseaux impulsionnels. Ils permettent en outre d'atteindre une efficacité énergétique importante[60], ce qui les rend d'autant plus attractifs pour l'électronique embarquée et la basse consommation.

Remarque

Le protocole employé par ces capteurs neuromorphiques est très souvent du type AER [a] tel que proposé dans la référence [61]. Il s'agit d'une représentation numérique, constituée d'une série de couples formés de l'adresse d'un neurone ayant émis une impulsion et de l'instant de cette impulsion.

a. De l'anglais *Address-Event Representation.*

Les exemples les plus probants de capteurs neuromorphiques sont certainement les rétines artificielles de type *Dynamic Vision Sensor*[62] (DVS), qui émulent le fonctionnement d'une rétine biologique. Chacun des éléments de la matrice de pixels d'une telle rétine fonctionne de façon asynchrone, indépendamment de ses voisins. Ces pixels sont sensibles aux *variations* de l'intensité lumineuse qu'ils reçoivent. Ils signalent par l'émission d'une impulsion une variation d'amplitude supérieure à un seuil fixé. Cette impulsion encode l'information sur seulement un bit et chaque pixel dispose en réalité de deux adresses, pour lui permettre de différencier une variation positive d'une variation négative. Ce type de capteurs présente deux grands avantages en comparaison des capteurs vidéo conventionnels exploitant une succession d'images *complètes* à fréquence *fixe* :

1. ils produisent des données plus parcimonieuses (seules les variations sont transmises), ce qui nécessite moins de puissance de calcul pour le traitement de ces dernières et permet de réduire la consommation énergétique[63] ;
2. leur caractère asynchrone permet d'obtenir une résolution temporelle plus courte (sous la milliseconde voire moins), notamment du fait d'une bande-passante moins sollicitée.

58. T. MASQUELIER et S. J. THORPE, PLoS Computational Biology, 2007.
59. A. VANARSE, A. OSSEIRAN et A. RASSAU, Frontiers in Neuroscience, 2016.
60. J. HASLER et H. MARR, Frontiers in Neuroscience, 2013.
62. P. LICHTSTEINER, C. POSCH et T. DELBRÜCK, IEEE Journal of Solid-State Circuits, 2008.
63. C. POSCH et coll., Proceedings of the IEEE, 2014.

Par exemple, le capteur DVS128 de la référence [62] présente une dynamique de 120 dB et une résolution temporelle de moins d'une milliseconde, pour une consommation énergétique de seulement 23 mW. Il est notamment utilisé en robotique[64] ainsi qu'en imagerie haute fréquence[65]. L'engouement autour des rétines artificielles est à l'origine de travaux visant à encore en améliorer les caractéristiques ou à les doter de nouvelles fonctionnalités[64,66]. L'une des principales limitations des réalisations existantes est leur résolution limitée (typiquement de 128×128 à 320×240 pixels), qui a pour origine l'utilisation de nœuds technologiques assez anciens (180 nm voire 350 nm).

Dans le domaine de l'acoustique, la cochlée artificielle AEREAR2[67] est un autre exemple de capteur neuromorphique prometteur. Son fonctionnement est inspiré de celui d'une cochlée biologique, modélisée par une cascade de filtres passe-bandes émettant des impulsions. Une nouvelle fois, la représentation parcimonieuse sous forme d'impulsions asynchrones permet de réduire la consommation ainsi que la puissance de calcul nécessaire à l'interprétation des signaux de sortie. Ainsi, le système AEREAR2 permet une audition binaurale sur (2×)64 voies pour une consommation de 52 mW et nécessite environ 40 fois moins de puissance de calcul pour effectuer une localisation de source par rapport aux algorithmes classiques par corrélation croisée[67]. Ce capteur a été utilisé pour identifier des orateurs[68] ou reconnaître des chiffres énoncés isolément[69].

Les capteurs olfactifs complètent la liste des capteurs neuromorphiques connaissant aujourd'hui un certain succès, en répondant au besoin de nez électroniques portatifs, fiable et à des prix abordables[70], notamment pour l'agriculture. Il a été démontré qu'utiliser une approche impulsionnelle couplée à un algorithme d'apprentissage de type *Spike-Timing Dependent Plasticity* était pertinent pour réduire la consommation énergétique, même en se contentant d'utiliser un nez électronique conventionnel[71]. Des exemples de réalisation de capteurs olfactifs impulsionnels[72] ont par ailleurs permis de constater que ces derniers produisent des signaux dont les signatures sont plus simples à identifier, donnant lieu à des taux de reconnaissance de presque 95 % pour une consommation de moins 7 mW.

Un argument opposable aux capteurs impulsionnels est qu'ils ne permettent pas de réutiliser directement les algorithmes de traitement des données conventionnels et bien établis : il est généralement nécessaire d'adapter ces algorithmes à la nature impulsionnelle des signaux. Toutefois, ce problème ne se pose plus lorsque ces capteurs servent à alimenter des réseaux de neurones impulsionnels, à condition de disposer d'une règle d'apprentissage adaptée à cette nature impulsionnelle. Par ailleurs, un avantage intéressant du fonctionnement impulsionnel

64. C. BRANDLI et coll., Frontiers in Neuroscience, 2014.
65. D. DRAZEN et coll., Experiments in Fluids, 2011.
66. T. SERRANO-GOTARREDONA et B. LINARES-BARRANCO, IEEE Journal of Solid-State Circuits, 2013.
67. S.-C. LIU et coll., IEEE Transactions on Biomedical Circuits and Systems, 2014.
68. C.-H. LI, T. DELBRÜCK et S.-C. LIU, IEEE International Symposium on Circuits and Systems, 2012.
69. M. ABDOLLAHI et S.-C. LIU, IEEE Biomedical Circuits and Systems Conference, 2011.
70. S.-W. CHIU et K.-T. TANG, Sensors, 2013.
71. H. Y. HSIEH et K. T. TANG, IEEE Transactions on Neural Networks and Learning Systems, 2012.
72. K. T. NG et coll., Sensors, 2009.

commun à ces différents capteurs est qu'il facilite la fusion multi-sensorielle, dont un exemple est l'exploitation conjointe de signaux visuels et auditifs pour faciliter la prise de décision avec des signaux d'entrée ambigus, bruités ou incomplets[73,74], ce qui demeure encore aujourd'hui un défi dans le domaine de la robotique.

1.2 Vers des réseaux de neurones artificiels plus efficients

Qu'il s'agisse des réseaux de neurones artificiels profonds ou des réseaux impulsionnels, l'un des plus gros obstacles à leur utilisation à grande échelle est le goulot d'étranglement de VON NEUMANN, présenté à la figure 1.6. Le problème résulte de la mauvaise adéquation entre l'architecture des ordinateurs usuels et le principe de fonctionnement des réseaux de neurones. Cette partie présente des pistes récentes de recherches conceptuelles pour pallier cette difficulté et permettre le déploiement des réseaux de neurones artificiels à grande échelle ou dans le domaine de l'électronique embarquée.

1.2.1 Adapter l'architecture matérielle aux opérations critiques

Une première stratégie d'amélioration possible est d'identifier les opérations critiques des réseaux de neurones, en termes de vitesse d'exécution et de consommation énergétique, et de concevoir des circuits et des architectures plus efficaces pour les réaliser.

Un exemple d'opération massivement exploitée par les réseaux de neurones est l'instruction « multiplie-et-accumule » (MAC[I])

$$a \leftarrow a + b \times c \tag{1.4}$$

qui est utilisée lors de l'inférence (basée sur l'équation (1.1)) pour la pondération par les poids synaptiques, mais également lors de la mise à jour de ces derniers dans le cas de certaines règles d'apprentissage très employées, telle que l'algorithme de rétropropagation du gradient (dérivé de la règle du perceptron, basé sur l'équation (1.3)).

Approche conventionnelle : le recours à un circuit intégré développé pour un client. Des travaux récemment publiés par SEO et coll. montrent qu'un circuit intégré développé pour un client (ASIC) permet des gains substantiels de vitesse d'exécution et de consommation énergétique dans le cadre de l'apprentissage de dictionnaires pour les codes parcimonieux[II], dont l'algorithme employé est proche de la rétropropagation du gradient. L'étude en simulation est basée sur un nœud technologique 65 nm. L'architecture exploite des blocs de mémoire vive statique (SRAM[III]) de taille réduite, placés au plus proche d'unités de calcul numériques. Ceci

73. V. CHAN, C. JIN et A. van SCHAIK, Frontiers in Neuroscience, 2012.
74. P. O'CONNOR et coll., Frontiers in Neuroscience, 2013.

I. *Multiply-Accumulate* en anglais.

II. Remarquons au passage que cette tâche peut également être mise en lien avec la biologie, comme l'explique par exemple la référence [75].

III. De l'anglais *Static Random Access Memory*.

permet de limiter les accès séquentiels à la mémoire (contrairement au cas d'une mémoire centrale), ainsi que de réaliser un certain nombre d'opérations en parallèle, ligne par ligne pour reprendre une vision matricielle. Sur le test de performance retenu par les auteurs, un tel système est presque 400 fois plus rapide qu'un processeur généraliste actuel, tout en affichant une réduction de sa consommation énergétique frôlant un facteur 7000.

L'un des points faibles de ce type d'approches est le faible potentiel de passage à l'échelle, lié à l'utilisation de mémoire vive statique. Comme nous l'avons observé précédemment, les réseaux de neurones artificiels nécessitent des quantités de mémoire importantes (notamment pour décrire leurs paramètres synaptiques) ce qui s'accorde mal avec l'utilisation d'une mémoire aussi coûteuse en surface de silicium que la mémoire vive statique.

Vers l'utilisation de nouveaux substrats matériels. Une piste prometteuse pour surmonter le problème de la densité d'intégration limitée de la mémoire vive statique est de se tourner vers les nouvelles mémoires résistives, dont le développement a connu une accélération au cours de la dernière décennie[76]. Ces nouveaux nanocomposants mémoire se répartissent en plusieurs familles, dont les principales seront présentées dans la partie 1.4. Ils offrent généralement un potentiel d'intégration bien supérieur à la mémoire vive statique, avec une taille de cellule mémoire pouvant descendre à 4 F^2 pour certaines[I], contre typiquement plus de 100 F^2 pour la mémoire vive statique. Ils sont donc particulièrement intéressants pour stocker les paramètres synaptiques dans des architectures spécialisées dans les réseaux de neurones artificiels, à condition de pouvoir être intégrés directement au cœur des ces dernières, par exemple au niveau des métallisations finales (BEOL[II]).

Utiliser une matrice de nanocomposants résistifs en lieu et place d'une matrice de cellules de mémoire vive statique permet par ailleurs de paralléliser encore davantage les opérations « multiplie-et-accumule ». En reprenant les notations de l'équation (1.1) page 11, si les poids synaptiques w_i sont représentés par la conductance électrique de ces composants et que les entrées x_i sont présentées sous la forme de tensions électriques entre les neurones d'entrée et chacun des neurones de sortie, alors la loi d'OHM assure les multiplications $w_i x_i$, tandis que la loi des nœuds permet d'obtenir leur somme sous la forme d'un courant électrique (figure 1.11).

Plusieurs travaux de simulation focalisés sur l'algorithme de rétropropagation du gradient prédisent des gains très significatifs, en termes de vitesse d'exécution et de consommation énergétique, en comparaison des processeurs généralistes mais également des circuits spécialisés uniquement CMOS[III]. Les travaux de SOUDRY et coll.[77] présentent une telle puce capable d'apprendre en ligne, avec des performances similaires à des systèmes purement logiciels, et qui requiert seulement entre 2 et 8 % de la surface et de la puissance statique des solutions matérielles entièrement à base de circuits CMOS disponibles dans la littérature. En sus du

76. H.-S. P. WONG et S. SALAHUDDIN, Nature Nanotechnology, 2015.
I. F désigne la dimension caractéristique du nœud technologique.
II. De l'anglais *Back-End-Of-Line*.
III. De l'anglais *Complementary Metal-Oxide-Semiconductor*.
77. D. SOUDRY et coll., IEEE Transactions on Neural Networks and Learning Systems, 2015.

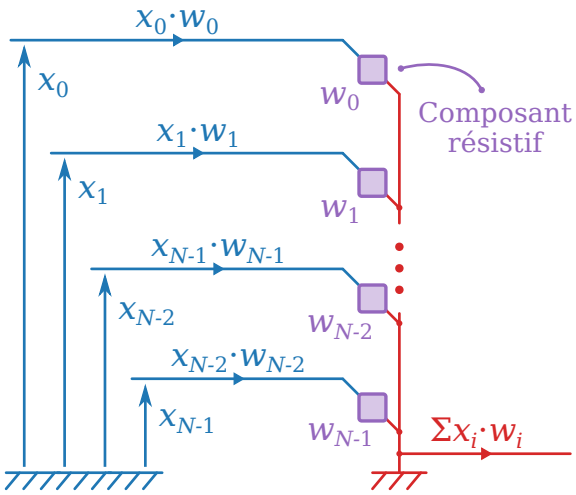


FIGURE 1.11 – Principe du calcul analogique d'un produit scalaire $\boldsymbol{x} \cdot \boldsymbol{w} = \sum_{i=0}^{N-1} x_i w_i$ en utilisant la loi d'OHM et la loi des nœuds. Le vecteur $\boldsymbol{x}$ est présenté sous la forme des tensions électriques aux bornes d'éléments résistifs, dont les valeurs de conductance représentent les composantes du vecteur $\boldsymbol{w}$. La valeur de sortie est représentée par la somme des courants circulant à travers les composants résistifs.

circuit intégré développé pour un client déjà présenté, SEO et coll.[78] propose également une architecture mêlant des unités de calcul CMOS 65 nm et des matrices de synapses réalisées avec des nanocomposants mémoires résistifs (de caractéristiques rapportées dans la littérature), qui améliore encore la vitesse d'exécution d'un facteur 23 et la consommation énergétique d'un facteur 5,4 par rapport à leur circuit intégré développé pour un client. Enfin, un dernier exemple est celui d'une étude de GOKMEN et coll.[79], visant à évaluer les caractéristiques des cellules mémoires non volatiles génériques requises pour fabriquer une puce accélératrice de l'algorithme de rétropopagation du gradient. Ces travaux couplent l'utilisation d'une matrice de nanocomposants mémoires résistifs, toujours pour paralléliser les opérations vectorielles et matricielles, à du calcul stochastique[I]. Ce dernier point permet de simplifier l'exécution de certaines multiplications, en les remplaçant par des opérations logiques bit à bit. Les résultats espérés à l'issue de ces simulations sont une accélération par rapport à une processeur haut de gamme actuel d'un facteur supérieur à 30 000 pour une consommation énergétique similaire (250 W). Cette solution repose néanmoins sur l'utilisation (partielle) d'un processeur graphique actuel performant pour contrôler la puce accélératrice.

Un constat qui ressort de ce type d'études est la bonne tolérance de telles architectures à la variabilité entre composants synaptiques ou bien au bruit pouvant affecter les signaux[77,79]. La raison à cela est qu'un réseau de neurones artificiels apprend *progressivement* sa tâche, en

78. J.-S. SEO et coll., IEEE Transactions on Nanotechnology, 2015.
79. T. GOKMEN et Y. VLASOV, Frontiers in Neuroscience, 2016.

I. Attention, le terme « calcul stochastique » désigne ici le fait de d'effectuer des opérations de calcul sur des signaux aléatoires codant l'information sous la forme d'une probabilité (typiquement un train de bits). Il ne s'agit pas de la branche des mathématiques portant sur l'étude des phénomènes aléatoires dépendant du temps.

s'adaptant aux ressources dont il dispose. Par exemple, à la manière d'un réseau de neurones biologiques, son organisation s'élabore en tenant compte (dans la mesure du possible) des composants éventuellement défectueux.

1.2.2 Réduire la précision utilisée pour décrire les poids synaptiques

L'inférence nécessite-t-elle absolument des variables de grande précision ? Une observation commune à plusieurs travaux récents est qu'une grande précision numérique sur les poids synaptiques n'est pas nécessairement requise lors de la phase d'inférence d'un réseau de neurones artificiels. Ainsi, le fait d'utiliser une représentation à virgule fixe sur 8 bits[80], voire ternaire[81] ou tout simplement binaire[82–84] pour décrire les poids synaptiques lors de la phase d'inférence ne dégrade pas significativement les performances en comparaison de systèmes équivalents exploitant une représentation à virgule flottante sur 32 ou 64 bits. Les avantages de ces approches sont multiples :

1. une forte réduction de l'empreinte mémoire [I], ce qui permet d'envisager déployer des réseaux de neurones artificiels réalistes sur des cibles embarquées[84] ;
2. un corollaire du premier point est la réduction de la bande-passante ainsi que de la consommation énergétique des accès à la mémoire[85] ;
3. il devient possible de remplacer certaines opérations arithmétiques coûteuses (notamment une partie des multiplications) par de simples décalages de registres. Dans le cas où à la fois les neurones et les synapses sont binaires, la possibilité est même offerte d'implémenter les opérations de convolution presque entièrement à l'aide d'opérations bit à bit de type « OU exclusif »[82,84], ce qui se traduit par une accélération sensible de l'inférence [II].

Dans le domaine industriel, accélérer les opérations arithmétiques sur des variables de précision limitée (notamment les entiers 8 bits) semble à ce jour être la stratégie retenue par plusieurs grands acteurs, qu'il s'agisse de processeurs graphiques [III] ou de circuits intégrés développés pour un client [IV], pour rendre l'étape d'inférence d'un réseau de neurones artificiels plus efficace.

80. P. GYSEL, M. MOTAMEDI et S. GHIASI, arXiv preprint arXiv :1604.03168, 2016.
81. Z. LIN et coll., arXiv preprint arXiv :1510.03009, 2015.
82. M. KIM et P. SMARAGDIS, arXiv preprint arXiv :1601.06071, 2016.
83. M. COURBARIAUX et coll., arXiv preprint arXiv : 1602.02830 v3, 2016.
84. M. RASTEGARI et coll., arXiv preprint arXiv :1603.05279, 2016.

I. Un exemple probant est le réseau de neurones artificiels AlexNet de la référence [84], qui passe d'une empreinte mémoire de 475 Mio avec une description des synapses en double précision à seulement 7,4 Mio dans le cas de synapses binaires.

85. M. HOROWITZ, IEEE International Solid-State Circuits Conference Digest of Technical Papers, 2014.

II. Les auteurs de la référence [84] rapportent notamment une accélération d'un facteur 58.

III. http://nvidianews.nvidia.com/news/new-nvidia-pascal-gpus-accelerate-deep-learning-inference

IV. http://www.forbes.com/sites/tiriasresearch/2016/05/26/google-builds-its-first-chip-just-for-machine-learning/

Quid de l'apprentissage ? La question se pose de savoir s'il est également envisageable de réduire la précision des poids synaptiques lors de la phase d'apprentissage d'un réseau de neurones artificiels sans dégrader les performances de ce dernier. Une première stratégie est de réduire la précision des poids synaptiques (possiblement jusqu'à les rendre binaires) *à l'issue* d'un classique apprentissage hors ligne, par exemple selon l'algorithme conventionnel de rétropropagation du gradient[80,82]. Cette étape peut être suivie d'une procédure non triviale d'apprentissage fin, telle qu'une version bruitée de l'algorithme de rétropropagation du gradient[82], dans le but d'améliorer légèrement les performances du système. Une seconde méthode est de rendre binaires ou ternaires les poids synaptiques dès l'apprentissage, de façon stochastique[81] ou non[83,84].

Si elle ne dégradent pas significativement les performances des réseaux de neurones artificiels sur lesquels elles sont appliquées tout en réduisant l'empreinte mémoire et la complexité de certaines opérations arithmétiques, ces différentes approches présentent néanmoins plusieurs inconvénients :

- dans tous les cas, l'apprentissage est hors-ligne et nécessite de conserver en mémoire la valeur pleine précision des paramètres synaptiques, même lorsque la projection de ces derniers vers une résolution plus faible est réalisée durant la phase d'apprentissage ;
- si le fait de rendre binaire ou ternaire des variables dès l'apprentissage peut accélérer la phase d'inférence, cela tend également à ralentir la convergence de l'apprentissage[83] ;
- les variantes stochastiques nécessitent, pour être intéressantes, d'avoir des générateurs de nombres aléatoires efficaces à disposition.

1.2.3 Adopter une approche impulsionnelle

Puisque la réduction de la consommation énergétique et le bio-mimétisme sont des axes porteurs de la recherche en architecture matérielle des réseaux de neurones artificiels[48], il semble naturel de chercher à réaliser des systèmes orientés vers les réseaux impulsionnels. Il s'agit de la stratégie adoptée par l'ensemble des réalisations matérielles que nous présentons dans la partie 1.3. Pour deux de ces systèmes, il est notamment possible de trouver dans la sous-partie 1.3.4 les valeurs de consommation particulièrement encourageantes rapportées dans la littérature dans le cadre d'une tâche pratique de classification d'images[86].

1.2.4 Le bruit et les non linéarités peuvent-ils être bénéfiques ?

Dans les implémentations logicielles utilisant des paramètres synaptiques de précision réduite, il a été récemment observé que les performances peuvent être améliorées, en comparaison d'un apprentissage totalement déterministe, lorsque du bruit est ajouté durant la phase de projection des poids synaptiques vers une précision plus faible[81,82]. MEROLLA et coll. ont par ailleurs montré que faire subir dès la phase d'apprentissage des distorsions non linéaires

86. S. K. ESSER et coll., Advances in Neural Information Processing Systems, 2015.

aux poids synaptiques d'un réseau de neurones artificiels profond pouvait le rendre également plus performant[24]. Ces conditions d'apprentissage améliorent les capacités de généralisation des réseaux de neurones artificiels et permettent d'éviter l'écueil du surapprentissage.

Le fait que les réseaux de neurones puissent être tolérants à des poids synaptiques de résolution et dynamique limitées, ainsi que peu sensibles à des fluctuations de ces derniers, est une caractéristique intéressante. Elle permet d'envisager l'amélioration de l'efficience des réseaux de neurones artificiels en ayant recours à des nanocomposants mémoires innovants pour stocker les paramètres synaptiques, même si ces nouvelles technologies mémoires souffrent généralement d'une variabilité entre composants encore significative.

1.2.5 Intérêt d'une programmation synaptique probabiliste

Inclure une part d'aléatoire lors de la phase d'apprentissage d'un réseau de neurones artificiels utilisant des poids synaptiques de précision réduite, comme le proposent certains des travaux susmentionnés[24,81,82], n'est pas sans rappeler les études théoriques de SENN et coll.[87]. Ces derniers démontrent la capacité d'apprentissage d'un réseau de neurones artificiels impulsionnels et de synapses binaires, en exploitant une version probabiliste de l'algorithme du perceptron. Malgré une vitesse de convergence diminuée dans le cas étudié, un apprentissage est bel et bien possible tant que les probabilités de transition des synapses binaires sont suffisamment faibles.

En s'inspirant de ces observations, et en faisant le constat qu'exploiter certains nanocomposants mémoires innovants de façon binaire plutôt que multivaluée permet de simplifier l'architecture électronique, SURI et coll. ont proposé d'utiliser ces derniers dans une architecture neuromorphique impulsionnelle, comme synapses binaires programmées selon une règle d'apprentissage *non supervisé* probabiliste[88]. Une première approche consiste à utiliser ces composants mémoires de façon déterministe et d'avoir recours à un générateur de nombre aléatoires extrinsèque pour mettre en place l'algorithme probabiliste ; les performances obtenues en simulation sont similaires à celles d'une solution purement logicielle. Une seconde approche proposée par les auteurs est d'exploiter directement les comportements stochastiques *intrinsèques* aux nanocomposants mémoires retenus, par exemple en se plaçant dans des conditions de programmation faible. Ceci permet de réduire la consommation énergétique des phases de programmation ainsi que l'aire des circuits associés, sans affecter les performances du système à l'issue de l'apprentissage. Une stratégie analogue d'apprentissage non supervisé basé sur une autre famille de nanocomposants mémoires filamentaires a également été testée avec succès en simultation par YU et coll.[89] Si cette deuxième approche permet de s'affranchir d'un générateur de nombres aléatoires externe, elle demeure difficilement envisageable dans la pratique avec de tels nanocomposants mémoires filamentaires, en raison du mauvais contrôle de la dispersion

87. W. SENN et S. FUSI, Physical Review E, 2005.
88. M. SURI et coll., IEEE Transactions on Electron Devices, 2013.
89. S. YU et coll., Frontiers in Neuroscience, 2013.

de leur caractéristiques lors de la fabrication.

Dans le cadre du projet d'architecture matérielle neuro-inspirée présentée à la partie 1.3.4, ESSER et coll. ont également présenté et mis en œuvre une stratégie de programmation synaptique stochastique[86]. L'algorithme de rétropropagation du gradient employé est interprété de façon probabiliste, afin de pallier le caractère binaire des poids synaptiques de l'architecture : les valeurs des entrées et des sorties représentent des probabilités de tir, et celles des synapses des probabilités d'existence d'une connexion. Comme dans la grande majorité des architectures neuro-inspirées matérielles à impulsions, l'apprentissage n'est pas réalisé directement sur puce. Néanmoins, contrairement aux approches de la partie 1.2.2, les possibilités de connexions synaptiques autorisées sont restreintes dès l'apprentissage afin de correspondre à celles de la future cible matérielle, sans que cela nuise aux performances d'inférence. Cette stratégie peut donc aider à simplifier la procédure d'apprentissage sur une architecture matérielle de faible précision synaptique. Ce dernier demeure toutefois hors-ligne et nécessite de disposer de générateurs efficaces de nombres aléatoire.

À propos de l'apprentissage probabiliste

Une (piste d') explication intuitive quant à la pertinence d'un apprentissage probabiliste dans le cas de poids synaptiques binaires est la suivante. Avec des synapses usuelles de grande résolution, à l'issue d'un pas d'apprentissage, un certain nombre d'entre elles ont *légèrement* évolué. Une variation égale à la somme de ces petites évolutions peut être obtenue avec des synapses binaires, si seule une certaine *proportion* de celles devant commuter commute effectivement. Une façon simple de mettre en place cela est de soumettre les synapses binaires concernées à une programmation probabiliste.

1.3 Les réalisations matérielles existantes

Depuis quelques années, la recherche sur les architectures matérielles de systèmes de calcul neuro-inspirés a fait l'objet d'un intérêt grandissant, menant à de multiples réalisations matérielles. Ces dernières peuvent être classées selon deux axes principaux :

- une direction pleinement « neuromorphique », visant à concevoir des systèmes matériels plus efficaces que les réseaux de neurones artificiels logiciels pour résoudre des tâches cognitives, aussi bien en termes de consommation énergétique que de puissance de traitement ;
- une direction davantage « neuromimétique », plus tournée vers la réalisation de systèmes matériels permettant d'émuler de façon biologiquement plausible des réseaux de neurones.

Il est à noter que ces deux approches ne sont pas orthogonales et peuvent s'influencer l'une l'autre.

Cette partie dresse un panorama des principales solutions matérielles de calcul neuro-inspiré existant aujourd'hui. Ces dernières couvrent un large spectre technologique, allant des circuits logiques programmables (FPGA[I]) aux circuits intégrés développés pour un client (ASIC) numériques, en passant par des grappes de systèmes sur puce (SoC[II]) ou encore un mélange de circuits numériques et analogiques (signaux mixtes). En revanche, l'ensemble de ces systèmes matériels vise à mettre en œuvre des réseaux neuronaux de type impulsionnel, pour des raisons de consommation énergétique mais également de simplification des communications entre les différents éléments, voire simplement dans le cas des réalisations « neuromimétiques » pour se rapprocher des comportements biologiques.

1.3.1 Solutions à base de circuits logiques programmables

Une première solution envisageable pour accélérer les réseaux de neurones artificiels est l'utilisation de circuits logiques programmables, dont la granularité fine des unités de calcul se prête bien aux tâches parallèles. Entièrement numériques, ces plateformes offrent les avantages d'être plus versatiles que les circuits intégrés développés pour un client, tant au niveau de la représentation arithmétique et de la précision des variables que des modèles neuronaux et synaptiques, tout en améliorant l'efficacité énergétique en comparaison des solutions exploitant des processeurs graphiques. Par ailleurs, utilisant des systèmes commerciaux de série, ces solutions sont sensiblement moins coûteuses à développer que des circuits intégrés développés pour un client, puisqu'une partie des coûts de conception est amortie sur un marché plus large.

1.3.1.1 Bluehive

Le système Bluehive[90] de l'université de Cambridge exploite un ensemble de 64 circuits logiques programmables Altera® Stratix IV, réalisés dans une technologie 40 nm. Ces puces intègrent des bus série haute vitesse ce qui les rend bien adaptées aux tâches exigeant une forte bande-pasante et une faible latence.

Les auteurs présentent une démonstration exploitant des neurones impulsionnels d'IZHIKEVICH[91], qui sont un bon compromis entre la plausibilité biologique et l'économie des ressources de calcul[92]. La simulation est effectuée à temps discret, avec un pas de temps de 1 ms, et en virgule fixe sur 16 bits. Un multiplexage temporel permet d'évaluer durant un pas de temps jusqu'à 100 000 neurones avec un seul pipeline d'exécution cadencé à 200 MHz.

L'architecture présente une topologie en maillage torique, un nœud pouvant compter jusqu'à 100 000 neurones, chacun connecté en moyenne à 1000 synapses. Ces dernières sont codées

I. De l'anglais *Field-Programmable Gate Array*.
II. De l'anglais *System on Chip*.
90. S. W. MOORE et coll., IEEE Annual International Symposium on Field-Programmable Custom Computing Machines, 2012.
91. E. M. IZHIKEVICH, IEEE Transactions on Neural Networks, 2003.
92. E. M. IZHIKEVICH, IEEE Transactions on Neural Networks, 2004.

sur 12 bits et sont assorties d'un délai axonal [I] sur 4 bits. La localité des connexions est exploitée pour répartir physiquement les neurones sur les différents réseaux programmables.

Les paramètres, en particulier synaptiques, sont stockés dans une mémoire externe et non dans un bloc mémoire réalisé au sein du réseau programmable. Si l'efficacité énergétique du système s'en trouve réduite, cette stratégie permet de modifier la liste d'interconnexions sans avoir à recompiler l'ensemble de l'architecture, un processus pouvant nécessiter plusieurs heures. Un intérêt supplémentaire à l'utilisation d'une telle banque de mémoire externe est de pouvoir utiliser cette dernière à la manière d'un routeur virtuel, réduisant ainsi le nombre de connexions physiques.

Sur un test d'envergure réduite, avec 2^{16} neurones répartis sur un ensemble de 4 circuits logiques programmables, les auteurs ont démontré que ce système est un à deux ordres de grandeur plus rapide qu'un processeur généraliste actuel, permettant d'atteindre le temps réel. Par ailleurs, du fait de l'utilisation limitée de la bande-passante disponible entre les circuits logiques programmables, ils prédisent un passage à l'échelle linéaire jusqu'à aux moins 64 de ces réseaux, permettant donc d'envisager la simulation d'environ 4,2 millions de neurones et 1000 fois plus de synapses.

1.3.1.2 NeuroFlow

NeuroFlow[93], de l'*Imperial College* de Londres, est une autre plateforme matérielle d'exécution de réseaux de neurones impulsionnels qui exploite des circuits logiques programmables. Une interface de programmation basée sur le langage de haut niveau PyNN[94], courant dans la communauté des neurosciences computationnelles, accompagne cette plateforme afin de la rendre accessible à des utilisateurs novices en langage de description matérielle.

Plusieurs modèles de neurones impulsionnels sont disponibles : « intègre-et-tire » avec fuites ou adaptatif exponentiel[95], d'IZHIKEVICH ou encore de HODGKIN et HUXLEY. La résolution des équations différentielles est compilée de façon matérielle en pipeline d'exécution. Ceci est plus efficace en termes d'utilisation des ressources logiques qu'une approche reposant sur l'émulation logicielle d'un processeur. Néanmoins, une modification de ces équations différentielles nécessite de recompiler l'ensemble de l'architecture, un processus dont la durée est de 10 à 20 h. L'arithmétique employée est à virgule flottante et la résolution est effectuée par pas de temps de 1 ms

De façon similaire au système Bluehive, les paramètres du réseau de neurones sont stockés dans une mémoire externe, ce qui permet de reconfigurer le réseau neuronal sans avoir à recompiler l'architecture et évite les connexions synaptiques physiques, rendant plus aisé le passage à l'échelle, au détriment de la consommation énergétique. Les poids synaptiques sont codés sur 16 bits et assortis d'un délai axonal sur 4 bits. Afin d'accélérer l'exécution, l'accumulation

I. Il s'agit du délai nécessaire à une impulsion présynaptique pour se propager jusqu'au neurone postsynaptique.

93. K. CHEUNG, S. R. SCHULTZ et W. LUK, Frontiers in Neuroscience, 2016.
94. A. DAVISON et coll., Frontiers in Neuroinformatics, 2009.
95. R. BRETTE et W. GERSTNER, Journal of Neurophysiology, 2005.

synaptique est effectuée en arithmétique à virgule fixe et les opérations mémoires associées ont lieu dans une banque implémentée directement sur le circuit logique programmable. La règle d'apprentissage employée est une variante simplifiée de la *Spike-Timing Dependent Plasticity* présentée à la figure 1.10, de type délai vers le plus proche voisin et centré sur les impulsion présynaptiques. Des variantes plus complexes sont envisageable, à condition d'y consacrer davantage de ressources matérielles. Les délais prennent des valeurs discrètes, rendant possible l'utilisation de tables de correspondance (LUT[I]). Par ailleurs, les mises à jour des poids synaptiques sont traitées par lots, afin de limiter le nombre d'accès mémoires.

Les auteurs ont mesuré la vitesse d'exécution sur un système NeuroFlow constitué d'un circuit logique programmable Xilinx® Virtex 6, dans une technologie 40 nm. Le réseau testé est constitué de 55 000 neurones d'IZHIKEVICH dont les poids synaptiques sont mis à jour toutes les millisecondes et s'exécute environ 2,8 fois plus vite que sur un processeur graphique Nvidia® Tesla C1060 et 33 fois plus rapidement que sur un processeur multicœur généraliste Intel® i7-920. Ces facteurs d'accélération sont toutefois à relativiser : le processeur graphique comme le processeur généraliste ne sont pas des modèles de générations actuelles. Néanmoins, NeuroFlow conserve un avantage certain sur le plan énergétique, consommant moins de 40 W, contre plus d'une centaine de watts pour les autres plateformes (exigeant en outre une durée d'exécution supérieure). Par ailleurs, il est possible de passer le système NeuroFlow à l'échelle : les auteurs rapportent avoir réussi à simuler un réseau similaire de 400 000 neurones en temps réel sur un ensemble de 6 circuits logiques programmables Altera® Stratix V (technologie 28 nm), avec la possibilité de monter jusqu'à presque 600 000 neurones en sacrifiant l'aspect temps réel. Le passage à une échelle supérieure demeure cependant encore à démontrer.

1.3.2 SpiNNaker, une grappe de systèmes sur puce pour simuler un cerveau

SpiNNaker[96] est un projet de l'université de Manchester, faisant partie de l'initiative *Human Brain Project*[II] de l'Union européenne. À ce titre, l'un de ses objectifs est de fournir un outil de simulation pour étudier l'échelle intermédiaire entre un ensemble de quelques neurones et les aires cérébrales complètes.

Afin de garantir une grande versatilité, le système est entièrement numérique. Il est conçu sur l'idée que pour les applications de type réseau de neurones, les ressources de calculs représentent un coût négligeable en regard de celui des moyens de communication nécessaires entre les différents nœuds du réseau, qui doivent être en mesure d'acheminer en de multiples endroits un très grand nombre de petits paquets d'information, selon des chemins asynchrones. Ceci se différencie notamment des solutions de calcul hautes performances plus conventionnelles, reposant sur des échange point à point de larges paquets de données. L'architecture est une

I. De l'anglais *Look-Up Table*.
96. S. B. FURBER et coll., Proceedings of the IEEE, 2014.
II. https://www.humanbrainproject.eu

grappe de systèmes sur puce ARM® conçus pour la basse consommation et interconnectés selon un maillage torique qui devrait à terme compter un million de cœurs de calcul.

Le système fonctionne en temps réel. S'il est fait usage au sein de chaque cœur d'un temps local discret et cadencé à 1 kHz, la gestion du temps entre ces mêmes cœurs est implicite et événementielle : les paquets sont acheminés de façon asynchrone en moins d'une milliseconde, par le biais d'un bus numérique utilisant le protocole *Address-Event Representation* (AER) mentionné précédemment. L'utilisation d'un tel protocole de routage permet de limiter le nombre de connexions physiques entre les nœuds de l'architecture. Associée à gestion asynchrone du temps global, elle garantit un fort potentiel de passage à l'échelle.

Les systèmes sur puces adoptés regroupent chacun 18 cœurs ARM968, cadencés à 200 MHz et utilisant une arithmétique d'entiers 32 bits. Les 18 cœurs d'une puce partagent 128 Mio de mémoire [I], et chacun d'entre eux dispose de 96 Kio de mémoire locale occupant une proportion significative de la surface de silicium utilisée (figure 1.13a p. 48). 16 cœurs sont utilisés pour effectuer les calculs, un 17e cœur est dédié au routage des signaux et le dernier est gardé en réserve en cas de défaut d'un des autres cœurs [II]. Les puces sont regroupées sur des circuits imprimés par ensembles de 48 puces, la communication entre ces circuits imprimés étant réalisée à l'aide de circuits logiques programmables. À terme, le système complet comportera 1200 de ces circuits imprimés, pour une consommation électrique estimée à 75 kW en crête.

Un cœur permet de simuler environ un millier de neurones « intègre-et-tire avec fuites », pour mille fois plus de synapses *non plastiques*. L'implémentation d'une plasticité synaptique peut nécessiter de réduire sensiblement ce nombre. Pour faciliter la programmation de cette architecture matérielle, une couche logicielle est chargée de répartir automatiquement le réseaux de neurones à simuler sur les différents processeurs, en s'appuyant sur la hiérarchie et les connexions synaptiques de ce dernier.

Une consommation énergétique de 8 nJ par événement synaptique a été observée avec réseau de 250 000 neurones « intègre-et-tire avec fuites » et 80 millions synapses, pour 1,8 milliards d'activations synaptiques par seconde[97]. Par ailleurs, des travaux récents ont montré qu'il est possible d'interfacer assez aisément cette plateforme avec un capteur neuromorphique, tel qu'une rétine de type *Dynamic Vision Sensor* [98,99] .

1.3.3 Les approches hybrides : une communication numérique mais des calculs analogiques

Une alternative à l'approche entièrement numérique pour résoudre les équations des modèles de neurones ou de synapses est d'exploiter le comportement *analogique* de certains com-

I. Prenant la forme d'une puce externe soudée directement sur la puce contenant les 18 cœurs de calcul.

II. Des puces avec seulement 17 cœurs fonctionnels sur les 18 attendus sont acceptées en sortie des chaînes de fabrication afin d'améliorer leur rendement.

97. E. STROMATIAS et coll., International Joint Conference on Neural Networks, 2013.

98. F. GALLUPPI et coll., Neural Information Processing : 19th International Conference, Proceedings, Part II, 2012.

99. G. ORCHARD et coll., IEEE International Symposium on Circuits and Systems, 2015.

posants ou circuits pour réaliser des opérations mathématiques [I].

Comme l'a fait remarqué Carver Mead à la fin des années 1980[14], cette approche est particulièrement pertinentes dans le cas des systèmes neuro-inspirés puisque de nombreuses quantités neurobiologiques présentent entre elles des dépendances exponentielles, qu'il est alors possible de modéliser en exploitant la dépendance exponentielle à la tension de grille du courant de drain d'un transistor à effet de champ lorsque ce dernier est utilisé dans son régime faible inversion [II]. À opération mathématique donnée, une approche analogique, dotée de primitives plus riches que de simples fonctions logiques, permet généralement de réduire le nombre de composants utilisés par rapport à un circuit numérique, qui transpose les opérations vers l'algèbre de BOOLE[100].

Un second avantage à utiliser des transistors dans le régime de faible inversion est à trouver du côté de la consommation énergétique. Puisqu'il s'agit d'exploiter les courants de fuite des transistors, les intensités électriques sont très significativement plus faibles que celles obtenues dans le cadre d'un usage « au-dessus du seuil ». Ceci laisse présager d'une consommation énergétique réduite en regard de celle d'un système entièrement numérique.

L'approche analogique n'est cependant pas idéale pour tout. En premier lieu, le passage à l'échelle lors d'un changement de nœud technologique est souvent moins aisé que dans le cadre d'une conception totalement numérique. Par ailleurs, les stratégies de résolution analogique sont généralement très affectées par la variabilité des composants ou encore la présence de bruit lors de l'exécution. Néanmoins, ces deux derniers points ne semblent pas nécessairement bloquants, dans la mesure où nous cherchons à réaliser une plateforme matérielle *capable d'apprendre*, avec les composants dont elle dispose. Enfin, il est plus difficile de faire évoluer les modèles de neurones et de synapses sur une architecture analogique que sur une architecture numérique.

1.3.3.1 Neurogrid

Neurogrid[101], développée par l'université de Stanford, est une plateforme matérielle neuromorphique mêlant calcul analogique et circuits numériques.

Les neurones sont réalisés en utilisant des circuits à transistors dans leur régime de faible inversion, afin de limiter la consommation énergétique et d'autoriser une forte densité d'intégration. Le circuit analogique décrivant un neurone nécessite toutefois 337 transistors, le modèle neuronal retenu (« intègre-et-tire » quadratique) adoptant une description poussée des mécanismes biologiques. Les déviations de ces transistors résultant des procédés de fabrication peuvent en partie être compensées, en jouant sur les tensions de polarisation. Néanmoins, ces

I. Cette approche n'est d'ailleurs pas sans rappeler la résolution d'équations différentielles au moyen de calculateurs analogiques à base d'amplificateurs opérationnels.

14. C. MEAD, Proceedings of the IEEE, 1990.

II. Ce régime de fonctionnement est également qualifié de « sous le seuil ».

100. R. SARPESHKAR, Neural Computation, 1998.

101. B. V. BENJAMIN et coll., Proceedings of the IEEE, 2014.

corrections sont de nature limitée en raison du partage des circuits de polarisation entre les différents circuits neuronaux.

Les synapses sont partagées au moyen d'une mémoire externe, ce qui permet d'éviter la présence d'une piste physique par synapse et s'avère particulièrement intéressant dans le cas d'un système dont la connectivité est parcimonieuse. Par ailleurs, les arbres dendritiques sont également partagés. Si cela permet de réduire toujours plus le nombre global de transistors, cela empêche néanmoins tout apprentissage synaptique sur puce.

Le système physique ayant servi à démontrer l'intérêt du système Neurogrid est une carte de 16 *neurocœurs* qui intègrent chacun 256×256 neurones. Ces *neurocœurs* emploient un nœud technologique 180 nm ; ils intègrent 23 millions de transistors dans une surface de 12×14 mm^2. Il est fait état d'une puissance absorbée par la carte de 3,1 W lors d'un test de performances utilisant un million de neurones connectés par environ 8000 synapses chacun et présentant un taux de tir moyen par neurone de 0,42 impulsion par seconde. Ceci représente une consommation énergétique inférieure à 1 nJ par activation synaptique.

1.3.3.2 HiAER-IFAT

L'université de Californie San Diego a également développé une puce à très haute densité d'intégration tournée vers les applications neuromorphiques[102]. Elle regroupe 32 *neurocœurs Integrate-and-Fire Array Transceiver* (IFAT) à signaux mixtes, avec des circuits de calcul analogiques mais un routage synaptique numérique et événementiel. La puce est réalisée selon un procédé 130 nm et occupe 5×5 mm^2. Chaque *neurocœur* incluant 2000 neurones, de type « intègre-et-tire » ; la puce regroupe au total 2^{16} neurones.

Les synapses sont analogiques et multiplexées temporellement afin de limiter le nombre de circuits physiques. Leur conductance peut être programmée selon une dynamique logarithmique, tout en ayant recours à seulement 3 transistors, polarisés sous le seuil[103]. Un intérêt majeur de ces synapses analogiques en comparaison de simples valeurs numériques stockées dans une banque mémoire est qu'elles présentent une dynamique temporelle biologiquement plausible lors de la réception d'une impulsion. L'utilisation de ces synapses requiert en revanche un contrôle de la durée des impulsions, par le biais de compteurs, internes à chaque *neurocœur* et pilotés numériquement, ce qui pourrait limiter les possibilités d'apprentissage effectué directement sur la puce.

Il est rapporté une consommation de 252 µW, dans le cadre d'un test de performance à 5 millions d'événements par seconde, dont 20 % pour la partie analogique, et les 80 % restants pour les circuits numériques.

Une extension de ce système a récemment été proposée[104], nommée *Hierarchical Address-Event Routing IFAT* (HiAER-IFAT), dont l'idée est d'étendre de façon hiérarchique le pro-

102. T. YU et coll., IEEE Biomedical Circuits and Systems Conference, 2012.
103. T. YU et G. CAUWENBERGHS, IEEE International Symposium on Circuits and Systems, 2010.
104. J. PARK et coll., IEEE Transactions on Neural Networks and Learning Systems, 2016.

tocole numérique *Address-Event Representation*, afin de rentre possible un passage à *très* grande échelle. Si ce protocole permet d'émuler un câblage synaptique virtuel, la quantité de connexions impliquées est en effet limitée par la bande-passante du bus numérique, partagée entre l'ensemble des synapses à multiplexer temporellement. Afin de maximiser le potentiel de passage à l'échelle, ce projet suggère d'interconnecter les différents nœuds de calcul selon une structure d'arbre plutôt que les maillage planaires ou toriques généralement employés par les autres réalisations matérielles.

Les auteurs de ce travail proposent également d'ajouter la gestion d'un délai axonal en exploitant un protocole numérique événementiel et compact, sur 6 bits (pour une résolution temporelle de 1 ms). Il est en outre envisagé de mettre en œuvre, dans le domaine des adresses et de façon hiérarchique, une règle d'apprentissage de type *Spike-Timing Dependent Plasticity* .

En résumé, le système HiAER-IFAT utilise un temps analogique continu au sein des nœuds de calcul mais bascule vers une représentation numérique du temps en dehors de ces derniers, dans le but de pouvoir émuler des systèmes biologiquement plausibles du point de vue de la dynamique corticale, et munie d'une connectivité synaptique versatile. Un démonstrateur physique a été réalisé, constitué de 4 modules indépendants, qui incluent chacun 32 neurocœurs IFAT, un circuit logique programmable (Xilinx® Spartan 6) ainsi que 2 puces de 256 Mio de mémoire externe. L'ensemble permet de simuler une population de 262×10^3 neurones pour 262 millions de synapses. Il est muni d'un 5^e circuit logique programmable pour régir les communications entre les différents modules ainsi que pour les synchroniser entre eux en distribuant une horloge globale (200 MHz). Ce dernier point pourrait constituer une limitation lors d'une extension à très grande échelle de ce système.

Les auteurs rapportent une consommation globale de 10 W pour 720 millions d'événements synaptiques par seconde, c'est-à-dire un coût énergétique de 14 nJ par événement synaptique. Les pistes suggérées pour se rapprocher toujours plus des 10 fJ par événement synaptique du cerveau humain sont par exemple d'améliorer l'intégration des circuits de routage et des communications à l'échelle d'une galette de silicium ou de rapprocher davantage circuits de calcul et mémoire en misant sur la co-intégration en trois dimensions.

1.3.3.3 ROLLS

Le processeur neuromorphique ROLLS (pour *Reconfigurable On-Line Learning Spiking*) est une puce intégrée à très grande échelle développée à l'université de Zurich, avec le double objectif de constituer un substrat d'expérimentation *in silico* pour les neurosciences et d'être une plateforme adaptée à l'exécution de tâches d'apprentissage automatique. Il s'agit d'une architecture à signaux mixtes, convertissant les courants des signaux sous le seuil exploités au sein de la puce en impulsion de tensions permettant une transmission numérique à l'extérieure de celle-ci, en utilisant un protocole de type *Address-Event Representation*.

Le traitement des données est effectué en temps réel, de façon implicite : les impulsions sont traitées à leur réception, ce qui impose que les constantes de temps des circuits analogiques

permettent de fonctionner aux échelles de temps du problème simulé. Il peut s'agir d'une contrainte de conception assez forte puisque les constantes de temps biologiques sont souvent bien plus longues que celles usuellement rencontrées en conception de circuit intégrés. Les circuits logiques et numériques sont quant à eux asynchrones.

La puce regroupe 256 neurones analogiques « intègre-et-tire » de type exponentiel adaptatif[95]. Les dimensions de certains transistors critiques en termes de variabilité globale du système sont agrandies (notamment dans le circuit de sommation des neurones), afin de limiter les effets de la dispersion des caractéristiques des composants lors de la fabrication.

Les neurones sont co-intégrés avec :

- 256 × 256 synapses analogiques, plastiques à long terme et bistables ;
- 256 × 256 synapses programmables et munies de circuits de plasticité synaptique à court terme ;
- 256 × 2 rangées d'intégrateurs pouvant servir de synapses virtuelles.

Qu'il s'agisse des synapses plastiques à long terme ou des celles munies d'un circuit de plasticité à court terme, une vingtaine de transistors sont utilisés par synapse : presque deux tiers de la surface du processeur sont occupés par les cellules mémoires. La figure 1.13d p. 48 montre ainsi l'importante surface de silicium requise par les circuits synaptiques.

Il est possible d'utiliser une variante partiellement supervisée de la règle d'apprentissage *Spike-Timing Dependent Plasticity*[105]. Il est à noter que les comportements stochastiques éventuellement requis sont obtenus en exploitant la variabilité des trains d'impulsions d'entrée et des membranes postsynaptiques, ce qui évite le recours à un circuit supplémentaire, dédié à la génération de nombres aléatoires.

Une démonstration du potentiel de ce processeur neuromorphique a été réalisée sur une tâche de classification d'images, en utilisant une rétine artificielle de type *Dynamic Vision Sensor* pour s'affranchir d'un coûteux prétraitement sur les signaux d'entrée[I].

La consommation de la puce, réalisée dans une technologie 180 nm et occupant une surface de 51,4 mm^2 pour 12 millions de transistors, a été mesurée à 4 mW lors de simulations de réseaux neuronaux typiques.

1.3.3.4 BrainScaleS-FACETS-HICANN

Le projet BrainScaleS[106], dont le nom complet est *Brain-inspired multiscale computation in neuromorphic hybrid systems*, est un autre volet du *Human Brain Project* européen, développé par l'université de Heidelberg. Il s'appuie sur le projet antérieur FACETS[II], qui a notamment donné naissance à la puce neuromorphique HICANN[III]. Ce circuit intégré à très grande échelle[IV],

105. J. M. BRADER, W. SENN et S. FUSI, Neural computation, 2007.

I. Les images étant fixes, il est néanmoins requis pour produire des impulsions d'agiter légèrement les images, ce qui a un effet similaire aux saccades d'une rétine biologique.

106. J. SCHEMMEL et coll., IEEE International Symposium on Circuits and Systems, 2010.

II. Acronyme de *Fast Analog Computing with Emergent Transient States*.

III. La signification de cet acronyme est *High Input Count Analog Neural Network*.

IV. *Very Large-Scale Integration circuit* en anglais.

à la différence de SpiNNaker et de ses cœurs de calcul numériques, exploite des circuits de calcul *analogique* à temps continu. BrainScaleS vise à fournir une plateforme de recherches pour l'étude de modèle neuronaux *biologiques* dans le cadre de topologies réalistes, présentant de l'ordre de 10 000 connexions synaptiques par neurone.

Le modèle neuronal retenu est de type « intègre-et-tire » adaptatif exponentiel[95], permettant de reproduire un grand nombre de comportements neuronaux observés en neurosciences expérimentales. Les équations différentielles de ce modèle sont émulées directement par la physique des transistors. Chaque circuit neuronal occupe une surface de $150 \times 10\ \mu m^2$ dans le procédé 180 nm employé.

Afin d'augmenter la densité d'intégration des circuits, ce système fonctionne en temps accéléré de 1000 à 100 000 fois par rapport au temps réel. Ceci permet en effet de limiter les dimensions des capacités électriques requises par les constantes de temps des équations résolues, Par ailleurs, les intensités électriques au cœur des circuits demeurent de fait généralement entre 0,1 et 1 µA, ce qui permet de simplifier la conception des circuits par rapport à des techniques plus avancées spécifiques à un usage profondément sous le seuil des transistors, telles qu'employées dans les autres architectures neuro-inspirées à signaux mixtes.

Pour limiter les transactions avec une mémoire distante, des synapses sont co-intégrées directement avec les neurones, au sein d'un *Analog Network Core* (ANC). Chaque cœur regroupe ainsi 256 neurones et 224 synapses. Le système est conçu pour permettre à un neurone de recevoir jusqu'à 14 336 connexions présynaptiques. Le contrôle des dendrites est effectué par le biais de 23 paramètres analogiques stockés dans des cellules mémoires *analogiques*. Les synapses sont quant à elles stockées de façon *numérique* sur 4 bits de mémoire vive statique, qui pilotent un convertisseur numérique-analogique en courant. Il est en outre possible de doubler la précision des synapses en divisant par deux leur nombre. Ces synapses sont dotées de deux types de plasticités, l'une à court terme, l'autre de type *Spike-Timing Dependent Plasticity*. Ces plasticités synaptiques sont réalisées par un mélange de circuits d'accumulation ou de corrélation analogiques et de circuits de lecture et de contrôle numériques[107,108]. Comme nous pouvons l'observer à la figure 1.13c p. 48, les circuits synaptiques occupent la majeure partie de la surface de silicium disponible.

La communication entre les *Analog Network Cores* est échelonnée selon deux niveaux. Un premier niveau, basé sur un protocole différentiel faible tension qui exploite des répéteurs de signaux et des boucles à verrouillage de délai *DLL*[I] permet d'interconnecter 352 puces HICANN à l'échelle d'une galette de silicium de 20 cm. Physiquement, ce niveau repose sur un pontage par fils d'or entre les puces d'une même galette. Une galette de silicium permet de disposer de 180 000 neurones et 40 millions de synapses. Un second niveau, pour les communications entre les galettes, utilise quant à lui un protocole numérique d'horodatage à temps discret. La mise en paquets est gérée par des circuits intégrés développés pour un client et des circuits logiques

107. J. SCHEMMEL et coll., IEEE International Joint Conference on Neural Network Proceedings, 2006.
108. J. SCHEMMEL et coll., IEEE International Symposium on Circuits and Systems, 2007.
I. De l'anglais *Delay-Locked Loop*.

programmables, permettant d'atteindre une précision temporelle inférieure à 4 ns

Il est à noter que de façon similaire au système NeuroFlow , la programmation de cette plateforme s'effectue par le biais de l'interface logicielle répandue PyNN[94].

1.3.4 TrueNorth, un circuit intégré développé pour un client, neuromorphique et numérique

À contre-courant de l'approche à signaux mixtes des précédents systèmes, TrueNorth[13] est une architecture impulsionnelle entièrement numérique réalisée par IBM®. Ce circuit intégré développé pour un client exploite un nœud technologique actuel 28 nm. La fréquence relativement faible nécessaire au bon fonctionnement du système permet le recours à une technologie de transistors *à faibles fuites*, ce qui limite la puissance statique dissipée par ces derniers. La puce regroupe 5,4 milliards de transistors (dont une partie significative est consacrée aux 428 millions de bits mémoires disponibles) sur 4,3 cm^2, selon un maillage planaire de 4096 *neurocœurs numériques*[I] permettant de simuler jusqu'à 1 million de neurones et 256 millions de synapses.

Chaque neurocœur occupe 240 × 390 μm^2 et comporte 256 axones d'entrée pour 256 dendrites de sortie, connectés par 256×256 synapses programmables non plastiques. Une différence majeure avec les autres réalisations neuromorphiques matérielle est le caractère *binaire* de la connectivité synaptique. Néanmoins, l'efficacité postsynaptique qui leur est associée peut prendre une valeur parmi quatre possibles, programmées à l'avance. Il est également possible de programmer un délai axonal sur 4 bits, avec une résolution temporelle de 1 ms.

Les neurocœurs comportent chacun un unique circuit neuronal numérique, multiplexé temporellement. La dynamique temporelle des neurones est discrétisée par pas de 1 ms à l'aide d'une horloge globale à 1 kHz, tandis que le reste de l'exécution s'effectue de façon parallèle et asynchrone. Le modèle neuronal adopté est une variante de neurone « intègre-et-tire » possédant un nombre important de comportements différents[110]. Le circuit numérique associé, afin d'être de complexité limitée, effectue des opérations arithmétiques à virgule fixe. Il est composé de 1272 portes logiques, dont 348 pour un générateur de nombre pseudo-aléatoires (PRNG[II]), qui permet d'ajouter des comportements stochastiques aux neurones ou aux synapses. En raison de la stratégie de multiplexage temporel employée, même dans le cas d'une telle approche entièrement numérique, la surface de silicium dédiée à la mémoire utile aux synapses demeure supérieure à celle utilisée par les circuits neuronaux (figure 1.13 p. 48).

La puce TrueNorth a montré une extrême efficacité énergétique pour l'inférence[13], avec par exemple le cas d'une application vidéo (30 images par seconde et 400 × 240 pixels) durant

13. P. A. MEROLLA et coll., Science, 2014.

I. La conception de ces derniers s'appuie sur un prototype antérieur présenté dans la référence [109]. Depuis, les auteurs rapportent une augmentation d'un facteur 15 de la densité d'intégration et une réduction d'un facteur 100 de la consommation électrique.

110. A. S. CASSIDY et coll., International Joint Conference on Neural Networks, 2013.

II. De l'anglais *Pseudo-Random Number Generator*.

laquelle la consommation électrique était de 63 mW, entraînant dissipation surfacique d'environ 20 mW·cm^{-2}. Sur un réseau de neurones s'activant en moyenne 20 fois par seconde et connectés en moyenne à 128 synapses chacun, les auteurs font mention d'une consommation énergétique de 26 picojoules par événement synaptique. Ce nombre est 176 000 fois plus faible que celui de la consommation d'un logiciel spécialisé pour simuler l'architecture TrueNorth, ou encore 769 fois plus faible que lors de l'exécution du même réseau neuronal sur la plateforme SpiNNaker de la partie 1.3.2. Si le système SpiNNaker avait déjà confirmé le potentiel de faible consommation énergétique d'une architecture matérielle numérique spécialisée dans les réseaux impulsionnels, en étant capable de classifier des chiffres manuscrits [I] à 50 chiffres par seconde pour un budget énergétique de 6 millijoules par chiffre[112], le système TrueNorth atteint le même taux de classification correcte (95 %, similaire à une implémentation logicielle), à une cadence 20 fois plus rapide (1000 chiffres par seconde) pour une consommation énergétique 1500 fois plus faible (4 microjoules par chiffre)[86]. Ces améliorations s'expliquent par la spécialisation matérielle plus avancée et la plus faible résolution synaptique employée.

Il a par ailleurs récemment été montré qu'il était possible d'exécuter des réseaux de neurones convolutifs sur une ou plusieurs puces TrueNorth[113]. Si une certaine stratégie de connexion et le sacrifice potentiel de neurones sont nécessaires pour mettre en place la topologie particulière des réseaux convolutifs, ce travail fait état de performances de l'ordre de l'état de l'art sur un ensemble représentatif de tests d'apprentissage automatique. Néanmoins, pour l'ensemble des ces démonstrations, l'apprentissage est toujours réalisé hors puce.

Un langage de programmation spécialisé a été développé[114], reposant sur l'idée d'assembler à plus grande échelle des réseaux de neurones de base déjà optimisés. Par ailleurs, il est à noter que le fait de disposer d'une plateforme entièrement numérique permet d'obtenir une correspondance exacte entre les simulations logicielles et l'exécution matérielle.

1.3.5 Conclusion

Ce panorama des grands projets actuels de réalisations matérielles d'architectures de calcul neuro-inspirées, dont les principales caractéristiques sont récapitulées dans la table 1.1, montre la diversité des approches explorées.

Les solutions retenues dépendent naturellement des objectifs du projet, selon qu'il penche du côté des neurosciences ou au contraire davantage vers l'ingénierie neuromorphique. Ces deux aspects ne sont cependant pas antinomiques, certaines architectures affichant dans leurs objectifs de faire avancer les deux domaines. Par ailleurs, il est toujours possible que certaines découvertes effectuées dans l'un des deux domaines apportent un éclairage nouveau dans l'autre, faisant progresser l'ensemble de ces champs disciplinaires. Les solutions envisagées dif-

I. Tirés de la base de données MNIST, devenue un test de performances classique en apprentissage automatique[111].

112. E. STROMATIAS et coll., International Joint Conference on Neural Networks, 2015.
113. S. K. ESSER et coll., arXiv preprint arXiv :1603.08270, 2016.
114. A. AMIR et coll., International Joint Conference on Neural Networks, 2013.

Système	Référence	Circuit logique programmable	Circuit intégré développé pour un client	Grappe de systèmes sur puce	Signaux numériques	Signaux mixtes	Apprentissage en ligne	Temps réel	Temps accéléré	Énergie par événement synaptique
Bluehive	[90]	✓	-	-	✓	-	?	✓	-	?
NeuroFlow	[93]	✓	-	-	✓	-	✓	✓	-	?
SpiNNaker	[96]	-	-	✓	✓	-	✓	✓	-	8 nJ
TrueNorth	[13]	-	✓	-	✓	-	-	✓	-	26 pJ
Neurogrid	[101]	-	✓	-	-	✓	-	✓	-	<1 nJ
BrainScaleS	[106]	-	✓	-	-	✓	✓	-	✓	<1 nJ
ROLLS	[115]	-	✓	-	-	✓	✓	✓	-	?
HiAER-IFAT	[104]	-	✓	-	-	✓	-	✓	-	14 nJ

TABLE 1.1 – Synthèse des principales réalisations neuromorphiques matérielles. Les points d'interrogations indiquent des données non disponibles dans la littérature.

fèrent également selon le budget alloué au projet dont elles découlent : une grappe de circuits logiques programmables est financièrement plus abordable que le développement complet d'un circuit intégré développé pour un client, tout en étant plus adaptée aux recherches exploratoires.

Calcul numérique contre calcul analogique. Dans l'ensemble, les solutions numériques Bluehive, NeuroFlow ou encore SpiNNaker confirment le caractère hautement versatile des simulations qu'elles permettent d'exécuter. Cependant, si leur efficacité énergétique pour les tâches cognitives est certes assurément meilleure que celle de plateformes reposant sur des processeurs graphiques ou plus généralistes, leur consommation énergétique demeure toutefois parmi les plus élevée des solutions présentées et se prête assez difficilement à un usage embarqué.

À l'inverse, les modèles de composants neuronaux qu'il est possible de simuler sur la plupart des architectures à signaux mixtes, qui effectuent une partie des calculs de façon analogique, telles que Neurogrid, ROLLS ou encore HiAER-IFAT, sont plus fortement contraints, par les circuits électroniques qui les décrivent *physiquement*. Néanmoins, ces circuits, qui reposent sur une utilisation en régime de faible inversion des transistors, permettent d'améliorer grandement l'efficacité énergétique de ces architectures. Avec des consommations électriques rapportées d'une dizaine de watts au maximum (voire de quelques milliwatts seulement) pour des réseaux neuronaux représentatifs d'usages courants, ces dernières laissent entrevoir une réponse à un usage embarqué d'accélérateurs électroniques pour les tâches cognitives.

La dernière grande architecture à signaux mixtes, la puce HICANN du projet BrainScaleS, poursuit un objectifs différant d'une utilisation embarquée puisqu'il s'agit plutôt de fournir une plateforme d'exploration biologique *in silico*. Une enveloppe limitée de la consommation électrique totale n'est alors pas nécessairement requise. Cette architecture présente deux particularités majeures, à savoir le recours à un temps accéléré afin de simplifier l'adéquation des constantes de temps biologiques à celles des circuits électroniques associés, et le fait d'interconnecter les *neurocœurs* directement à niveau de la galette de silicium pour faciliter le passage à l'échelle.

Enfin, l'architecture entièrement numérique TrueNorth qui complète l'échantillon des solutions présentées apporte une certaine surprise. L'utilisation d'un nœud technologique bien plus petit que les architectures à signaux mixtes, dans une technologie faibles fuites et associée à un multiplexage temporel de certains circuits coûteux en surface de silicium, permettent à cette architecture d'afficher un nombre de neurones par *neurocœur* similaire aux approches à signaux mixtes, tout en ayant une meilleure efficacité énergétique.

Remarque

Nous pouvons constater à l'issue de ce diaporama que la majorité des modèles de neurones impulsionnels retenus par les systèmes sont de type « intègre-et-tire » avec fuites, bien que possiblement modifié, comme par exemple le modèle exponentiel adaptatif [95].

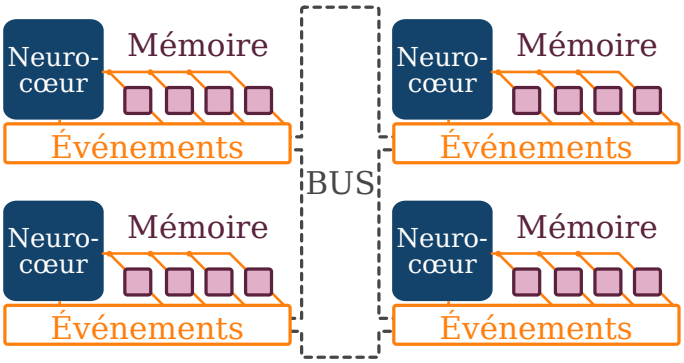


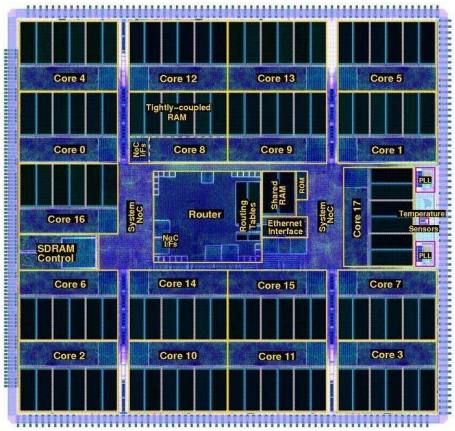
FIGURE 1.12 – Vision schématique d'architecture neuromorphique (de type impulsionnel) telle qu'imaginée dans la plupart des projets de recherche en cours. En comparaison de l'architecture de VON NEUMANN (figure 1.6b), la mémoire est placée plus proche des unités de calcul (les *neurocœurs*), afin d'obtenir une meilleure adéquation de la topologie matérielle avec la structure hautement parallèle des couches d'un réseau de neurones artificiels (figure 1.6a). La communication entre ces blocs hybrides « *neurocœur* et mémoire » est assurée par le biais d'un bus numérique de type événementiel.

Communications à grande échelle. Un consensus général est l'utilisation d'un protocole numérique pour les communications à grande échelle, au moins hors des *neurocœurs* et notamment sous la forme de diverses variantes du protocole *Address-Event Representation*. La figure 1.12 schématise cette idée de relier entre elles, au moyen d'un bus événementiel, des unités de calcul neuronal associées à de la mémoire pour la partie synaptique. L'un des intérêts d'un bus de communication numérique est de pouvoir limiter la quantité de connexions physiques requise pour être en mesure de relier une paire arbitraire de neurones, dont le nombre explose lors d'un passage à plus grande échelle. Plusieurs architectures présentées profitent par ailleurs de la nature numérique de certaines de leurs communications pour mettre en place une gestion simple du délai axonal (Bluehive, NeuroFlow, TrueNorth), voire potentiellement de la plasticité synaptique directement dans le domaine des adresses (HiAER-IFAT).

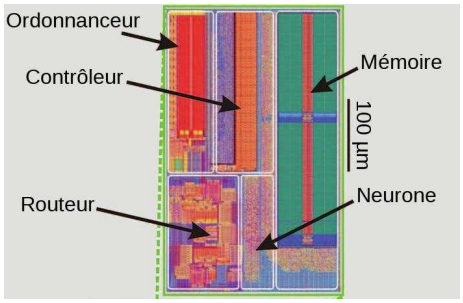
À la recherche de synapses compactes, co-intégrables et plastiques. Une observation à l'issue de ce panorama est que la réalisation de synapses artificielles matérielles pose encore certains défis, notamment pour les plateformes visant un usage embarqué. Les solutions actuellement retenue sont de plusieurs types. Les architectures à base de circuits logiques programmables utilisent des banques de mémoires externes, ce qui simplifie leur utilisation en permettant de modifier plus facilement la topologie du réseau simulé mais sacrifie la consommation énergétique. Les autres architectures tendent à utiliser de la mémoire cache, voire des cellules synaptiques analogiques co-intégrées avec les neurones d'un *neurocœur*. Néanmoins, les solutions actuellement retenues nécessitent une surface de silicium conséquente, comme le montrent les photographies de puces de la figure 1.13. Dans chacun des cas présentés, la surface consacrée aux cellules mémoires d'usage synaptique est non négligeable et dépasse celle allouée aux circuits neuronaux. Par ailleurs, pour les architectures disposant de synapses

plastiques, la plasticité des cellules mémoires est encore limitée, et nécessite généralement d'accroître la complexité des circuits associés, au détriment de la densité d'intégration. Ainsi, il existe un besoin de synapses artificielles qui conserveraient une forte densité d'intégration tout en disposant d'un comportement plastique et en étant co-intégrable avec les circuits neuronaux.

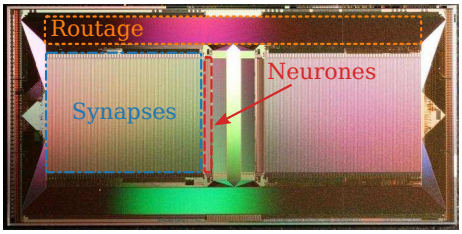
(a) SpiNNaker



(b) TrueNorth



(c) HICANN



(d) ROLLS

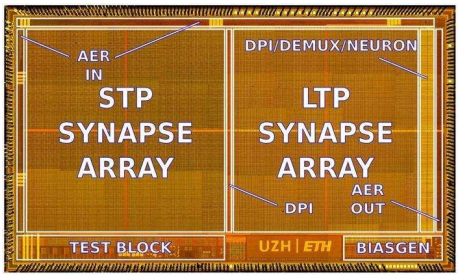


FIGURE 1.13 – Photographies de puces neuro-inspirées : *(a)* SpiNNaker[96], *(b)* TrueNorth[13], *(c)* HICANN[106] et *(d)* ROLLS[115]. Dans chacun des cas, la surface de silicium allouée à la mémoire (permettant notamment de décrire les synapses) est supérieure à celles de circuits neuronaux.

1.4 Synapses artificielles : les nanocomposants potentiels

Dans cette partie, nous nous intéressons à des nanocomposants mémoires innovants constituant de potentielles synapses artificielles.

Depuis une petite dizaine d'années, il y a un intérêt certain pour le développement de nouvelles mémoires non volatiles. Ceci fait notamment suite à la démonstration expérimentale[116] par les laboratoires HP® en 2008 d'un composant proche d'une prédiction théorique établie presque 40 ans auparavant par Leon CHUA[117] à propos d'un quatrième dipôle passif élémentaire : le *memristor*. Nous retiendrons de ce composant la nature programmable de sa résistance électrique en fonction de son histoire antérieure[I]. La définition généralement ad-

116. D. B. STRUKOV et coll., Nature, 2008.
117. L. CHUA, IEEE Transactions on Circuit Theory, 1971.
I. Ce comportement est d'ailleurs à l'origine de son nom, contraction de *memory resistor*.

mise d'un *memristor* est celle d'un dipôle électrique passif dont la caractéristique courant-tension présente une courbe d'hystérésis pincée et confinée aux 1er et 3e quadrants en convention récepteur[118]. Cette définition englobe davantage de familles de composants que simplement celle du démonstrateur des laboratoires HP®, puisqu'un nombre important de nouvelles technologies mémoires faisant actuellement l'objet de recherches et de développements technologiques sont basées sur une modulation de résistance électrique plutôt qu'un stockage de charges.

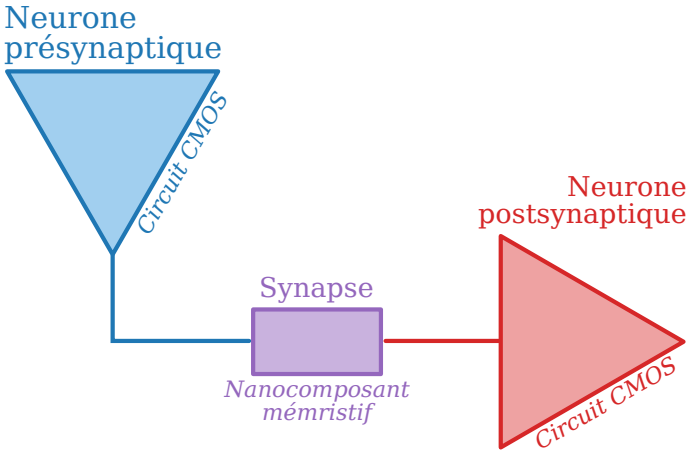


FIGURE 1.14 – Schéma de la topologie générique des neurones et synapses artificiels envisagée : des nanocomposants mémristifs sont utilisés comme synapses plastiques pour relier entre eux des circuits en technologie CMOS qui réalisent les fonctions neuronales souhaitées.

Ces nouvelles cellules mémoires, que nous qualifierons de « mémristives » puisque leur résistance électrique est modulée selon l'historique des impulsions électriques qu'elles reçoivent, présentent un comportement rappelant fortement le rôle attendu d'une synapse artificielle au sein d'une architecture neuro-inspirée. Des travaux théoriques et expérimentaux ont par ailleurs rapidement remarqué que de tels composant mémristifs permettaient d'obtenir une plasticité synaptique de type *Spike-Timing Dependent Plasticity* sans nécessiter de circuits numériques dédiés[119], différentes variations de ce type de plasticité synaptique pouvant être obtenus en jouant sur les formes d'onde des impulsions de programmation[120].

Dans cette partie nous passons en revue différentes technologies mémristives, dans la grande majorité non volatiles, dans le but d'évaluer leur potentiel pour un usage synaptique. L'idée est d'utiliser ces composants pour relier électriquement les neurones (figure 1.14) d'un même *neurocœur* au sein d'une architecture neuro-inspirée. Nous cherchons un composant programmable électriquement présentant une faible consommation électrique, permettant d'obtenir une forte densité d'intégration et co-intégrable avec un procédé de fabrication CMOS standard. Il est à noter qu'une utilisation synaptique d'un composant mémoire n'exige pas

118. L. CHUA, Applied Physics A, 2011.
119. B. LINARES-BARRANCO et T. SERRANO-GOTARREDONA, Nature Precedings, 2009.
120. T. SERRANO-GOTARREDONA et coll., Frontiers in Neuroscience, 2013.

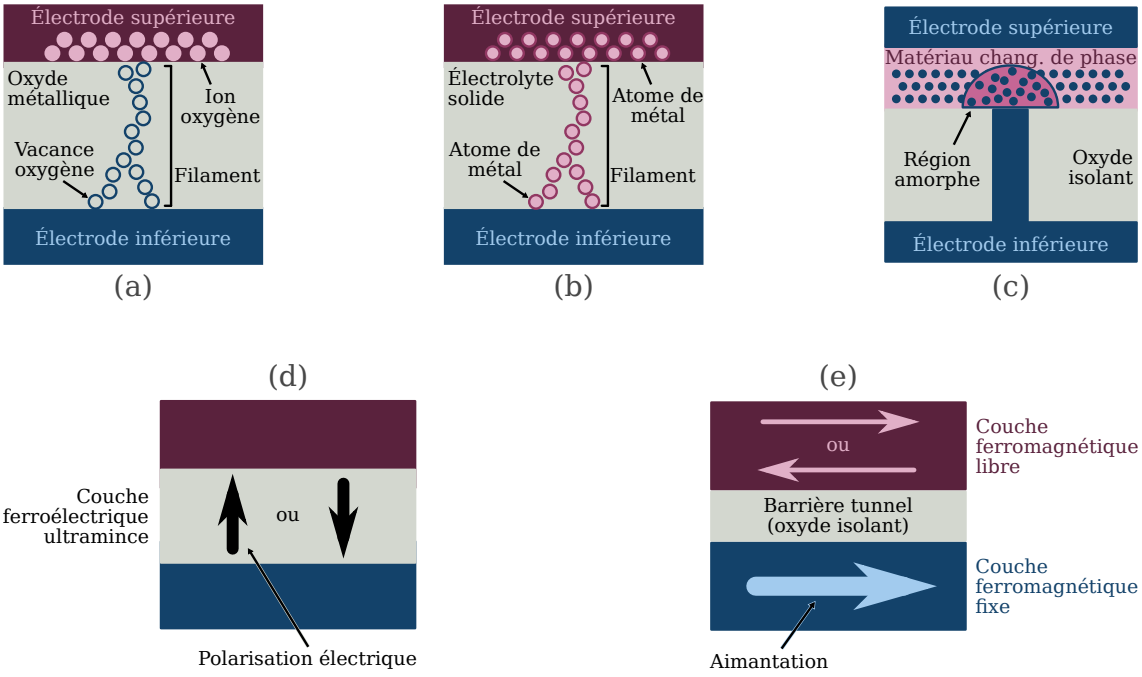


FIGURE 1.15 – Synapses artificielles : schémas de principe des principales technologies mémoires potentielles. *(a)* Cellule à base d'oxydes de métaux de transition. *(b)* Cellule électrochimique métallique. *(c)* Cellule à changement de phase. *(d)* Jonction tunnel ferro-électrique. Des domaines de polarisations opposées peuvent coexister au sein de la couche ferro-électrique. *(e)* Jonction tunnel magnétique. Les deux états magnétiques sont exclusifs.

nécessairement les mêmes performances qu'une utilisation plus conventionnelle : à priori la fréquence d'écriture d'une cellule mémoire n'est pas la même, tout comme la rétention peut éventuellement être réduite selon l'usage ou encore les étapes de lecture ou d'écriture être plus lentes en raison des constantes de temps usuelles des tâches cognitives et du caractère parallèle d'une architecture neuro-inspirée.

1.4.1 Cellules à base d'oxydes de métaux de transition

Il s'agit de la famille de composants du démonstrateur expérimental de 2008 des laboratoires HP®. Ces composants mémristifs sont composés de deux électrodes métalliques entourant une couche d'oxyde métallique (figure 1.15a). L'application d'un champ électrique provoque le déplacement de vacances d'oxygène, qui finissent par former un ou plusieurs filaments conducteurs à travers l'oxyde, faisant passer le composant d'un état OFF de forte résistance R_{OFF} à un état ON de faible résistance R_{ON}. Les matériaux employés (HfO_x, TaO_x, TiO_x ou encore AlO_x), présentent l'avantage d'être d'un usage courant dans l'industrie des semi-conducteurs[76]. Si la cellule mémoire est de composition simple, il existe encore toutefois des discussions sur sa physique exacte, notamment autour de la forme des filament ou le rôle des électrodes[121].

Cette famille de composants a été l'objet d'études poussées, visant à évaluer les débouchés industriels potentiels[121,122]. L'un de ses avantages est notamment la facilité d'intégration verticale[123], dont la mise en œuvre à grande échelle bénéficierait fortement à la densité d'intégration surfacique.

L'un de inconvénients majeurs de cette technologie mémoire est le manque de reproductibilité de ses transitions. La nature filamentaire de ces dernières entraîne non seulement une forte variabilité des caractéristiques entre composants, mais également entre les différents cycles de programmation d'un même composant[76].

Il a été démontré expérimentalement, à l'échelle d'un composant, la possibilité d'obtenir un comportement plastique graduel, pouvant permettre de mettre en place une règle d'apprentissage de type *Spike-Timing Dependent Plasticity* par l'application d'impulsions de tension[124].

D'autres travaux expérimentaux font état d'un rapport des résistances entre états OFF et ON de 3 ordres de grandeur[89]. Les auteurs font également observer la pertinence d'une programmation binaire probabiliste, au travers d'impulsions d'amplitude réduite, obtenant des performances d'apprentissage similaires à celles d'un système analogue reposant sur une évolution graduelle des poids synaptiques. Néanmoins, l'origine exacte de ce comportement stochastique est encore mal comprise.

En 2013, ALIBART et coll. ont réalisé la première démonstration expérimentale d'apprentissage *in situ* à l'échelle d'un système, avec un perceptron utilisant des cellules TiO_2 comme

121. R. WASER et M. AONO, Nature materials, 2007.
122. H. S. P. WONG et coll., Proceedings of the IEEE, 2012.
123. B. PRINCE, , 2014.
124. S. YU et coll., IEEE Transactions on Electron Devices, 2011.

synapses artificielles[125]. S'agissant d'une première démonstration, l'échelle du système reste toutefois faible.

Plus récemment, PREZIOSO et coll. ont également réussi à construire un perceptron matériel exploitant également des synapses TiO_2 et capable d'apprentissage *in situ*[126]. Les avancées technologiques de ces travaux sont multiples : la fabrication des cellules mémoires peut être réalisée à moins de 300 °C, ces dernières s'affranchissent de la nécessité d'un transistor de sélection, présentent un rapport de résistances R_{OFF}/R_{ON} de 4 ordres de grandeur ($R_{ON} \simeq 1$ à 10 kΩ) et une variabilité réduite. Les dimensions latérales des cellules mémoires demeurent toutefois assez larges (200×200 nm^2), tout comme les impulsions de lecture et d'écriture employées restent longues (500 µs).

Par ailleurs, GAO et coll. ont montré qu'il était possible de fabriquer de telles cellules mémoires ne nécessitant pas d'étape de formation initiale des filaments (*forming* en anglais)[127]. Cependant, l'échelle du système présenté est celle de la preuve de concept, avec une matrice de 2×2 composants.

1.4.2 Cellules électrochimiques métalliques

Cette famille de composants mémristifs est d'une structure assez similaire à celle des cellules à base d'oxydes de métaux de transition, faite de deux électrodes qui entourent un électrolyte solide, l'une des électrodes (qualifiée d'« active ») étant constituée d'un métal pouvant s'oxyder en ions métalliques[128]. L'application d'un champ électrique entraîne la migration d'ions depuis l'électrode active vers l'autre électrode, ce qui finit par former un pont conducteur entre les deux (figure 1.15b). Ce processus de migration ionique est affecté de fluctuations stochastiques, ce qui rend la formation du pont conducteur différente lors de chaque répétition et entraîne, comme pour les cellules à base d'oxydes de métaux de transition, une certaine non reproductibilité des transitions de résistance[76]. Néanmoins, des travaux récents font état de démonstrateurs technologiques industriels[129], avec une puce de 2 Gio réalisée dans une technologie 27 nm avancée, présentant un rapport de résistances R_{OFF}/R_{ON} de 3 ordres de grandeur ($R_{ON} \simeq 0{,}01$ à 1 MΩ).

Dans le cadre d'une utilisation synaptique, plusieurs travaux expérimentaux[130,131] ont démontré à l'échelle d'un composant la possibilité d'obtenir un comportement plastique apparenté à celui de la *Spike-Timing Dependent Plasticity*. La consommation énergétique des impulsions de programmation appliquées sur la cellule mémoire est en revanche élevée, aux alentours du millijoule.

Des résultats de simulations réalisées par SURI et coll. ont montré qu'il était possible, en

125. F. ALIBART, E. ZAMANIDOOST et D. B. STRUKOV, Nature Communications, 2013.
126. M. PREZIOSO et coll., Nature, 2015.
127. L. GAO et coll., Nanotechnology, 2015.
128. R. WASER et coll., Advanced Materials, 2009.
129. J. ZAHURAK et coll., IEEE International Electron Devices Meeting, 2014.
130. S. H. JO et coll., Nano Letters, 2010.
131. C. DU et coll., Advanced Functional Materials, 2015.

utilisant des synapses électrochimiques métalliques, d'obtenir un apprentissage à l'échelle d'un système complet, sur une tâche audio mais également vidéo[88,132]. Ces travaux présentent le caractère notable de proposer une utilisation binaire des synapses artificielles retenues, associée à une version probabiliste de règle d'apprentissage de type *Spike-Timing Dependent Plasticity*. Par ailleurs, les auteurs ont étudié la possibilité de mettre en œuvre cette programmation probabiliste en exploitant les comportements stochastiques intrinsèques des synapses. La mauvaise compréhension de l'origine physique de ces fluctuations aléatoires en rend néanmoins l'exploitation difficile, ce qui tend à favoriser une solution exploitant un circuit de génération de nombres pseudo-aléatoires externe aux synapses.

Plus récemment, LA BARBERA et coll. ont observé expérimentalement l'existence de plusieurs régimes plastiques au sein de cellules utilisant un électrolyte de sulfure d'argent (Ag_2S)[133]. En collaboration avec les auteurs précédents, nous avons depuis découvert une forme supplémentaire de plasticité intrinsèque offerte par ces nanocomposants. Nous présenterons au chapitre 4 comment exploiter cette dernière pour mettre en œuvre une version originale de *Spike-Timing Dependent Plasticity* à l'aide d'impulsions de programmation simples. Dans ce même chapitre, nous verrons également comment la compréhension fine des transitions entre les multiples comportements plastiques à notre disposition permet de réaliser un apprentissage à l'échelle du système.

1.4.3 Cellules à changement de phase

Ces composants mémristifs sont constitués de deux électrodes qui entourent un verre de chalcogénure. Le profil temporel de température appliqué à ce dernier durant une phase de programmation détermine la proportion de matériau amorphe et celle de matériau cristallin (figure 1.15c). Le phase cristalline conduisant mieux le courant électrique que la phase amorphe, la résistance du composant sera d'autant plus élevée que la part amorphe sera importante. Le rapport R_{OFF}/R_{ON} rencontré est typiquement de 2 à 3 ordres de grandeur ($R_{ON} \simeq 1\,k\Omega$ à $1\,M\Omega$).

Le caractère analogique et bien contrôlé de la transition entre les deux niveaux extrêmes de résistance électriques rend cette technologie particulièrement intéressante en termes de densité d'intégration, qu'il s'agisse d'en faire des synapses artificielles ou simplement des cellules mémoires multi-niveau. Par ailleurs, le courant de programmation présente l'attrait de diminuer avec l'aire de la cellule mémoire[76], sachant qu'il s'agit d'une technologie présentant un potentiel important de passage à l'échellle[134]. Enfin, les matériaux nécessaires ont déjà été l'objet d'études poussées[135,136].

Un inconvénient majeur des cellules mémoires à changement de phase est l'existence d'une

132. M. SURI et coll., IEEE International Electron Devices Meeting, 2012.
133. S. LA BARBERA, D. VUILLAUME et F. ALIBART, ACS Nano, 2015.
134. S. RAOUX et coll., IBM Journal of Research and Development, 2008.
135. G. W. BURR et coll., Journal of Vacuum Science & Technology B, 2010.
136. H. S. P. WONG et coll., Proceedings of the IEEE, 2010.

dérive au cours du temps vers l'état de forte résistance[76]. Il existe un compromis à faire entre la résolution de l'information stockée dans une cellule et la rétention à long terme de cette dernière.

Des simulations informatiques ont montré qu'il était possible d'obtenir un apprentissage sur une tâche vidéo à l'échelle d'un système utilisant des synapses artificielles analogiques à changement de phase munies d'une règle d'apprentissage de type *Spike-Timing Dependent Plasticity*[137,138]. Afin de réaliser la dépression synaptique, chaque synapse est constituée d'un couple de cellules à changement de phase, utilisé de façon différentielle pour coder le poids synaptique. L'inconvénient de cette approche est de diviser par deux la densité d'intégration synaptique.

Récemment, BURR et coll. ont réalisé la démonstration expérimentale d'un perceptron à 3 couches, utilisant 165 000 synapses également constituée d'une paire de cellules à changement de phase par synapse[139]. En utilisant une variante de l'algorithme de rétropropagation du gradient, les auteurs ont obtenu sur un sous-ensemble de la base MNIST des performances similaires à celles de solutions logicielles. Toutefois, la stratégie de codage différentiel des poids synaptiques utilise une étape de rafraîchissement pouvant potentiellement nuire à la vitesse ou la performance énergétique de l'ensemble.

1.4.4 Jonctions tunnel ferro-électriques

Ces composants mémristifs sont constitués de deux électrodes conductrices enserrant une fine barrière de matériau ferroélectrique (figure 1.15d), qui présente une polarisation électrique au repos, tel que le $BaTiO_3$ ou le $PbTiO_3$. Il est possible de moduler cette polarisation électrique par l'application d'une tension électrique sur la barrière[140].

Il a été fait démonstration de composants avec des dimensions latérales de 20 nm[141], ou encore présentant des rapport de résistance électrique OFF/ON de l'ordre de la centaine[142] (avec $R_{ON} \simeq 100\,k\Omega$). Une utilisation binaire n'est pas la seule envisageable, CHANTHBOUALA et coll. ayant par exemple montré la possibilité d'un usage plus analogique et tourné vers les architectures neuromorphiques[143]. LECERF et coll. ont également présenté un démonstrateur technologique de réseau de neurones artificiels exploitant de tels composants comme synapses capables d'apprendre[144].

L'un des inconvénients majeurs des cellules mémoires ferro-électriques usuelles est la nature destructive de leur lecture, qui nécessite de reprogrammer à l'identique une cellule venant

137. O. BICHLER et coll., IEEE Transactions on Electron Devices, 2012.
138. M. SURI et coll., International Electron Devices Meeting, 2011.
139. G. W. BURR et coll., IEEE Transactions on Electron Devices, 2015.
140. M. Y. ZHURAVLEV et coll., Physical Review Letters, 2005.
141. A. GRUVERMAN et coll., Nano Letters, 2009.
142. A. CHANTHBOUALA et coll., Nature nanotechnology, 2012.
143. A. CHANTHBOUALA et coll., Nature materials, 2012.
144. G. LECERF et coll., IEEE International Symposium on Circuits and Systems, 2014.

d'être lue[145]. Néanmoins, l'exploitation d'effets d'interface au niveau de la barrière peut permettre d'outrepasser cette limitation et de mettre en place une lecture non destructive, en exploitant notamment l'effet de résistance électrique tunnel géante[146]

Certains procédés de fabrication rapportés dans la littérature[142] font état d'étape à plus de 700 °C, ce qui pourrait constituer un défi pour co-intégrer ces composants avec une technologie CMOS.

1.4.5 Jonctions tunnel magnétiques à transfert de spin

Ces composants mémoires innovants relèvent du domaine de l'électronique de spin, qui exploite le moment magnétique intrinsèque des électrons en complément de leur charge électrique. Programmables en courant[147,148], ces nanocomposants sont des candidats sérieux pour constituer l'un de piliers des futures mémoires informatiques non volatiles[149], combinant notamment vitesses de lecture et d'écriture, courant de programmation pouvant passer à l'échelle et potentiel de co-intégration avec un procédé de fabrication CMOS conventionnel[150].

À la différence des autres technologies présentées précédemment, il s'agit de composants binaires offrant seulement deux états de résistance (figure 1.15e), ce qui peut sembler préjudiciable à une utilisation comme synapses capables d'apprendre. En utilisant la modélisation du comportement naturellement stochastique de ces nanocomposants présentée au chapitre 2, nous verrons dans le chapitre 3 que cette technologie mémoire est au contraire particulièrement adaptée à une tâche cognitive dynamique, à condition d'adopter une règle d'apprentissage probabiliste qui exploite ce comportement aléatoire intrinsèque.

1.4.6 Conclusions

Synapses analogiques, multi-niveaux ou binaires ? Les différentes candidats mémristifs pour réaliser des nanosynapses artificielles permettent d'obtenir des comportements multiples, allant d'une transition analogique à un simple comportement binaire, offrant au passage la possibilité d'une utilisation multi-niveaux. À l'échelle du composant, la tendance pour un usage synaptique était de favoriser les composants analogiques. Cependant les récents travaux théoriques sur l'utilisation de poids synaptiques de résolution très réduite ou les succès de réalisations matérielles utilisant également des synapses de faible résolution redonnent de l'attrait aux composants binaires ou faiblement multi-niveaux qui offriraient par ailleurs de meilleures caractéristiques.

145. D. S. JEONG et coll., Reports on Progress in Physics, 2012.
146. V. GARCIA et coll., Nature, 2009.
147. J. C. SLONCZEWSKI, Journal of Magnetism and Magnetic Materials, 1996.
148. L. BERGER, Journal of Magnetism and Magnetic Materials, 1996.
149. N. D. RIZZO et coll., IEEE Transactions on Magnetics, 2013.
150. X. FONG et coll., Proceedings of the IEEE, 2016.

Apprentissage probabiliste. Dans le cas d'un composant synaptique binaire, l'une des limitations potentielles est le recours à un générateur de nombres (pseudo-)aléatoires pour mettre en œuvre une règle d'apprentissage probabiliste, nécessaire pour ce type de composants. En effet, les circuits usuels de tels générateurs sont coûteux en surface de silicium, comme il est par exemple possible de le constater avec la puce neuromorphique TrueNorth, sans parler de la hausse de la consommation énergétique qu'ils peuvent entraîner.

Utiliser la physique des composants. Une solution au point précédent, suggérée par SURI et coll. est de chercher à mieux exploiter le comportement intrinsèque des nanocomposants. Certains offrent en effet de pouvoir exploiter des comportements stochastiques pour mettre en place la règle d'apprentissage probabiliste. Cependant, une barrière à ce type d'approche est généralement la mauvaise compréhension du phénomène physique à l'origine de ces comportements stochastiques, ce qui en rend la modélisation et l'exploitation plus complexe.

Plus généralement, plus les comportements riches mais contrôlables offerts par un nanocomposant synaptique sont exploités, meilleur est le potentiel d'intégration de ces composants. Ce type de stratégie peut néanmoins exiger de trouver un compromis entre les circuits de commandes associés à l'exploitation d'un comportement intrinsèque complexe, et l'influence de ces derniers sur la consommation énergétique, la surface de silicium ou encore la tolérance à la variabilité entre composants.

1.5 Bilan

Pour faire la synthèse de cet état des lieux des solutions neuro-inspirées, nous pouvons commencer par observer que les tâches cognitives sont un domaine en pleine expansion en informatique, et que les approches par réseaux de neurones artificiels semblent être une solution mieux adaptée pour les traiter que les approches algorithmiques plus conventionnelles, notamment en raison de leur robustesse aux données floues, bruitées ou incomplètes. Ces réseaux de neurones artificiels ont connus de multiples évolutions depuis leur formulation originelle, prenant aujourd'hui la forme de réseaux profonds, convolutifs ou impulsionnels.

L'une des caractéristiques notables des réseaux de neurones est de réaliser finalement peu de calcul en regard du nombre de transactions mémoires. Il s'agit par ailleurs de systèmes dont l'exécution des opérations est fortement parallèle, et cadencée lentement dans les cas biologiques. Il se trouve que l'architecture de VON NEUMANN des ordinateurs conventionnels est peu adaptée à l'exécution de réseaux de neurones artificiels, puisque leur conception s'appuie au contraire sur un faible nombre de cœurs de calcul dont l'exécution des opérations et des transactions mémoires est séquentielle, cadencée à haute fréquence pour en autoriser le multiplexage temporel. Ceci entraîne, notamment sur les problèmes de grande taille, une lenteur et une consommation énergétique importantes des réseaux de neurones logiciels. Des architectures matérielles dédiées à l'exécution de réseaux de neurones sont donc l'objet de recherches,

avec une tendance générale pour les solutions impulsionnelles du fait de bénéfices énergétiques mais également en raison de leur caractère biologiquement plus plausible.

L'une des difficultés majeures réside dans le nombre de connexions synaptiques *plastiques* qui explose avec celui des neurones, nécessitant de disposer de quantités importantes de mémoire, si possible « intelligente », c'est-à-dire permettant de réaliser tout ou partie de la plasticité synaptique requise en exploitant leur comportement physique. Le développement de nanocomposants mémoires innovants pour les usages synaptique est ainsi un domaine de recherches actuellement très actif, les candidats à l'étude étant nombreux. Si originellement, les synapses analogiques étaient l'objet de plus d'attention, de récentes avancées sur l'utilisation de synapses de résolution réduite donnent un potentiel nouveau à certaines réalisations technologiques. Par ailleurs, les composants présentant des comportements complexes peuvent se révéler d'autant plus utiles qu'ils permettent de réaliser intrinsèquement tout ou partie de l'apprentissage, offrant des avantages en termes de densité d'intégration ou encore d'efficacité énergétique

Dans les chapitres suivants, nous étudierons l'utilisation synaptique de deux familles de nanocomposants mémoires innovants. Tout d'abord les jonctions tunnel magnétiques à transfert de spin (présentées dans la partie 1.4.5), une technologie binaire mais offrant un comportement naturellement stochastique que nous chercherons, à la lumière des travaux mentionnés dans la partie 1.2.5, à exploiter pour mettre en œuvre une règle d'apprentissage probabiliste. Les cellules électrochimiques métalliques Ag_2S (cf. partie 1.4.2), qui constituent la seconde technologie mémristive étudiée, sont quant à elles des nanocomposants qui possèdent plusieurs comportements plastiques analogiques, intrinsèques et riches. Nous proposerons une stratégie tirant profit de ces derniers pour simplifier la mise en œuvre d'un apprentissage non supervisé par un système neuro-inspiré doté de telles synapses artificielles.

Chapitre 2

La jonction tunnel magnétique : une cellule mémoire stochastique

« Le hasard est plus docile qu'on ne pense. Il faut l'aimer. Et dès qu'on l'aime, il n'est plus hasard, ce gros chien imprévu dans le sommeil des jeux de quilles. »

René DAUMAL

“CE CHAPITRE *réinterprète le comportement naturellement stochastique des jonctions tunnel magnétiques de seconde génération comme une* fonctionnalité. *Les principaux phénomènes physiques régissant le comportement de ces nanocomposants binaires y sont décrits, avant de mettre l'accent sur la modélisation haut niveau de ces éléments, étape nécessaire à la conception de circuits et d'architectures électroniques. Un intérêt particulier est porté au délai précédant la commutation de ce type de jonctions, puisqu'il s'agit de la grandeur aléatoire intrinsèque qui sera exploitée par la suite pour effectuer une programmation synaptique probabiliste. Ce chapitre présente notamment un modèle analytique développé durant la thèse à l'aide de simulations physiques Monte-Carlo, rapide à simuler et capable de décrire la distribution statistique de ce délai sur l'ensemble des régimes de programmation possibles, palliant ainsi certaines lacunes des travaux existant dans la littérature.”*

Introduction

Dans ce chapitre, nous allons voir que les jonctions tunnel magnétiques à transfert de spin possèdent un comportement naturellement stochastique, intéressant pour en faire des synapses artificielles capables d'apprendre. En effet, si le caractère binaire de ces nanocomposants mentionnés dans la partie 1.4.5 les rendait initialement peu attractifs comme synapses artificielles au sein d'architectures neuro-inspirées, la situation évolue. Nous avons ainsi vu dans la partie 1.2.2 que des synapses artificielles de faible résolution ne sont pas préjudiciables aux performances d'inférence d'un réseau de neurones artificiels et peuvent améliorer durant l'apprentissage les capacités de généralisation de ce dernier. Par ailleurs, les travaux présentés dans la partie 1.2.5 ont montré la possibilité de réaliser avec succès un apprentissage sur puce qui exploite des nanocomposants mémoires innovants comme synapses artificielles binaires, à condition de recourir à une programmation probabiliste. Pour être intéressante, cette approche nécessite toutefois de disposer de générateurs de nombres aléatoires économes en énergie et en surface de silicium, ce qui demeure actuellement un défi. Nous avons par exemple vu dans la partie 1.3.4 qu'un quart des circuits constituant les neurones de la puce neuromorphique TrueNorth y est consacré.

Si le comportement aléatoire intrinsèque des jonctions tunnel magnétiques à transfert de spin est généralement considéré comme un défaut à combattre, il est néanmoins modélisé seulement de façon partielle dans les modèles analytiques de la littérature, malgré une origine physique bien connue. Dans ce chapitre, nous présentons un travail de modélisation haut niveau de ce comportement, qui s'appuie sur des simulations de la dynamique magnétique de ces nanocomposants. Ceci nous permettra par la suite de mettre en place dans le chapitre 3 une règle d'apprentissage probabiliste qui *exploite* le comportement stochastique de telles synapses artificielles plutôt que de reposer sur des circuits externes de génération de nombres aléatoires. Ceci est d'autant plus intéressant qu'il s'agit-là d'une technologie mémoire innovante ayant atteint la maturité industrielle : des puces de mémoire conventionnelles utilisant de jonctions tunnel magnétiques à transfert de spin sont actuellement mises sur le marché par des acteurs comme la société Everspin® [I].

2.1 Principes de base et technologie des jonctions tunnel magnétiques

2.1.1 Physique d'une jonction tunnel magnétique

Cette partie présente les grands principes du fonctionnement des jonctions tunnel magnétiques. Sont ainsi présentés de façon phénoménologique les deux effets physiques majeurs qui régissent le comportement de ces composants : la magnétorésistance tunnel et l'effet de transfert de spin.

Nous nous intéresserons ici aux jonctions tunnel magnétiques dites « de seconde généra-

I. https://www.everspin.com/news/everspin-256mb-st-mram-perpendicular-mtj-sampling

tion », pilotées en courant, et non plus en champ magnétique comme pour les jonctions de la « première génération ». Cette deuxième génération de composants permet en effet de passer à l'échelle, contrairement à l'ancienne génération pilotée en champ magnétique, et d'atteindre une énergie de programmation sensiblement plus faible dans le cas des nœuds technologiques avancés[151].

2.1.1.1 Structure de base

La structure schématique d'une jonction tunnel magnétique (MTJ[I]) est constituée de deux couches ferromagnétiques entourant une couche isolante (non ferromagnétique). Cette dernière est néanmoins suffisamment fine (typiquement entre 0,5 et 2 nm d'épaisseur) pour permettre le passage d'un courant électrique par effet tunnel à travers la structure.

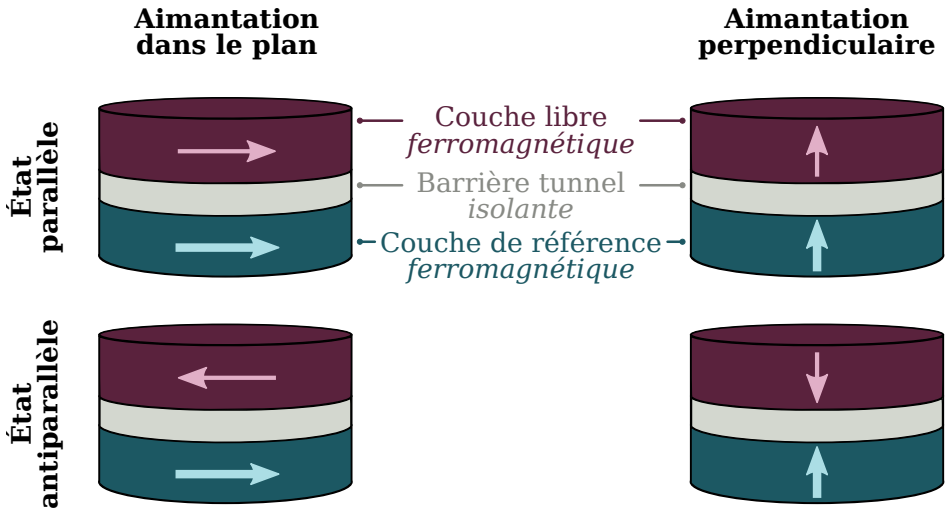


FIGURE 2.1 – Les trois principales couches d'une jonction tunnel magnétique. Les flèches représentent les différentes orientations stables de l'aimantation des couches ferromagnétiques. *À gauche* : une configuration où l'aimantation est « dans le plan ». *À droite* : une configuration où l'aimantation est « perpendiculaire » (ou « hors du plan »).

Comme cela est schématisé à la figure 2.1, le moment magnétique global de l'une des couches, dite « de référence », est orienté de façon fixe selon une direction et un sens donnés (généralement au moyen de couches antiferromagnétiques supplémentaires, non représentées sur le schéma en question), tandis que la seconde couche ferromagnétique est qualifiée de « libre », son moment magnétique global pouvant basculer entre deux orientations stables. Les deux configurations stables d'une jonction tunnel magnétique sont qualifiées de « parallèle » (P) et d'« antiparallèle » (AP) selon l'orientation relative du moment magnétique global de la couche libre et de celui de la couche de référence.

151. C. AUGUSTINE et coll., IEEE Sensors Journal, 2012.

I. De l'anglais *Magnetic Tunnel Junction*.

Il existe deux grandes familles de jonctions tunnel magnétiques, selon que leurs couches ferromagnétiques sont aimantées dans le plan ou bien de façon perpendiculaire à ce dernier, ce qui dépend de l'axe d'anisotropie magnétique dominant.

D'un point de vue énergétique, les deux configurations magnétiques stables d'une jonction tunnel magnétique forment un paysage en double puits. Les deux minimums sont de même énergie, comme schématisé à la figure 2.2. Pour basculer d'un puits à l'autre, la configuration magnétique doit franchir une barrière énergétique E_0 qui dépend de l'anisotropie magnétique et de la température du composant. Cette commutation peut être induite par simple activation thermique ou bien être le résultat d'agents extrinsèques, tels que le couple exercé par transfert de spin, que nous détaillerons dans la partie 2.1.1.3.

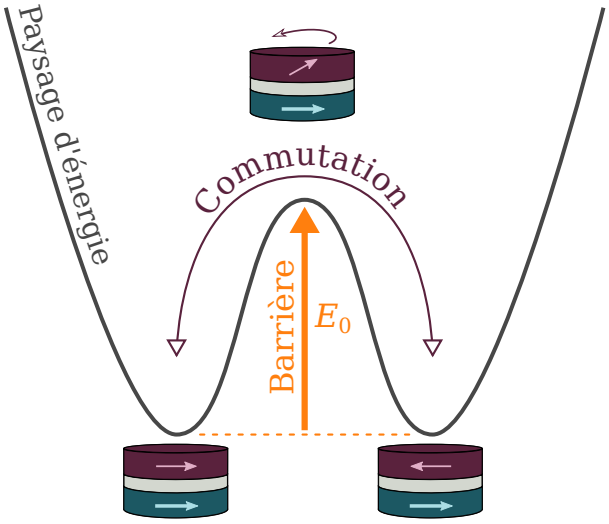


FIGURE 2.2 – Vision schématique du profil énergétique d'une jonction tunnel magnétique. Pour basculer d'un état stable à l'autre, la jonction doit franchir une barrière d'énergie E_0.

2.1.1.2 Magnétorésistance tunnel

Parce que l'aimantation d'une couche ferromagnétique est liée aux densités d'état de ses électrons de valence, le passage d'un électron par effet tunnel à travers la barrière isolante dépend du moment magnétique intrinsèque[I] de ce dernier et de l'orientation relative des couches ferromagnétiques de la jonction.

En faisant l'hypothèse que le spin des électrons est conservé durant leur traversée par effet tunnel[152] et parce que ces derniers possèdent deux orientations de spin, il est possible de considérer le courant électronique total circulant à travers une jonction tunnel magnétique comme la somme des contributions de deux canaux parallèles indépendants[153]. C'est ce que schématise

I. Que nous qualifierons simplement de « spin » par la suite.
152. S. YUASA et coll., Nature materials, 2004.
153. M. JULLIÈRE, Physics Letters A, 1975.

la figure 2.3. L'existence d'une aimantation non nulle au sein d'un matériau ferromagnétique est associée à une dissymétrie en spin des densités électroniques autour du niveau d'énergie de Fermi E_F. Les électrons dont le moment magnétique intrinsèque est parallèle à l'aimantation locale sont qualifiés de « majoritaires », alors que les électrons dont le moment magnétique est orienté de façon opposée sont quant à eux qualifiés de « minoritaires ». Lors de son passage par effet tunnel, un électron incident dont le moment magnétique est d'orientation opposée à celle de l'aimantation de la couche ferromagnétique d'arrivée aura une probabilité plus importante d'être réfléchi qu'un électron de moment magnétique parallèle à cette dernière, en raison de la plus faible densité d'états disponibles dans la première situation.

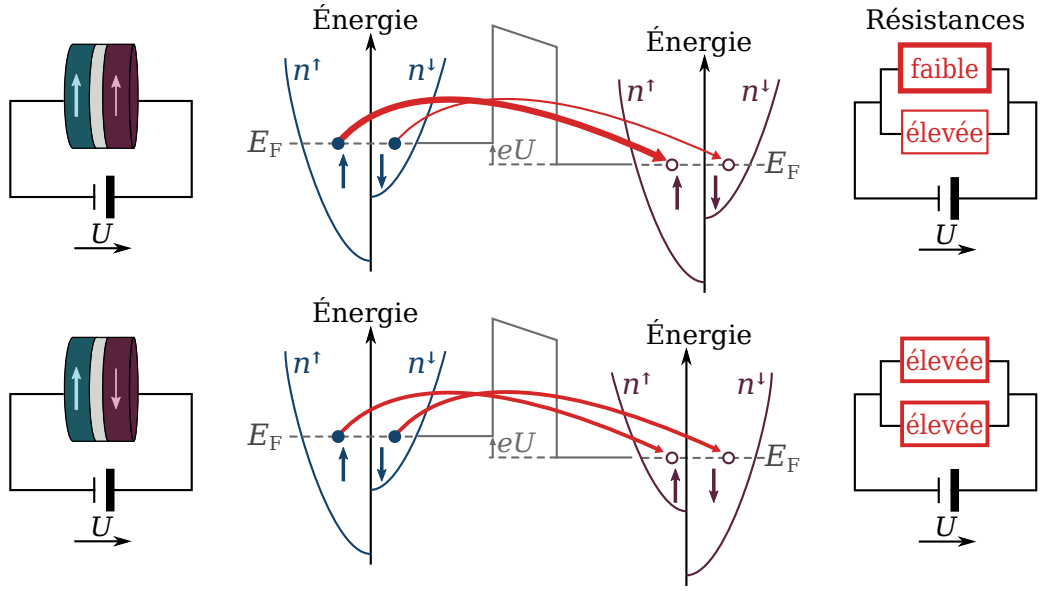


FIGURE 2.3 – Modèle à deux courants de la magnétorésistance tunnel. *À gauche* : la tension U appliquée sur la jonction tunnel magnétique. *Au centre* : une schématisation des bandes d'énergie associées aux deux canaux de conduction par effet tunnel. E_F désigne le niveau d'énergie de Fermi et e la charge électrique élémentaire tandis que $n^\uparrow$ et $n^\downarrow$ sont les densités d'états électroniques associées aux électrons de spin $\uparrow$ et $\downarrow$, les flèches faisant référence à la direction de l'aimantation. *À droite* : l'association de résistances équivalente aux deux canaux de conduction.
Dans le cas d'une configuration magnétique parallèle (*en haut*), les électrons de spin majoritaire d'une couche sont également de spin majoritaire dans la seconde couche, ce qui entraîne une probabilité de passage par effet tunnel importante pour ce canal, qui apparaît alors comme de résistance faible. À l'inverse, dans le cas d'une configuration magnétique antiparallèle (*en bas*) il n'existe pas de canal de conduction de faible résistance, entre deux densités d'états majoritaires : les électrons de spin majoritaire d'une couche ferromagnétique sont de spin minoritaire dans l'autre.

Dans le cas d'une jonction tunnel magnétique, lorsque les aimantations des deux couches ferromagnétiques sont parallèles, les électrons de spin majoritaire provenant d'une première couche ferromagnétique ont ainsi une probabilité plus importante d'atteindre par effet tunnel

la seconde couche ferromagnétique que les électrons de spin minoritaire. À l'inverse, lorsque les aimantations des deux couches sont antiparallèles, les électrons de spin minoritaire de la première couche ferromagnétique ont une probabilité plus élevée que les électrons de spin majoritaire d'atteindre par effet tunnel la seconde couche ferromagnétique. Le courant électronique est donc plus faible dans le second cas et la résistance électrique de la jonction tunnel magnétique plus élevée. Ce modèle simple permet d'expliquer pourquoi la résistance R_{AP} dans la configuration antiparallèle est plus importante que la résistance R_P dans la configuration parallèle[153].

La magnétorésistance tunnel (*TMR*[I]) est une métrique employée pour rendre compte de l'amplitude de ce phénomène. Elle est définie par

$$TMR = \frac{R_{AP} - R_P}{R_P}. \tag{2.1}$$

Plus la magnétorésistance tunnel est de valeur élevée, plus grande est la différence de résistance électrique entre les configurations parallèle et antiparallèle, et plus la lecture de l'état de la jonction tunnel magnétique est aisée. Expérimentalement, les valeurs *à température ambiante* sont de nos jours généralement autour de 100 à 200 % pour les technologies industriellement matures, et peuvent monter jusqu'à 600 % dans certaines réalisations académiques[154]. Ces valeurs sont sensiblement plus élevées que celles obtenues au sein de dispositifs exploitant la magnétorésistance géante[155] (GMR) ou la magnétorésistance anisotropique[156] (AMR), les effets les plus étudiés en électronique de spin avant la découverte de la magnétorésistance tunnel.

Le rapport R_{OFF}/R_{ON} des jonctions tunnel magnétiques (ici R_{AP}/R_P) est de l'ordre de quelques fois l'unité, ce qui est sensiblement plus faible que la plupart des autres nanocomposants mémoires émergents. Par ailleurs, une jonction tunnel magnétique dans son état antiparallèle peut difficilement être considérée comme « non passante » puisque sa résistance R_{AP} ne dépasse pas quelques dizaines de kiloohms, là où les autres technologies de mémoires émergentes présentées au chapitre 1 offrent des états non passants dont la résistance est supérieure de plusieurs ordres de grandeur. Si ces chiffres peuvent s'avérer suffisants dans le cadre d'un usage mémoire des jonctions tunnel magnétiques, ils annoncent à priori des défis accrus quant à une utilisation de ces composants dans le domaine de l'électronique neuromorphique.

Il est à noter que la magnétorésistance tunnel dépend de la tension appliquée aux bornes de la jonction, majoritairement du fait de la dépendance en tension de la résistance R_{AP} dans la configuration antiparallèle[152,158]. La figure 2.4a présente un modèle analytique de cette dépendance, généralement utilisé dans les modèles compacts de jonctions tunnel magnétiques[157]. La figure 2.4b montre quant à elle la caractéristique courant-tension associée. Ces courbes il-

I. De l'anglais *Tunnel MagnetoResistance*.

154. S. IKEDA et coll., Applied Physics Letters, 2008.
155. A. BARTHELEMY et coll., Journal of Magnetism and Magnetic Materials, 2002.
156. T. MCGUIRE et R. L. POTTER, IEEE Transactions on Magnetics, 1975.
158. S. ZHANG et coll., Physical Review Letters, 1997.
157. Y. ZHANG et coll., IEEE Transactions on Electron Devices, 2012.

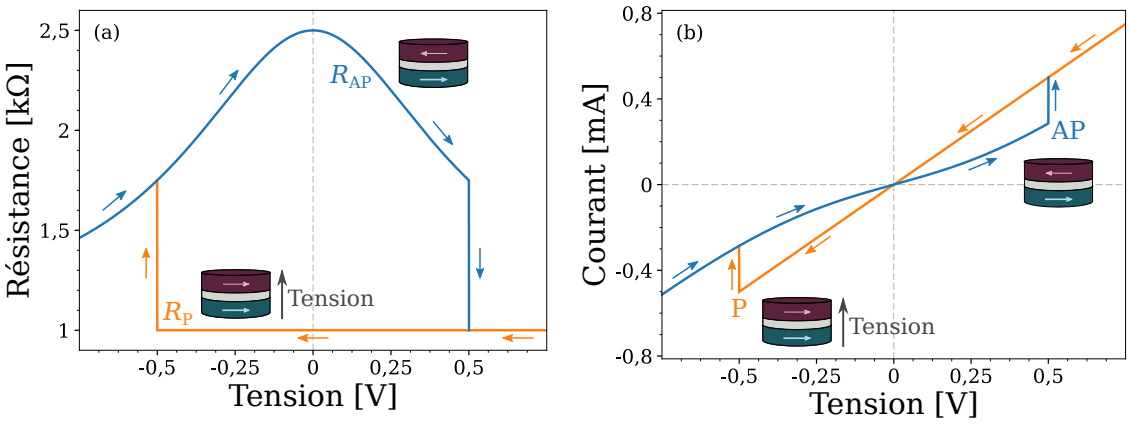
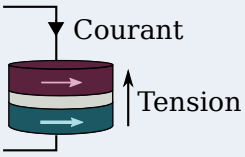


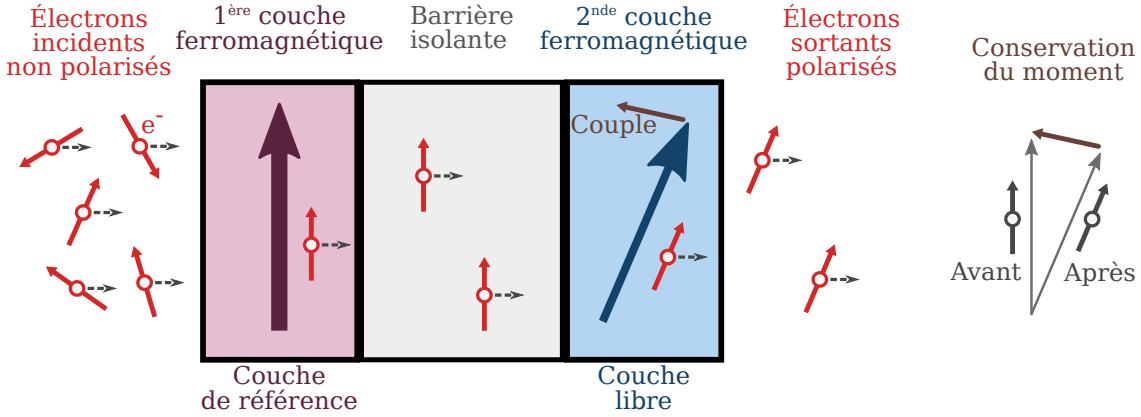
FIGURE 2.4 – Caractéristiques quasi-statiques *(a)* résistance-tension et *(b)* courant-tension d'une jonction tunnel magnétique, calculées à l'aide du modèle de la référence [157]. Les flèches indiquent le sens du cycle d'hystérésis, mais seules les transitions verticales sont réellement unidirectionnelles.

lustrent par ailleurs qu'il est possible de faire commuter la configuration d'une jonction tunnel magnétique en utilisant un courant électrique : un courant positif (en convention récepteur sur la figure) suffisamment élevé peut faire basculer la jonction d'une configuration antiparallèle à une configuration parallèle, tandis qu'un courant négatif d'amplitude assez forte peut faire l'inverse. Les détails de la physique et des comportements associés à la commutation résistive des jonctions tunnel magnétiques sont présentés dans la partie 2.2.2, mais nous pouvons d'ores et déjà remarquer que ce comportement n'est pas sans rappeler celui d'un composant mémristif (bipolaire), du fait du cycle d'hystérésis courant-tension pincé. Les caractéristiques particulières à la jonction tunnel magnétiques sont sa nature binaire (il n'y a pas d'état intermédiaire stable entre les configurations parallèle et antiparallèle), et le fait que le retournement soit stochastique, comme cela sera présenté plus en détails dans la partie 2.2.2.

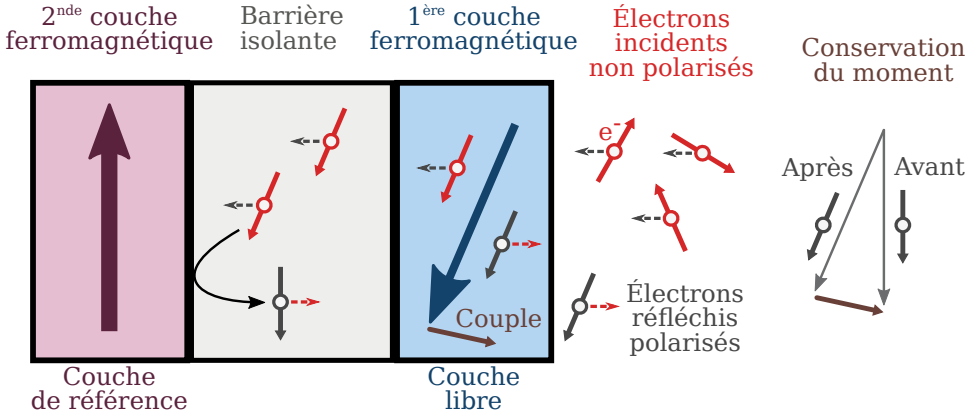
Conventions électriques



La convention récepteur adoptée à la figure 2.4 (reproduite ci-contre) est conservée pour l'intégralité du présent document. La tension appliquée aux bornes d'une jonction tunnel magnétique est donc la soustraction du potentiel électrique de l'électrode de la couche de référence à celui de l'électrode de la couche libre. Par ailleurs, le courant est compté positivement s'il rentre par la couche libre (celle du haut sur le schéma).



(a) Couple de transfert de spin exercé sur la couche ferromagnétique libre, dû à la traversée d'électrons (courant positif).



(b) Couple de transfert de spin exercé sur la couche ferromagnétique libre, dû à la réflexion d'électrons (courant négatif).

FIGURE 2.5 – *(a)* Dans le cas d'un courant positif circulant à travers la jonction tunnel magnétique (en convention récepteur sur la figure 2.4), la première couche ferromagnétique traversée par les électrons incidents est la couche de référence et joue le rôle de polariseur. Après leur passage par effet tunnel, les électrons voient leur moment magnétique se réaligner selon l'aimantation de la couche libre, ce qui se traduit en l'application d'un couple de transfert de spin (schématisé sur la droite de la figure).
(b) Lorsque le courant est de signe opposé, les électrons réfléchis par la couche de référence sont à l'origine du couple de transfert de spin exercé sur la couche libre. Le couple est question est alors de signe opposé à celui présenté en (a).
N.B. : l'épaisseur relative des différentes couches n'est pas à l'échelle.

2.1.1.3 Transfert de spin

Une explication phénoménologique du couple « de transfert de spin » (STT[I]) est schématisée à la figure 2.5. Chaque électron est porteur d'un moment magnétique, induit par son moment angulaire de spin. Les moments magnétiques des électrons de conduction s'alignent avec l'aimantation locale de la première couche ferromagnétique qu'il rencontrent, c'est-à-dire la couche de référence dans le cas d'un courant positif, qui se comporte comme un polariseur en spin. Cette polarisation en spin est conservée durant le passage par effet tunnel à travers la barrière isolante. La seconde couche ferromagnétique rencontrée, ici la couche libre, est alors soumise à l'incidence d'un courant d'électrons polarisés[152]. À mesure que les électrons traversent la couche libre, les moments magnétique portés par ces derniers voient leur composante transverse se modifier, les alignant avec l'aimantation de la couche libre. En raison de la conservation du moment cinétique, cette composante est transférée à l'aimantation de la couche libre, ce qui a pour conséquence d'exercer un couple sur cette dernière, comme schématisé sur la figure 2.5a. Ce couple est dit « de transfert de spin » et son amplitude est proportionnelle au flux d'électrons incidents, c'est-à-dire au courant traversant la jonction. Lorsque le courant est de signe opposé, la couche libre est également soumise à un couple de transfert de spin, de sens opposé au cas précédent et provenant cette fois des électrons réfléchis par la couche de référence. Cette situation est esquissée à la figure 2.5b.

Il est possible de trouver dans la littérature davantage de détails quant au phénomène de transfert de spin, par exemple dans la référence [159].

Il est à noter que si les moments magnétiques des deux couches ferromagnétiques d'une jonction tunnel magnétique sont parfaitement parallèles ou antiparallèles, alors quelle que soit l'amplitude du courant appliqué, le couple exercé par transfert de spin (figure 2.5) est nul. Un léger déséquilibre, généralement d'origine thermique, est donc nécessaire pour exploiter le couple de transfert de spin. Par ailleurs, nous pouvons retenir qu'un courant positif favorise la configuration magnétique parallèle (P), tandis qu'un courant négatif favorise à l'inverse la configuration magnétique antiparallèle (AP).

2.1.2 Usages possibles des jonctions tunnel magnétiques et considérations technologiques

Cette partie présente rapidement quelques usages possibles des jonctions tunnel magnétiques, conventionnels et moins conventionnels. Elle détaille également succinctement la technologie des jonctions tunnel magnétiques, du point de vue des matériaux ainsi que le futur potentiel d'intégration de cette technologie.

I. De l'anglais *Spin-Transfer Torque*.

2.1.2.1 Exemples d'usages

Au sein de puces de mémoire non volatile. Une endurance extrêmement élevée, une lecture et une écriture rapides, et un potentiel d'intégration important, le tout couplé à une technologie compatible avec les procédés de fabrication CMOS concourent à faire des jonctions tunnel magnétiques à transfert de spin des candidates de choix pour intégrer les nouvelles générations de mémoires non volatiles[150]. Des puces de mémoire vive magnétorésistive (MRAM [I]) basées sur ces nanocomposants commencent ainsi à arriver sur le marché industriel. Le leader de ce marché est la société Everspin®, qui produit déjà des puces de 64 Mib[149] et a récemment annoncé la disponibilité d'échantillons de puces de 256 Mib [II]. De nouveaux acteurs font également leur apparition, avec par exemple la société Avalanche Technology®, qui s'est lancée dans la fabrication de puces de mémoire vive magnétorésistive sur des galettes de 300 mm [III], ainsi que des entreprises de taille plus importante telles que Samsung® et IBM® [IV]. Si pour l'instant les marchés visés sont des niches industrielles, du fait de coûts de production encore élevés, les générations futures sont pressenties pour toucher à terme le marché du stockage de masse, d'un volume bien plus conséquent, à condition que leur densité d'intégration s'améliore (entraînant une baisse du coût de cette technologie) [V].

Vers des usages moins conventionnels. Les usages envisagés pour les jonctions tunnel magnétiques ne se limitent pas simplement à celui de puces de mémoire non volatile. Existent ainsi dans la littérature un certain nombre d'idées qui profitent par exemple des caractéristiques des jonctions tunnel magnétiques pour proposer de bousculer la hiérarchie mémoire classique, que ce soit pour développer de nouvelles fonctionnalités ou réduire la consommation énergétique. Il est par exemple question d'essayer de remplacer au sein des puces actuelles certains blocs mémoires, tels que des registres[160] ou encore une partie de la mémoire cache[161,162], par des jonctions tunnel magnétiques, dans le but d'exploiter leur caractère non volatile pour réduire de façon drastique la consommation statique, qui est un problème grandissant à mesure que le nœud technologique diminue. Certaines propositions exploitent ces nanocomposants pour concevoir des circuits logiques élémentaires de type *logic-in-memory*[163], potentiellement bénéfiques pour traiter simultanément des données volatiles et des données plus stationnaires. Un autre usage pressenti est la possibilité de mettre « instantanément » sous tension et hors ten-

150. X. FONG et coll., Proceedings of the IEEE, 2016.
I. De l'anglais *Magnetoresistive Random Access Memory*.
149. N. D. RIZZO et coll., IEEE Transactions on Magnetics, 2013.
II. https://www.everspin.com/news/everspin-256mb-st-mram-perpendicular-mtj-sampling
III. http://www.eetimes.com/document.asp?doc_id=1327122
IV. http://www.mram-info.com/ibm-demonstrated-11nm-stt-mram-junction-says-time-stt-mram-now
V. http://www.eetimes.com/author.asp?section_id=36&doc_id=1325705
160. Y. GANG et coll., IEEE Transactions on Magnetics, 2011.
161. S. SENNI et coll., IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 2016.
162. E. KITAGAWA et coll., IEEE International Electron Devices Meeting, 2012.
163. W. S. ZHAO et coll., IEEE Transactions on Circuits and Systems I : Regular Papers, 2014.

sion des circuits où les cellules de mémoire vive statique (SRAM[I]), volatiles, sont remplacées par des jonctions tunnel magnétiques. Cette stratégie peut notamment s'appliquer aux processeurs généralistes[164], mais également aux tables de correspondance des circuits logiques programmables (FPGA[II])[165].

2.1.2.2 Un composant intrinsèquement probabiliste

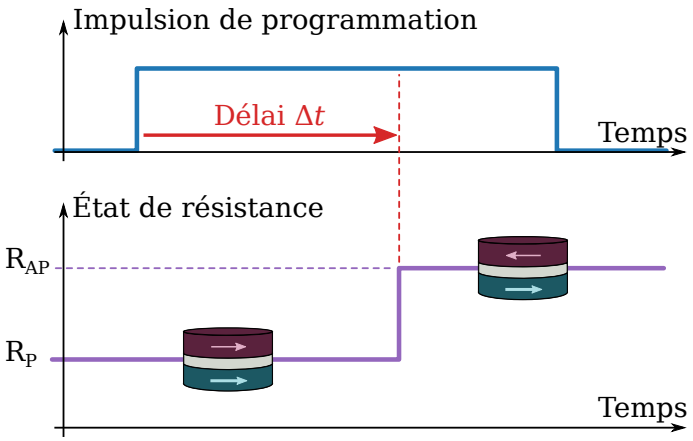


FIGURE 2.6 – Schéma du délai Δt avant la commutation d'une jonction tunnel magnétique à transfert de spin lors de l'application d'une impulsion de programmation.

Les usages précédents se focalisent sur le comportement statique des jonctions tunnel magnétiques, bien que ces composants possèdent également un comportement dynamique riche[166]. Notamment, lors de la programmation d'une jonction tunnel magnétique, le délai Δt précédant la commutation (schématisé à la figure 2.6) est une quantité aléatoire[167,168]. Là où les usages présentés précédemment s'efforcent généralement de contrecarrer les éléments aléatoires, certains travaux cherchent au contraire à tirer parti de cette caractéristique des jonctions tunnel magnétiques pour en faire des générateurs de nombres aléatoires[169,170]. Il existe par ailleurs dans la littérature des travaux qui reposent sur l'hypothèse d'un accès aisé à un grand nombre de sources de nombres aléatoires[171]. Cette hypothèse deviendrait réaliste avec des générateurs de nombres aléatoires de dimensions nanométriques réalisés à l'aide de jonctions tunnel magnétiques.

Nous verrons dans la suite de ce chapitre que la trajectoire du retournement de l'aimantation de la couche libre d'une jonction tunnel magnétique dépend du courant injecté. Ce faisant,

I. De l'anglais *Static Random-Acces Memory*.
164. B. JOVANOVIĆ, R. M. BRUM et L. TORRES, Analog Integrated Circuits and Signal Processing, 2014.
II. De l'anglais *Field-Programmable Gate Array*.
165. W. S. ZHAO et coll., ACM Transactions on Embedded Computing Systems, 2009.
166. N. LOCATELLI, V. CROS et J. GROLLIER, Natural Materials, 2014.
167. Y. HIGO et coll., Applied Physics Letters, 2005.
168. R. HEINDL et coll., Journal of Applied Physics, 2011.
169. A. FUKUSHIMA et coll., Applied Physics Express, 2014.
170. A. FUKUSHIMA et coll., IEEE Magnetics Conference, 2015.
171. J. S. FRIEDMAN et coll., IEEE Transactions on Circuits and Systems I : Regular Papers, 2016.

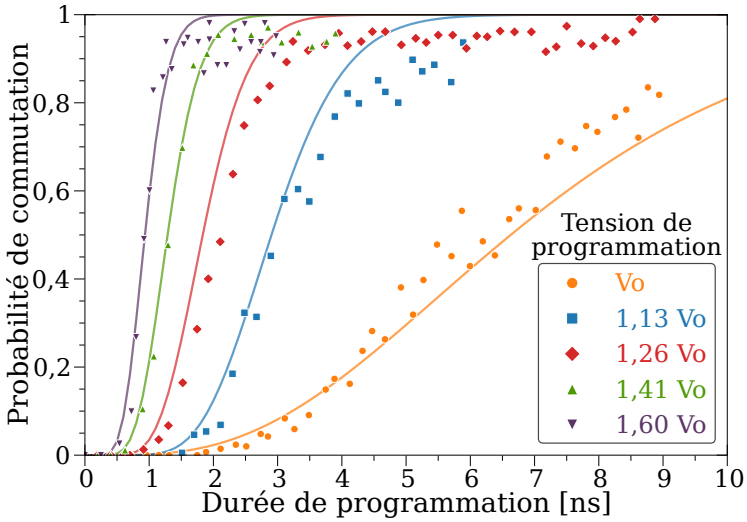


FIGURE 2.7 – Dépendance, avec la tension de programmation appliquée, de la fonction de répartition du délai avant commutation d'une jonction tunnel magnétique. *Symboles* : relevés expérimentaux basés sur les travaux [172, 173]. *Lignes* : ajustements du modèle analytique présenté dans ce chapitre.

une modification des signaux électriques appliqués sur une jonction tunnel magnétique permet de moduler la distribution statistique du délai avant commutation. C'est ce que présente la figure 2.7, où la fonction de répartition du délai (c'est-à-dire la probabilité de ce dernier d'être inférieur à la durée de l'impulsion de programmation) est modifiée selon la tension de programmation employée.

D'autres travaux cherchent à exploiter un phénomène de « résonance stochastisque » en utilisant des signaux électriques d'amplitude extrêmement réduite et des jonctions tunnel magnétiques qui présentent une barrière énergétique faible[174]. Dans notre cas, l'idée est davantage de pouvoir implémenter avec des jonctions plus conventionnelles et à priori pour n'importe quel régime de courant, une règle d'apprentissage probabiliste, en s'inspirant des travaux de la référence [88], sans toutefois nécessiter le recours à un générateur de nombres aléatoires extrinsèque aux éléments synaptiques. Une façon d'y arriver sera présentée dans la partie 2.2.

2.1.2.3 Technologie et potentiel de passage à l'échelle

En termes de matériaux employés, les couches ferromagnétiques sont le plus souvent constituées d'un alliage CoFeB. La barrière isolante était quant à elle à l'origine généralement réalisée en oxyde d'aluminium mais ce matériau est désormais remplacé par de l'oxyde de magnésium (cristallin), ce qui permet d'augmenter sensiblement les valeurs de magnétorésistance tunnel[175–177]. La couche de référence est en général couplée à une troisième couche ferro-

174. A. MIZRAHI et coll., arXiv preprint arXiv :1605.06932, 2016.
175. S. IKEDA et coll., Japanese Journal of Applied Physics, 2005.
176. D. DJAYAPRAWIRA et coll., Applied Physics Letters, 2005.
177. J.-G. ZHU et C. PARK, Materials Today, 2006.

magnétique en CoFe placée au sein d'une structure antiferromagnétique de synthèse (SAF[I]) CoFeB/Ru/CoFe, dont le but est d'annuler l'influence du champ magnétique dipolaire de la couche de référence sur la couche libre. Cette structure antiferromagnétique de synthèse est elle-même fixée par l'ajout d'une couche antiferromagnétique PtMn, qui complète la pile standard d'une jonction tunnel magnétique. Un exemple d'un tel empilement est présenté à la figure 2.8.

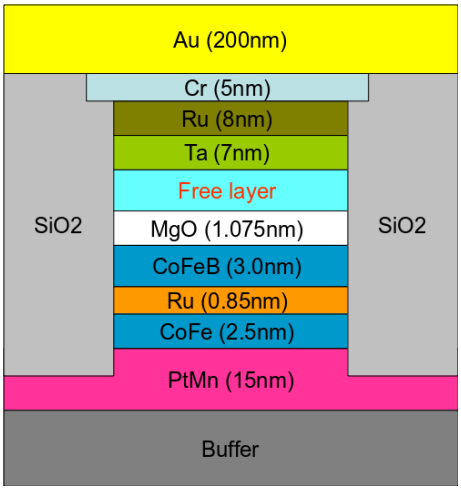


FIGURE 2.8 – Exemple de constitution d'une technologie de jonctions tunnel magnétiques réelles, notamment utilisée dans le cadre des mesures expérimentales de la référence [178]. La situation est donc sensiblement plus complexe que ce peut laisser imaginer la figure 2.1 page 61.

L'une des propriétés intéressantes des jonctions tunnel magnétiques est leur compatibilité avec les procédés de métallisation finale (BEOL[II]) au sein d'un processus de fabrication CMOS. Ces structures peuvent notamment supporter un recuit à 350 °C [179,180]. Du point de vue de la réalisation technologique, une étape exigeante est la gravure des jonctions, qui ne doit pas endommager la barrière isolante.

Le potentiel d'intégration des jonctions tunnel magnétiques à transfert de spin (STT-MTJ) a été démontré par plusieurs réalisations qui utilisent ces éléments au sein de puces mémoires standards. Comme mentionné auparavant, la société Everspin® a par exemple déjà lancé sur le marché une puce *standalone* de 64 Mib dans une technologie 90 nm[149,181]. Des travaux sur des puces similaires ont également été publiés par différents groupes de recherche [182–184]. Du côté de l'électronique embarquée, Toshiba® a notamment présenté une puce mémoire de 1 Mib

I. De l'anglais *Synthetic Anti-Ferromagnet*.
II. De l'anglais *Back-End-Of-Line*.
179. S. IKEDA et coll., Nature materials, 2010.
180. K. IKEGAMI et coll., IEEE International Electron Devices Meeting, 2014.
181. T. ANDRE et coll., Custom Integrated Circuits Conference, IEEE, 2013.
182. R. BEACH et coll., IEEE International Electron Devices Meeting, 2008.
183. S. CHUNG et coll., IEEE International Electron Devices Meeting, 2010.
184. D. C. WORLEDGE et coll., IEEE International Electron Devices Meeting, 2010.

dans une technologie 65 nm[185].

Contrairement à la mémoire flash, ainsi qu'à d'autres alternatives de mémoires non volatiles, les jonctions tunnel magnétiques à transfert de spin utilisent des tensions de programmation inférieures ou égales aux tensions logiques. Par ailleurs, les courants de programmation passent à l'échelle avec le nœud technologique[186] allant du milliampère à moins d'une dizaine de microampères[187]. Les réalisations récentes utilisent des structures à anisotropie magnétique perpendiculaire, ce qui réduit l'amplitude du courant de programmation nécessaire[162,188,189], ainsi que les contraintes lithographiques. Pour fixer les idées, prenons l'exemple des réalisations de Toshiba® [162,185] dans une technologie 30 nm à anisotropie magnétique perpendiculaire : la tension de programmation est de 0,6 V, le courant de programmation associé vaut quant à lui 50 µA et la durée des impulsions de programmation est de 3 ns seulement.

Il est à noter que les circuits de lecture et d'écriture de jonctions tunnel magnétiques à transfert de spin se sont fortement développés au cours des dernières années. Les circuits de lecture utilisent notamment des amplificateurs de détection (*sense amplifiers* en anglais) spécifiquement conçus pour cette technologie mémoire[185,190], tandis que les circuits d'écriture récents tentent de réduire le caractère aléatoire de la programmation par des techniques *self-enabled*[191].

2.2 Modélisation haut niveau du comportement stochastique d'une jonction tunnel magnétique à transfert de spin

2.2.1 Encore un autre modèle ?

La conception de circuits électroniques fait un usage important de modèles compacts, basés idéalement sur des équations analytiques afin de permettre des durées de simulation raisonnables. Des efforts conséquents ont ainsi été réalisés pour fournir des modèles compacts décrivant le comportement des jonctions tunnel magnétiques[157,192–199], notamment la nouvelle génération à transfert de spin (STT-MTJ), puisque ces composants sont pressentis pour être les cellules élémentaires de nouvelles mémoires non volatiles[200,201].

185. H. NOGUCHI et coll., Symposium on VLSI Technology, 2013.
186. Z. DIAO et coll., Journal of Physics : Condensed Matter, 2007.
187. J. J. NOWAK et coll., IEEE Magnetics Letters, 2016.
188. J. H. KIM et coll., Symposium on VLSI Technology : Digest of Technical Papers, 2014.
189. L. THOMAS et coll., Journal of Applied Physics, 2014.
190. W. S. ZHAO et coll., IEEE Transactions on Magnetics, 2009.
191. Y. LAKYS et coll., IEEE Transactions on Magnetics, 2012.
192. S. LEE et coll., Japanese Journal of Applied Physics, 2005.
193. M. E. BARAJI et coll., Journal of Applied Physics, 2009.
194. W. GUO et coll., Journal of Physics D : Applied Physics, 2010.
195. S. LEE et coll., Solid-State Electronics, 2010.
196. W. S. ZHAO et coll., Nanoscale Research Letters, 2011.
197. Y. ZHANG et coll., IEEE Transactions on Magnetics, 2013.
198. P. WANG et coll., IEEE/ACM International Conference on Computer-Aided Design, 2012.
199. G. D. PANAGOPOULOS, C. AUGUSTINE et K. ROY, IEEE Transactions on Electron Devices, 2013.
200. C. CHAPPERT, A. FERT et F. N. VAN DAU, Nature materials, 2007.
201. T. KAWAHARA et coll., Microelectronics Reliability, 2012.

Parmi les modèles compacts de jonctions tunnel magnétiques, différentes approches coexistent. Tout d'abord, certains modèles résolvent directement les équations dynamiques décrivant le comportement magnétique de ces composants[193,194,198,199]. Si elle permet une description temporelle très précise, cette approche présente l'inconvénient d'être particulièrement coûteuse en temps de calcul et ne convient pas à des simulations impliquant un grand nombre de jonctions tunnel magnétiques. À l'inverse, d'autres modèles s'appuient sur des équations analytiques plus simples afin d'accélérer les simulations. Malheureusement, les modèles analytiques de haut niveau de la commutation des jonctions tunnel magnétiques à transfert de spin disponibles dans la littérature[202–204] ne permettent pas de décrire le comportement de ces composants dans tous les régimes de courant de programmation possibles, ni de basculer de l'un à l'autre de façon réaliste. En effet, il existe un régime intermédiaire de courant, au sein duquel la physique de la commutation magnétique s'avère complexe[186,205] et qu'aucun des modèles précédemment mentionnés ne permet de décrire de façon satisfaisante. Le modèle de Sun notamment, utilisé pour les régimes de plus fort courant, présente une divergence mathématique dans cette zone de courant intermédiaire. Pour cette raison, les modèles analytiques actuels tendent à ignorer ce régime intermédiaire[157,196,197], nonobstant le fait que ce dernier soit particulièrement pertinent en termes d'applications[186].

Par ailleurs, le délai avant commutation Δt d'une jonction tunnel magnétique est une quantité aléatoire, comme cela a été mentionné dans la partie 2.1.2.2. Or les modèles actuels font généralement usage d'une description simplifiée (parfois à l'extrême) de la distribution de probabilité de ce délai avant commutation[197,206,207]. Ceci peut être un problème pour étudier la fiabilité de circuits mémoires conventionnels, ou pour concevoir des circuits *exploitant* la nature stochastique de la commutation des jonctions tunnel magnétiques, qu'il s'agisse de réduire la consommation énergétique par une programmation probabiliste[207] ou de les utiliser comme synapses binaires soumises à une règle d'apprentissage probabiliste[88]. Une modélisation suffisamment précise du délai avant commutation de ces nanocomposants est souhaitable lors de la conception de ce type de circuits.

Dans cette partie, nous présentons un modèle analytique qui a été développé durant la thèse pour décrire la distribution statistique du délai avant commutation Δt de jonctions tunnel magnétiques à aimantation dans le plan. Les principaux intérêts de ce nouveau modèle sont le fait d'être beaucoup plus rapide à simuler que des simulations plus proches de la physique tout en procurant une description statistique réaliste, ainsi que le fait d'assurer la continuité entre les modèles existant dans la littérature pour les régimes de courant extrêmes, incompatibles entre eux.

202. L. NÉEL, Annales de géophysique, 1949.
203. W. F. BROWN, Physical Review, 1963.
204. J. Z. SUN, Physical Review B, 2000.
205. D. BEDAU et coll., Applied Physics Letters, 2010.
206. Y. ZHANG et coll., IEEE Conference on Nanotechnology, 2013.
207. W. WU et coll., IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 2012.

2.2.2 Bases analytiques

2.2.2.1 L'équation de Landau-Lifschitz-Gilbert-Slonczewski

Une approximation usuelle pour décrire la dynamique d'une jonction tunnel magnétique est de supposer que le moment magnétique total de la couche libre se comporte comme un « macrospin » $\mathbf{m_f}$, présentant un processus de retournement cohérent entre les deux configurations stables (parallèle ou antiparallèle au moment magnétique total de la couche de référence). Si cette hypothèse décrit assez fidèlement le comportement des jonctions tunnel magnétiques dont l'aimantation est dans le plan, ceci est moins avéré dans le cas des jonctions à anisotropie magnétique perpendiculaire. Pour ces dernières, les prédictions d'un modèle macrospin perdent en précision lorsque les dimensions latérales de la jonction dépassent la cinquantaine de nanomètres. Au-delà de cette limite, il est souhaitable d'inclure les effets de retournement par nucléation et de sous-volumes magnétiques si l'on veut pouvoir rendre compte des écarts constatés expérimentalement avec les prédictions de retournement d'un modèle macrospin[208].

Le mouvement de précession du macrospin $\mathbf{m_f}$ peut être décrit à l'aide de l'équation de Landau-Lifshitz-Gilbert, à laquelle sont ajoutés d'autres termes pour rendre compte de l'agitation thermique ainsi que du couple de transfert de spin (STT) :

$$\frac{\mathrm{d}\mathbf{m_f}}{\mathrm{d}t} = \underbrace{-|\gamma|\mu_0 \mathbf{m_f} \times (\mathbf{H_{eff}} + \mathbf{h_{sto}})}_{\text{Précession}} + \underbrace{\frac{\alpha}{M_s V} \mathbf{m_f} \times \frac{\mathrm{d}\mathbf{m_f}}{\mathrm{d}t}}_{\text{Amortissement}} + \underbrace{V \boldsymbol{T}_S}_{\text{STT}} \tag{2.2}$$

Le champ effectif $\mathbf{H_{eff}}$ regroupe les différents termes d'anisotropie magnétique, ainsi qu'un possible champ extérieur. Dans ce travail, le champ extérieur est supposé nul puisque nous nous focaliserons sur des jonctions tunnel magnétiques pilotées électriquement et dont le retournement est effectué en ayant recours au couple de transfert de spin. Les paramètres γ, μ_0 et V sont respectivement le rapport gyromagnétique d'un électron, la perméabilité magnétique du vide et le volume de la couche libre de la jonction. Les paramètres décrivant les matériaux sont l'aimantation magnétique à saturation M_s ainsi que le coefficient d'amortissement de Gilbert α (sans unité). Pour modéliser les effets de l'agitation thermique sur le macrospin $\mathbf{m_f}$, un terme aléatoire de Langevin $\mathbf{h_{sto}}$ est ajouté, dont les coordonnées cartésiennes suivent des processus aléatoires gaussiens indépendants, sans corrélation et chacun de moyenne nulle [186,203,209].

Dans l'équation (2.2) précédente, le terme $\boldsymbol{T}_S$ rend compte du couple de transfert de spin de Slonczewski[147,186,210]

$$\boldsymbol{T}_S = -\frac{\mu_B}{|e|} \times \frac{J_s}{M_s^2 t_f} \times P \times \eta(\mathbf{m_f}, \hat{\mathbf{u}}_{inc}) \times \mathbf{m_f} \times (\mathbf{m_f} \times \hat{\mathbf{u}}_{inc}), \tag{2.3}$$

où $|e|$, μ_B et t_f désignent respectivement la charge électrique élémentaire, le magnéton de Bohr

208. J. Z. Sun et coll., Physical Review B, 2011.
209. J. L. García-Palacios et F. J. Lázaro, Physical Review B, 1998.
147. J. C. Slonczewski, Journal of Magnetism and Magnetic Materials, 1996.
210. M. D. Stiles et J. Miltat, Spin Dynamics in Confined Magnetic Structures III, 2006.

et l'épaisseur de la couche libre. La densité du courant traversant la jonction selon sa normale est notée J_s. Le terme P est la polarisation en spin du courant, lié à la magnétorésistance tunnel selon la relation $TMR = 2P^2/(1-P^2)$. Le terme $\eta(\mathbf{m_f}, \hat{\mathbf{u}}_{\text{inc}})$ décrit l'efficacité angulaire du transfert de spin. Nous considérerons une valeur unitaire constante, à la manière de plusieurs travaux de la littérature [186,211]. L'amplitude du couple de Slonczewski est proportionnelle à celle de la densité du courant. Le terme de couple de transfert de spin qualifié de *field-like*, de plus petite amplitude que celui de Slonczewski [I], est négligé.

Comme cela a été déjà mentionné auparavant, des études poussées ont montré un bon accord expérimental avec les prévisions issues d'un modèle macrospin, bien que certains écarts aient néanmoins pu être observés dans le cas des jonctions à aimantation hors du plan, notamment dans le régime des faible probabilités de retournement[208].

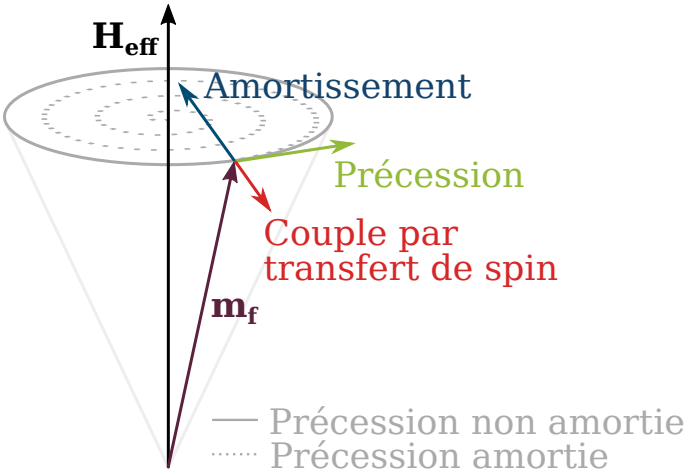


FIGURE 2.9 – Schéma général de la trajectoire suivie par un moment magnétique $\mathbf{m_f}$ en présence d'un champ effectif $\mathbf{H_{eff}}$ stationnaire (pour plus de clarté, le terme stochastique de Langevin $\mathbf{h_{sto}}$ est ici supposé nul). L'ellipse en trait plein présente un cas sans amortissement, tandis que la spirale en pointillé décrit un cas analogue en présence d'amortissement (positif). Trois flèches colorées indiquent les directions des trois composantes de couple qui apparaissent dans l'équation (2.2). Il faut en outre noter que le couple de transfert de spin peut être dans le sens opposé, selon le signe du courant. *N.B.* : les différentes flèches ne sont pas à l'échelle.

La figure 2.9 schématise l'effet de chacune des trois différentes composantes présentées dans l'équation (2.2) :

- le couple de précession, à cause duquel le moment magnétique $\mathbf{m_f}$ a tendance à décrire une trajectoire cyclique autour du champ effectif $\mathbf{H_{eff}}$;
- le couple d'amortissement, caractérisé par le coefficient de Gilbert α, à l'origine de la relaxation du moment magnétique $\mathbf{m_f}$ vers le champ effectif $\mathbf{H_{eff}}$;

211. J. Z. SUN, IBM Journal of Research and Development, 2006.

I. Au moins inférieur d'un facteur trois dans la littérature[212].

– le couple de transfert de spin, qui peut être interprété, selon le signe du courant qui traverse la jonction, comme une composante s'opposant à l'amortissement ou au contraire renforçant ce dernier.

Nous pouvons donc constater que le couple de transfert de spin est l'effet physique qui permet d'agir sur l'aimantation de la couche libre d'une jonction tunnel magnétique par une simple injection de courant à travers la jonction. Si le courant est injecté dans le sens adéquat et d'une amplitude suffisante pour contrebalancer l'amortissement, il devient possible au moment magnétique $\mathbf{m_f}$ de basculer vers la seconde configuration magnétique stable. Par ailleurs, à la suite de la commutation, si le courant est maintenu, ce dernier agit alors comme un amortissement supplémentaire.

Remarque

Il est à noter que le moment magnétique de la couche de référence est soumis à des effets similaires, mais de bien moindre amplitude en raison de l'ingénierie des couches employées pour fixer son orientation magnétique. Ceci permet de considérer qu'il n'est pas possible d'observer son retournement lors de l'application des impulsions de programmation.

2.2.2.2 Une valeur de densité de courant critique pour le renversement de l'aimantation

Comme cela a été présenté dans la partie 2.2.2.1, et à condition de bien choisir le signe du courant injecté, le couple de transfert de spin peut s'opposer à l'amortissement de la trajectoire du moment magnétique $\mathbf{m_f}$ de la couche libre. La densité de courant critique[I] J_{c0} peut être définie formellement comme la densité de courant, injectée à une température T de 0 K , pour laquelle le moment magnétique $\mathbf{m_f}$ de la couche libre est à la limite de la stabilité[204].

Physiquement, en négligeant pour le moment les fluctuations thermiques et en considérant une situation initiale du moment magnétique $\mathbf{m_f}$ légèrement hors d'équilibre, deux situations peuvent être rencontrées selon l'amplitude de la densité de courant injectée J_s :

$|J_s| < J_{c0}$: le couple d'amortissement surpasse celui dû au transfert de spin et la trajectoire du moment magnétique $\mathbf{m_f}$ est amortie vers la position initialement stable de ce dernier (comme le cas en pointillé de la figure 2.5a page 66) ;

$|J_s| > J_{c0}$: le couple de transfert de spin surpasse l'amortissement, déstabilisant l'état d'équilibre actuel. Le rayon de précession du moment magnétique $\mathbf{m_f}$ s'accroît jusqu'au basculement de ce dernier vers la configuration magnétique opposée, qui devient alors le nouvel état d'équilibre.

Dans le cas d'une jonction tunnel magnétique dont l'aimantation est dans le plan, l'amplitude de la densité de courant critique s'exprime [204,211], pour une transition de la configuration

I. Ou plus précisément l'amplitude de cette dernière.

parallèle à la configuration antiparallèle :

$$J_{c0}^{P\to AP} = \frac{2|e|}{\hbar} \times \frac{1+P^2}{P} \times \alpha\, t_f \mu_0 M_s \left(H_k + \frac{M_s}{2} \right), \tag{2.4}$$

et pour une transition opposée, de la configuration antiparallèle à la configuration parallèle :

$$J_{c0}^{AP\to P} = \frac{2|e|}{\hbar} \times \frac{1-P^2}{P} \times \alpha\, t_f \mu_0 M_s \left(H_k + \frac{M_s}{2} \right), \tag{2.5}$$

où $|e|$, $\hbar$ et t_f désignent respectivement la charge électrique élémentaire, la constante de Planck réduite et l'épaisseur de la couche libre. Le terme H_k désigne l'amplitude du champ d'anisotropie. Il s'agit de la somme des contributions de différentes anisotropies, par exemple d'un terme d'anisotropie magnétique lié aux matériaux tel que l'anisotropie magnétocristalline, et d'un terme d'anisotropie de forme provenant de la géométrie du composant, notamment dans le cas d'une jonction dont la section cylindrique n'est pas circulaire.

Les équations correspondantes pour le cas d'une jonction tunnel magnétique à anisotropie magnétique perpendiculaire peuvent être trouvées dans la référence [213].

L'existence de deux équations distinctes, (2.4) et (2.5) selon la transition considérée, pour l'amplitude de la densité de courant critique J_{c0} provient de la dépendance angulaire du transfert de spin. Cependant, il est montré dans la littérature[214,215] et expérimentalement[180] que ces deux valeurs correspondent à la même tension critique $V_{c0} = J_{c0}^{P\to AP} R_P = J_{c0}^{AP\to P} R_{AP}$ (comme cela est présenté sur la figure 2.4).

2.2.3 Simulations numériques du retournement de la couche libre

Simuler la trajectoire du moment magnétique $\mathbf{m_f}$ de la couche libre d'une jonction tunnel magnétique dans le cadre du modèle de retournement cohérent décrit précédemment est une façon simple d'obtenir une valeur du délai avant commutation Δt. Cette partie présente rapidement l'approche Monte-Carlo adoptée pour obtenir une estimation de la distribution statistique de ce délai avant commutation Δt. Cette estimation servira ensuite dans la partie 2.2.4 à élaborer un modèle analytique qui soit plus rapide à simuler que les simulations Monte-Carlo précédentes, ceci afin de permettre notamment l'étude de systèmes impliquant un nombre conséquent de jonctions tunnel magnétiques, tout en conservant un degré de réalisme physique satisfaisant quant à la grandeur qui nous intéresse, à savoir le délai avant commutation Δt.

2.2.3.1 Dispositifs simulés

Nous visons des applications réalistes des jonctions tunnel magnétiques, sans dispositif de refroidissement complexe. Les simulations sont donc réalisées à température ambiante (300 K).

En sus de nous placer dans une hypothèse de programmation à courant stationnaire, nous

214. K. BERNERT et coll., Physical Review B, 2014.
215. J. Z. SUN, Proceedings of the SPIE, 2016.

supposons également qu'aucun champ magnétique extérieur n'affecte le moment magnétique $\mathbf{m_f}$ de la couche libre, ce qui revient à supposer les aspects technologiques suffisamment bien maîtrisés pour être capables d'écranter totalement le champ magnétique dipolaire de la couche de référence.

Si la tendance actuelle des applications mémoires est de privilégier les jonctions tunnel magnétiques à anisotropie perpendiculaire[157,188,216] pour des raisons de meilleur passage à l'échelle de la consommation et de la durée de rétention, ainsi que de la densité d'intégration[150], les simulations présentées ci-après se focalisent au contraire sur des dispositifs à aimantation dans le plan. Non que les avantages des jonctions à anisotropie perpendiculaire ne soient pas probants, mais plutôt parce qu'il n'est pas évident que les applications neuromorphiques nécessitent le même cahier des charges que les applications mémoires conventionnelles, notamment en terme de durée de rétention. Or à volume constant, la rétention mémoire peut être modifiée simplement en changeant le rapport d'aspect des jonctions à aimantation dans le plan[I], alors que les dispositifs à anisotropie perpendiculaire nécessitent d'ajuster l'anisotropie d'interface des matériaux employés[150], ce qui est potentiellement moins aisé. Par ailleurs, il n'y a à priori pas d'obstacle majeur à l'adaptation de la méthodologie présentée ci-après (et du modèle analytique en découlant) au cas des dispositifs à anisotropie perpendiculaire.

Les dimensions latérales des jonctions simulées (table 2.1 page ci-contre) les rangent dans une technologie du nœud 45 nm.

2.2.3.2 Méthodologie

L'équation (2.2) est résolue numériquement en utilisant un schéma à point milieu[217,218], pour des valeurs stationnaires de la densité de courant J_s, s'étendant de valeurs sensiblement en deçà de la valeur critique J_{c0} à bien au-delà. Ces simulations étant de type Monte-Carlo, pour chaque valeur de J_s, de multiples trajectoires de retournement du moment magnétique $\mathbf{m_f}$ de la couche libre sont simulées (typiquement entre 250 et 25 000). La moyenne arithmétique permet de construire un estimateur empirique du délai moyen avant retournement $<\Delta t>$, tandis qu'un histogramme fournit une estimation empirique de sa distribution statistique. L'étude des résultats de ces simulations Monte-Carlo est l'objet des parties 2.2.4.1 et 2.2.4.2.

Il est à noter qu'un ensemble de géométries et de barrières énergétiques a été simulé, afin de pouvoir construire un modèle analytique de haut niveau présentant une certaine robustesse face à des variations de géométrie et de matériaux des jonctions tunnel magnétiques. Il s'agit d'obtenir un modèle qui puisse être utilisé en ayant simplement connaissance d'un nombre restreint de paramètres. L'idée est d'être capable de prédire les bonnes tendances selon le régime de programmation employé, sans immanquablement nécessiter le recours à un ajustement

216. M. GAJEK et coll., Applied Physics Letters, 2012.

I. Nous négligerons à ce propos tout anisotropie magnétocristalline et considérerons l'anisotropie de forme comme l'unique source d'anisotropie magnétique.

217. C. SERPICO, I.D. MAYERGOYZ et G. BERTOTTI, Journal of Applied Physics, 2001.

218. L. BANAS, Numerical Analysis and Its Applications, 2005.

numérique (ce dernier restant toutefois envisageable si le besoin s'en fait sentir et que des données expérimentales sont disponibles). Les jeux de paramètres restreints ainsi utilisés dans les simulations sont présentés dans la table 2.1. Les valeurs retenues sont représentatives des ordres de grandeur rencontrés dans la littérature.

Étiquette légendes		A	B	C						
Demi-grand axe a	(nm)	50	37	50	40	47	50	54	60	60
Demi-petit axe b	(nm)	20	19	20	20	18	20	21,5	20	20
Épaisseur t_f	(nm)	2	1,9	2	2	1,9	2	2,2	1,5	2
Saturation magnétique $\mu_0 M_s$	(T)	1,5	1,8	1,8	1,5	1,8	1,2	1,2	1,5	1,5
Résistance état parallèle R_P	(kΩ)	..... 1								
Résistance état antiparallèle R_{AP}	(kΩ)	.....2,5								
Température T	(K)	..... 300								
Coefficient de Gilbert α	(sans unité)	..... 0,01								

TABLE 2.1 – Paramètres de la couche ferromagnétique libre des composants simulés : caractéristiques (*haut*) et communs (*bas*). Seuls les composants utilisés dans les figures sont dotés d'une étiquette.

Deux types de commutations sont possibles : de la configuration magnétique parallèle vers celle antiparallèle (P→AP) et réciproquement (AP→P). Comme nous avons par exemple pu le voir dans le cas de la densité de courant critique avec les équations (2.4) et (2.5), les formules associées à chacun des cas se différencient seulement par un facteur multiplicatif (tel que $1+P^2$ ou $1-P^2$), appliqué à la densité de courant injectée J_s et lié à l'efficacité du transfert de spin. Afin de diviser par deux le nombre de simulations Monte-Carlo à réaliser, nous simulerons un cas exempt de ce facteur multiplicatif, et les axes de densités de courant seront ensuite remis à l'échelle selon le facteur multiplicatif du type de commutation souhaitée.

2.2.3.3 Exemple de résultats

Un exemple de trajectoire illustrant le processus de retournement est présenté à la figure 2.10. La trajectoire du moment magnétique $\mathbf{m_f}$[I] de la couche libre (ligne bleue en trait plein) a été obtenue en résolvant numériquement l'équation (2.2) en présence d'un courant constant et à une température supérieure à zéro kelvin. Les oscillations en spirales de la trajectoire sont dues au terme de précession de l'équation (2.2). La trajectoire est par ailleurs bruitée par l'agitation thermique à laquelle est soumis le moment magnétique. L'orbite de précession du moment magnétique $\mathbf{m_f}$ augmente progressivement, s'éloignant de plus en plus de la position initiale, jusqu'à provoquer la commutation vers la configuration magnétique opposée. Après le retournement, le courant injecté ne s'oppose plus à l'amortissement de la trajectoire, mais agit au contraire comme un amortissement supplémentaire : le moment magnétique $\mathbf{m_f}$ converge rapidement vers sa nouvelle position d'équilibre stable.

I. Plus précisément de la pointe du moment magnétique $\mathbf{m_f}$.

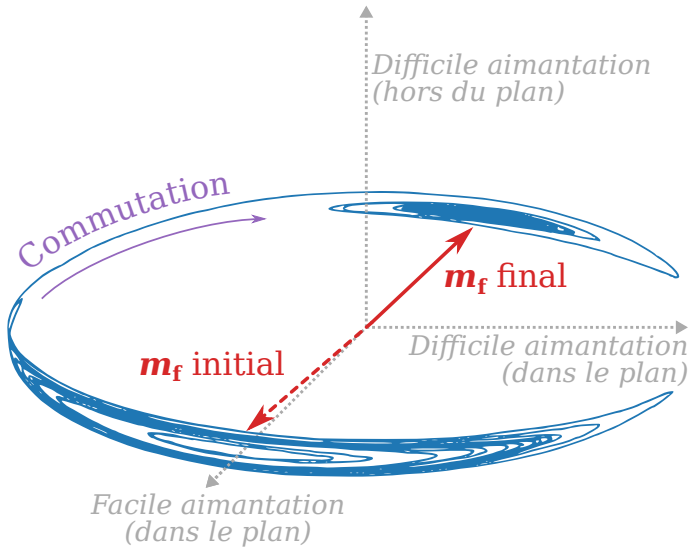


FIGURE 2.10 – Exemple, dans le cas d'une jonction tunnel magnétique dont l'aimantation est dans le plan, de la trajectoire suivie par le moment magnétique $\mathbf{m_f}$ de la couche libre, régie par l'équation (2.2), lors de l'application d'un courant constant à une température (absolue) non-nulle.

Nous pouvons trouver certains exemples de l'analyse des résultats fournis par ce type de simulations aux figures 2.13 page 89 et 2.15 page 94.

Remarque

Rassembler les données de chacune des courbes de la figure 2.13 a nécessité plusieurs jours de calculs sur un cœur de processeur moderne (Intel® Xeon E5-1620), et ce malgré l'utilisation d'un langage compilé pour les portions critiques (en termes de vitesse d'exécution) du code de simulation. La raison de ceci est liée au fait que le pas de temps des simulations ne peut pas être augmenté proportionnellement à la durée moyenne de ces dernières, qui augmente lorsque le courant injecté diminue. Le pas de temps employé durant ces simulations doit en effet rester suffisamment faible pour décrire la dynamique de précession du moment magnétique $\mathbf{m_f}$ de la couche libre, qui ne dépend du courant injecté. Or le temps moyen d'incubation avant retournement à simuler peut quant à lui devenir bien plus élevé, de sept à huit ordres de grandeur dans le cas des plus faibles valeurs de densité de courant J_s simulées.

Ceci souligne l'importance et l'intérêt de développer un modèle analytique du comportement stochastique des jonctions magnétiques qui soit plus rapide à simuler dans le cadre de la conception et de l'optimisation de circuits, ainsi que de simulations à l'échelle d'un large système.

2.2.3.4 Nomenclature

La plupart des symboles utilisés dans les équations du présent chapitre sont regroupés dans la table 2.2 page suivante. Ces notations seront conservées tant que faire se peut dans les chapitres suivants.

Symbole	Description (unité)	Formule
$\hbar$	constante de Planck réduite (J·s)	
$\|e\|$	charge d'un proton (C)	
μ_0	perméabilité magnétique du vide (H·m^{-1})	
μ_B	magnéton de Bohr (A·m^2)	
γ	rapport gyromagnétique (rad·s^{-1}·T^{-1})	
k_B	constante de Boltzmann (J·K^{-1})	
a	demi-grand axe de la couche libre (m)	
b	demi-petit axe de la couche libre (m)	
t_f	épaisseur de la couche libre (m)	
P	polarisation en spin du courant	
α	coefficient d'amortissement de Gilbert	
M_s	aimantation à saturation (A·m^{-1})	$M_s = \frac{\|\|\mathbf{m_f}\|\|}{V}$
T	température (K)	
V	volume de la couche libre (m^3)	$V = \pi a b t_f$
N_{a,b,t_f}	coefficients du champ démagnétisant[219] selon les directions de a, b et t_f	
f_0	facteur de fréquence du modèle de Néel-Brown[220] (Hz)	$f_0 = (\|\gamma\| N_a \mu_0 M_s)/\pi$
H_k	amplitude du champ d'anisotropie[221] (A·m^{-1})	$H_k = (N_b - N_a) M_s$
E_0	barrière énergétique à courant nul[221] (J)	$E_0 = (\mu_0 M_s H_k V)/2$
ϕ_0	écart-type angulaire dans le plan (rad)	$\phi_0^2 = k_B T/(\mu_0 (N_b - N_a) M_s^2 V)$
τ_0	temps caractéristique du modèle de Sun[204] (s)	$\tau_0 = \ln\left(\frac{\pi}{2\phi_0}\right) / \left(\alpha \|\gamma\| \mu_0 (H_k + \frac{M_s}{2})\right)$
$\mathscr{D}$	paramètre de diffusion[209] (T^2·s·m^3)	$\mathscr{D} = \frac{\alpha}{1+\alpha^2} \cdot \frac{k_B T}{\|\gamma\| M_s}$
R_P	résistance électrique dans la configuration parallèle (Ω)	
R_{AP}	résistance électrique dans la configuration antiparallèle (Ω)	
TMR	magnétorésistance tunnel	Éq. (2.1)
$\mathbf{m_f}$	moment magnétique global de la couche libre (A·m^2)	
$\hat{\mathbf{u}}_{inc}$	vecteur unitaire de la direction des électrons polarisés incidents	
$\eta(\mathbf{m_f}, \hat{\mathbf{u}}_{inc})$	efficacité angulaire du transfert de spin	
J_s	densité de courant (A·m^{-2})	
J_{c0}	amplitude de la densité de courant critique à 0 K (A·m^{-2})	…Éq. (2.4) et (2.5)
$\mathbf{H_{eff}}$	champ effectif (A·m^{-1})	
$\mathbf{h_{sto}}$	terme aléatoire de Langevin (A·m^{-1})	Voir éq. (2.3) de [209]

TABLE 2.2 – Symboles récurrents dans les formules du chapitre. Les formules sont données sous les hypothèses des simulations Monte-Carlo réalisées.

2.2.4 Développement d'un modèle analytique

Dans un premier temps, nous nous intéressons à modéliser le délai moyen avant commutation $<\Delta t>$ sur l'ensemble des régimes de courant de programmation, présentant les modèles qui existent dans la littérature, avant d'introduire le complément que nous proposons. Dans un second temps, notre intérêt se porte sur la description statistique du délai avant commutation Δt, effectuée au moyen d'une loi de probabilité paramétrique qui unifie les résultats parcellaires disponibles dans la littérature et ceux obtenus par le biais des simulations physiques présentées précédemment. Les paramètres de la loi de probabilité retenue s'appuient notamment sur le modèle du délai moyen avant commutation $<\Delta t>$ développé auparavant.

2.2.4.1 Délai moyen avant commutation

Dans cette partie, nous introduisons les deux équations classiquement utilisées pour modéliser le délai moyen avant retournement de l'aimantation de la couche libre d'une jonction tunnel magnétique soumise à un courant constant : le modèle de Néel-Brown et le modèle de Sun. Nous présentons par ailleurs deux nouvelles équations, inspirées de considérations physiques, qui définissent les limites de validité de ces deux modèles, que nous noterons J_{NB} et J_{Sun}. Enfin, est également présenté un nouveau modèle analytique du délai moyen avant commutation, dans la zone de courant intermédiaire entre les régimes de Néel-Brown et de Sun, où aucun des deux précédents modèles n'est satisfaisant.

À faible courant : modèle de Néel-Brown et limite associée. Lorsque la densité de courant J_{s} injectée selon la normale à la jonction tunnel magnétique est suffisamment faible, le processus de retournement du moment magnétique $\mathbf{m_f}$ de la couche libre est dominé par un phénomène d'activation thermique au-dessus de la barrière énergétique E_0 séparant les deux configurations magnétiques de la jonction. Dans ce régime, le courant injecté a un effet analogue à un abaissement de la barrière énergétique[222] (ce qui peut également être interprété comme une augmentation de la température effective de la couche libre).

Si l'amplitude de la densité de courant J_{s} est significativement inférieure à sa valeur critique J_{c0}, la résolution d'une équation de Fokker-Plank tirée de l'équation (2.2) permet d'obtenir le modèle dit de Néel-Brown[202,203,220,222] pour le délai moyen avant retournement $<\Delta t>$ (à la température T) :

$$<\Delta t> = f_0^{-1} \times \exp\left(\frac{E_0}{k_{\mathrm{B}}T}\left(1-\frac{J_{\mathrm{s}}}{J_{\mathrm{c0}}}\right)\right), \tag{2.6}$$

où k_{B} désigne la constante de Boltzmann et f_0 est une fréquence liée à celle de la précession de Larmor du moment magnétique $\mathbf{m_f}$[220].

222. Z. Li et S. Zhang, Physical Review B, 2004.
220. D. M. Apalkov et P. B. Visscher, Physical Review B, 2005.

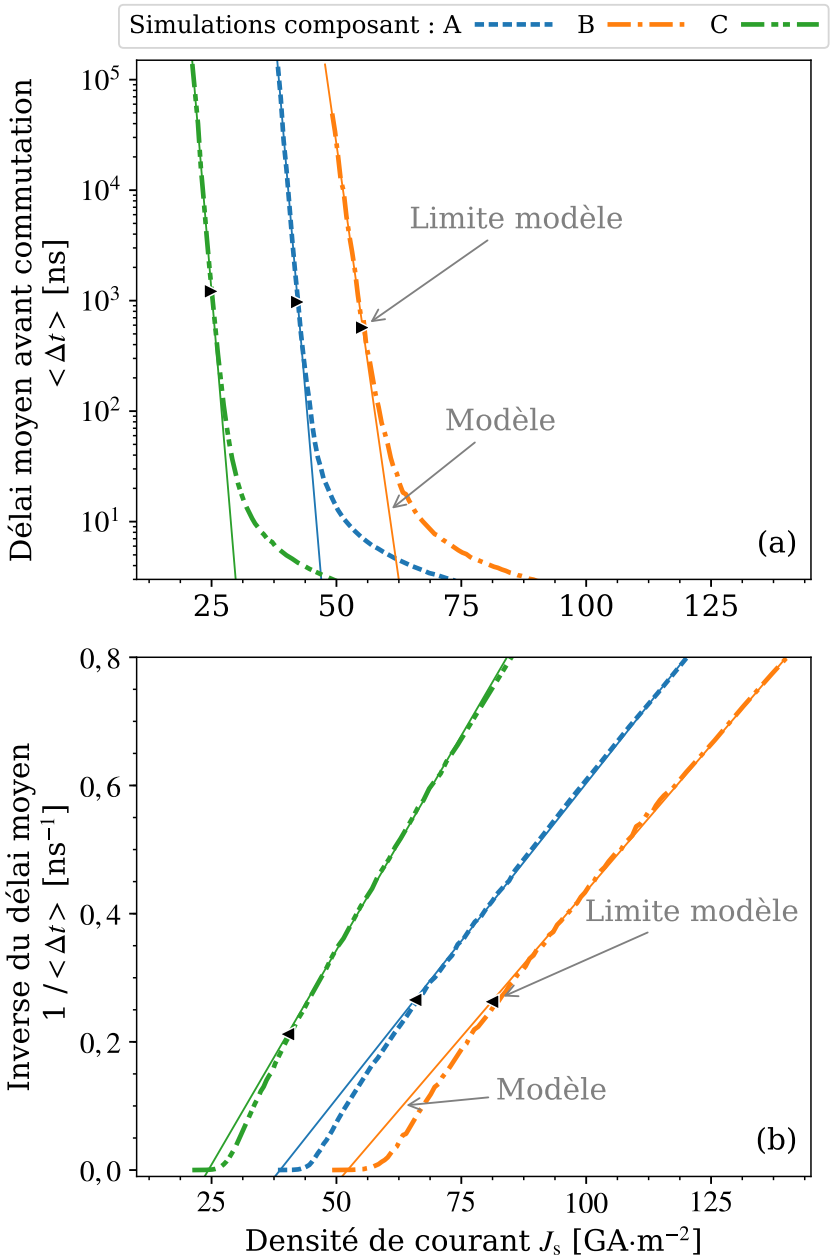


FIGURE 2.11 – *(a)* : délai moyen avant commutation en fonction de la densité de courant injectée, obtenu par les simulations physiques (lignes en pointillé) et prédit par le modèle de Néel-Brown (lignes en trait plein), dans le cas de faibles densités de courant (par rapport à la valeur critique J_{c0}), pour différentes géométries (couleurs). Le symbole ▶ indique la valeur de J_{NB} et montre son bon accord avec la frontière de la zone où le modèle de Néel-Brown commence à décrocher.
(b) : inverse du délai moyen avant commutation, en fonction de la densité de courant injectée, obtenu par les simulations physiques (lignes en pointillé) et prédit par le modèle de Sun (lignes en trait plein), dans le cas de fortes densités de courant (par rapport à la valeur critique J_{c0}), pour les mêmes géométries que la figure (a). Le symbole ◀ indique la valeur de J_{Sun}, également en bon accord avec la zone où le modèle de Sun montre ses limites.
Le détail des géométries A, B et C est donné dans la table 2.1.

Exemple

Afin d'illustrer la façon dont le délai moyen avant commutation peut aisément être ajusté sur plusieurs décades dans le cas du régime de faible courant, considérons une valeur classique de 1 ns pour f_0^{-1} et prenons l'exemple d'une barrière énergétique E_0 de 40 $k_B T$ (il s'agit d'une valeur permettant la rétention de l'état avec une probabilité de 50 % durant approximativement 7,5 années). Dans ce cas, le modèle de Néel-Brown prédit un délai moyen avant commutation $<\Delta t>$ d'environ 0,2 µs lorsque la densité de courant J_s est égale à $\frac{3}{4} J_{c0}$. Si maintenant nous réduisons la densité de courant J_s d'un tiers (c'est-à-dire $J_s = \frac{1}{2} J_{c0}$), alors $<\Delta t>$ atteint désormais presque 0,5 s !

L'équation (2.6) a été obtenue en faisant l'hypothèse d'une barrière énergétique importante. En deçà d'une valeur particulière de barrière énergétique, il est donc nécessaire de considérer le modèle de Néel-Brown comme non valide. Dans ce but, nous introduisons une valeur limite J_{NB} de la densité de courant, pour laquelle la barrière énergétique effective atteint une valeur E_{NB} bien choisie, et en deçà de laquelle le modèle de Néel-Brown n'est plus considéré comme pertinent :

$$E_0 \left(1 - \frac{J_{NB}}{J_{c0}}\right) = E_{NB}. \tag{2.7}$$

Nous pouvons observer sur la figure 2.11 qu'avec une échelle des ordonnées logarithmique, le modèle de Néel-Brown s'ajuste bien linéairement dans le domaine des faibles courants avec les simulations physiques, avant de décrocher pour les valeurs plus élevées de courant. Nous pouvons également constater que l'expression (2.7) prédit assez précisément la densité de courant au-delà de laquelle le modèle de Néel-Brown montre ses limites. La valeur de E_{NB} employée, déterminée empiriquement et identique pour l'ensemble des géométries présentées, est de 7,25 $k_B T$.

À fort courant : modèle de Sun et limite associée. Lorsque l'amplitude de la densité de courant J_s est significativement plus élevée que la valeur critique J_{c0}, la dynamique du retournement du macrospin $\mathbf{m_f}$ est surtout dictée par l'action du couple de transfert de spin, qui s'oppose à l'amortissement magnétique. Ceci résulte en une commutation quasi-adiabatique, l'effet de l'agitation se résumant alors essentiellement à fixer la position initiale lorsque le courant est appliqué. Le délai moyen avant commutation $<\Delta t>$ suit alors le modèle de Sun[172,186,204]

$$<\Delta t> = \tau_0 \times \frac{J_{c0}}{J_s - J_{c0}}, \tag{2.8}$$

172. T. DEVOLDER et coll., Physical review letters, 2008.

avec dans le cas d'une jonction tunnel magnétique dont l'aimantation est dans le plan :

$$\tau_0 = \ln\left(\frac{\pi}{2\phi_0}\right) \times \frac{1}{\alpha\gamma\mu_0\left(\frac{M_s}{2} + H_k\right)} \tag{2.9}$$

où ϕ_0 est l'écart-type de l'angle initial aléatoire entre le moment magnétique $\mathbf{m_f}$ et sa position stable. Cet angle initial dépend de l'agitation thermique à laquelle est soumise la couche libre de la jonction magnétique : avant qu'un courant ne soit appliqué, il fluctue naturellement, en suivant une loi gaussienne de moyenne nulle et de variance $\phi_0^2 = k_B T/(\mu_0 H_k M_s V)$. Le retournement dans ce régime de courant est parfois qualifié de « précessionnel », la trajectoire n'étant quasiment pas affectée par les fluctuations thermiques.

Encore une fois, l'équation correspondante dans le cas d'une jonction à anisotropie magnétique perpendiculaire peut être trouvée dans la référence [213].

Exemple

Si nous considérons une jonction tunnel magnétique avec les mêmes paramètres que celle de l'exemple précédent et que nous supposons une valeur de 4° pour l'écart-type ϕ_0, alors pour une amplitude de densité de courant $J_s = 2\ J_{c0}$, le délai moyen avant commutation prédit par le modèle de Sun est autour de 3 ns. Dans ce régime de plus fort courant, le délai observé avant la commutation de la couche libre est ainsi beaucoup plus court en moyenne que celui dans le régime de plus faible courant.

Le modèle de Sun sort de sa zone de validité lorsque la densité de courant devient suffisamment faible pour que la composante brownienne du mouvement du moment magnétique $\mathbf{m_f}$ ait une influence significative sur la trajectoire de ce dernier. Un résultat classique[223,224] est que la moyenne quadratique d'une quantité régie par un mouvement brownien est proportionnelle au temps. Dans le cas d'une jonction tunnel magnétique à aimantation dans le plan et de section elliptique, considérons $m_{f_b}^{\text{th}}$, la projection sur l'axe de difficile aimantation dans le plan (figure 2.10) d'un macrospin entièrement régi par l'agitation thermique : en nous basant sur la remarque précédente et des considérations dimensionnelles, nous pouvons nous attendre à ce que sous le coup de son agitation brownienne

$$< m_{f_b}^{\text{th}}(t)^2 >= \left(|\gamma| M_s V\right)^2 \times \frac{2\mathscr{D}}{V} \times t, \tag{2.10}$$

où t est le temps écoulé depuis l'instant initial et $\mathscr{D}$ est un paramètre de diffusion obtenu par application du théorème de fluctuation-dissipation[209].

En faisant l'hypothèse de nous placer dans le cas de petits angles du macrospin $\mathbf{m_f}$ par rapport à sa position initiale, nous pouvons alors obtenir une estimation de l'échelle de temps

223. A. EINSTEIN, Annalen der Physik, 1905.
224. G. E. UHLENBECK et L. S. ORNSTEIN, Physical Review, 1930.

Δt_{sto} nécessaire à cette composante stochastique pour avoir une influence significative sur le retournement du macrospin $\mathbf{m_f}$

$$\Delta t_{\text{sto}} = \frac{V\phi_0^2}{2\mathscr{D}|\gamma|^2}. \tag{2.11}$$

En égalisant cette échelle de temps de la dérive brownienne avec les prédictions du délai moyen avant commutation prédit par modèle de Sun à l'équation (2.8), cela nous suggère une limite de validité J_{Sun} de ce modèle satisfaisant la relation

$$\tau_0 \times \frac{J_{\text{c0}}}{J_{\text{Sun}} - J_{\text{c0}}} = \Delta t_{\text{sto}} \tag{2.12}$$

et en-dessous de laquelle le modèle de Sun arrête d'être réellement pertinent.

Comme nous pouvons le constater à la figure 2.11b, dont l'échelle des ordonnées est l'inverse du délai moyen avant commutation, le modèle de Sun s'ajuste lui aussi linéairement avec les résultats des simulations physiques, dans le cas des plus fortes densités de courant. Cette figure montre également que l'expression proposée de J_{Sun} fournit une prédiction satisfaisante de la limite de validité du modèle de Sun.

Comportement dans le régime de courant intermédiaire. Entre les deux régimes de courant présentés précédemment existe une zone intermédiaire, dans laquelle l'influence des fluctuations thermiques sur la *trajectoire* du retournement [I] ne peut plus être négligée, ce qui rend la description analytique de la dynamique du macrospin $\mathbf{m_f}$ beaucoup plus ardue que dans les régimes plus extrêmes présentés ci-avant[186]. Il n'existe ainsi pas dans la littérature de modèle analytique de haut niveau tiré de la physique qui permette de décrire le délai moyen avant commutation de la couche libre d'une jonction tunnel magnétique dans ce régime de courant. Par ailleurs, il est malaisé de faire naïvement la jonction entre le modèle de Néel-Brown et le modèle de Sun, puisque ce dernier diverge mathématiquement lorsque $J_{\text{s}} = J_{\text{c0}}$. Néanmoins, les limites J_{NB} et J_{Sun} proposées n'étant pas abruptes, il semble raisonnable de songer que la transition d'un comportement extrême à l'autre puisse être décrite de manière douce. C'est également ce que suggèrent les résultats des simulations physiques, où nous n'observons pas de changement brutal de comportement du délai moyen avant commutation (figures 2.11 et 2.13).

Suite aux constatations précédentes, nous proposons pour modéliser le délai moyen avant commutation dans le régime de courant intermédiaire, d'avoir recours à une expression mathématique dont les deux termes principaux tendent à reproduire qualitativement les comportements des modèles employés dans les régimes extrêmes [II]

$$<\Delta t> = f_0^{-1} \exp\left(\frac{E_0}{k_{\text{B}}T}\left(1 - \frac{J_{\text{s}}}{J_{\text{c0}}}\right)\right) + \text{w}_{\text{r}}(J_{\text{s}} - J_{\text{NB}}) \cdot \tau_0 \left(C_1 \frac{(1-\varepsilon)J_{\text{NB}}}{J_{\text{s}} - (1-\varepsilon)J_{\text{NB}}} + C_0\right). \tag{2.13}$$

I. Et non plus seulement le point de départ de cette dernière.
II. À savoir le modèle de Néel-Brown et le modèle de Sun.

Afin de repousser légèrement hors de la zone intermédiaire la divergence du terme mimant le modèle de Sun, il a été substitué $(1-\varepsilon)J_{\mathrm{NB}}$ à J_{c0}, avec $\varepsilon > 0$. Empiriquement, une valeur de 0,01 pour ε donne de bons résultats. Les constantes (réelles) C_0 et C_1 sont choisies de façon à ce que pour $J_{\mathrm{s}} = J_{\mathrm{Sun}}$, l'expression (2.13) ait la même valeur que le modèle de Sun (équation (2.8)) ainsi que la même dérivée première.

Pour plus de clarté de l'expression mathématique présentant la fenêtre d'apodisation définie par la fonction $\mathrm{w_r}$, nous introduisons les notations

$$\begin{aligned} x &= J_{\mathrm{s}} - J_{\mathrm{NB}}, \\ x_{\mathrm{c}} &= J_{\mathrm{c0}} - J_{\mathrm{NB}}, \\ x_{\mathrm{lim}} &= J_{\mathrm{Sun}} - J_{\mathrm{NB}}, \\ \Delta x &= \frac{x_{\mathrm{c}}(x_{\mathrm{lim}} - x_{\mathrm{c}})}{4x_{\mathrm{lim}}}. \end{aligned}$$

La fonction d'apodisation $\mathrm{w_r}$ est alors définie sur l'intervalle $[0, x_{\mathrm{lim}}]$ par

$$\mathrm{w_r}(x) = \begin{cases} \mathrm{p}(x) \times \frac{1}{2}\left(1 - \cos\left(\frac{\pi x}{\Delta x}\right)\right) & , 0 \le x < \Delta x \\ \mathrm{p}(x) & , \Delta x \le x \le x_{\mathrm{lim}} \end{cases}, \tag{2.14}$$

avec le polynôme

$$\mathrm{p}(x) = p_2\left(\frac{x}{x_{\mathrm{lim}}}\right)^2 + p_1\left(\frac{x}{x_{\mathrm{lim}}}\right) + p_0 \tag{2.15}$$

Une valeur empiriquement appropriée pour le coefficient p_0 est 1,66, d'où découlent $p_1 = -1{,}32$ et $p_2 = 0{,}66$ afin d'assurer une valeur unitaire et une dérivée première nulle de $\mathrm{w_r}$ en $x = x_{\mathrm{lim}}$.

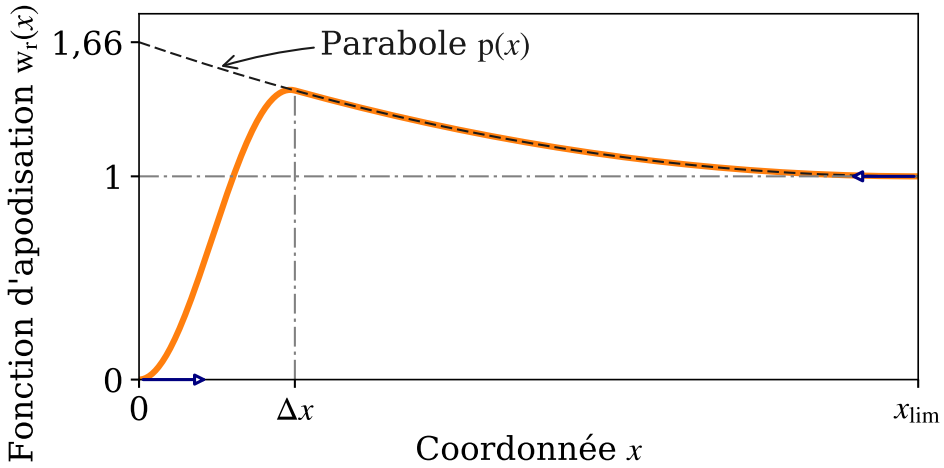


FIGURE 2.12 – Esquisse (*en trait plein*) de la fenêtre d'apodisation définie par $\mathrm{w_r}$ dans l'équation (2.12). La composante parabolique $\mathrm{p}(x)$ est celle de l'équation (2.15). La fonction $\mathrm{w_r}$ est de tangente horizontale aux deux extrémités de son domaine de définition, comme cela y est illustré par les flèches de couleur sombre.

La figure 2.12 présente une esquisse de la fenêtre d'apodisation définie par w_r. De droite à gauche, elle croît paraboliquement selon $p(x)$, jusqu'à atteindre $x = \Delta x$, moment à partir duquel elle décroît jusqu'à s'éteindre totalement en $x = 0$. Les valeurs nulles de w_r et de sa dérivée première en $x = 0$ permettent d'assurer une jonction continûment dérivable avec le modèle de Néel-Brown. La composante parabolique $p(x)$ contrebalance le fait que la valeur de $<\Delta t>$ soit légèrement sous-estimée par le terme similaire au modèle de Sun, terme dont par ailleurs la singularité à été repoussée jusqu'à ce que le modèle de Néel-Brown prenne le pas.

Ce modèle complet du délai moyen avant commutation $<\Delta t>$ est confronté sur la figure 2.13 aux résultats issus des simulations physiques, avec une attention particulière sur la zone intermédiaire dans l'insert. Les frontières J_{NB} et J_{Sun} de cette zone intermédiaire sont par ailleurs indiquées par le biais de symboles triangulaires. Notre modèle s'ajuste particulièrement bien avec les résultats des simulations physiques, et ce même dans la zone intermédiaire où les autres modèles décrochent.

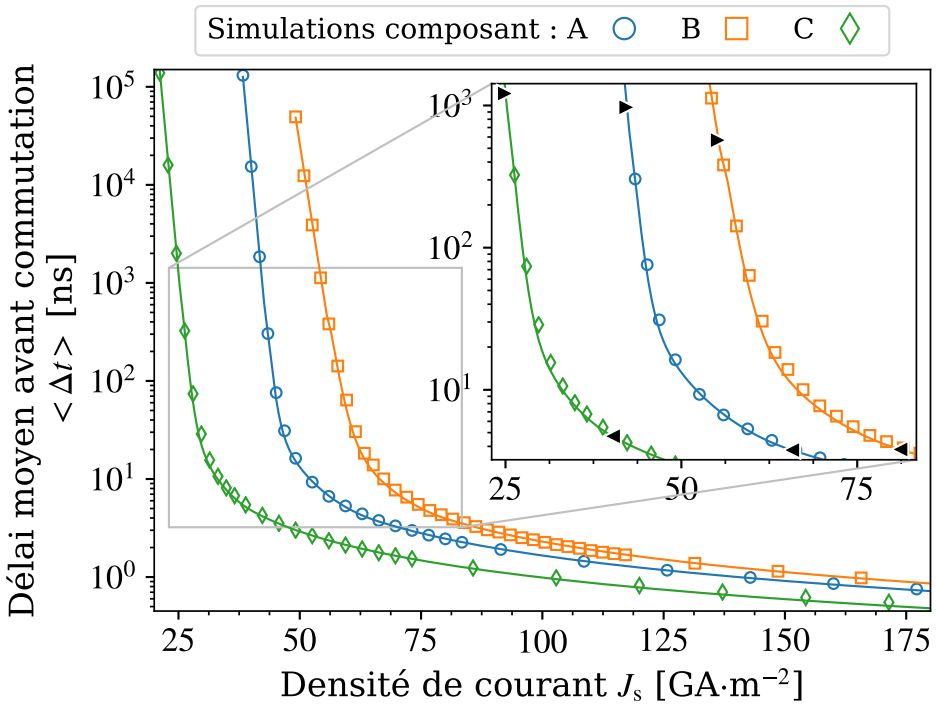


FIGURE 2.13 – Évolution du délai moyen avant commutation $<\Delta t>$ en fonction de la densité de courant J_s injectée. Les commutations sont de type antiparallèle vers parallèle (AP→P) et l'ensemble des régimes de courant sont présents. Le détail des géométries A, B et C est présenté dans la table 2.1. *Symboles* : résultats des simulations physiques. *Lignes* : prédictions de notre modèle analytique. *Insert* : zoom sur la zone de courant intermédiaire délimitée par J_{NB} (►) et J_{Sun} (◄).

2.2.4.2 Statistique du délai avant commutation

Dans cette nouvelle partie, nous nous focaliserons sur la fonction de densité de probabilité qui régit le comportement de la variable aléatoire qu'est le délai moyen avant commutation de la couche libre d'une jonction tunnel magnétique. L'un des défis est notamment d'obtenir un modèle satisfaisant sur l'ensemble des régimes de courant.

Modèle existant dans la littérature pour le régime de faible courant. Dans le régime de faible courant, il a été démontré théoriquement[222] et observé expérimentalement[167] que l'occurrence du retournement de l'aimantation de la couche libre d'une jonction tunnel magnétique suit un processus de Poisson : la fonction de densité de probabilité (PDF[I]) du délai Δt avant commutation est ainsi

$$f_{\mathrm{sw}}(t) = \frac{1}{<\Delta t>} \exp\left(-\frac{t}{<\Delta t>}\right), \tag{2.16}$$

où $<\Delta t>$ est donné par l'équation (2.6). Autrement dit, le délai avant commutation Δt suit une loi exponentielle dont la moyenne est donnée par le modèle de Néel-Brown[186,205]. Avec une telle fonction de densité de probabilité, si une impulsion de programmation (à courant constant) de durée t_{p} est appliquée sur la jonction, la probabilité de commutation $\mathscr{P}_{\mathrm{sw}}$ de cette dernière est donnée par

$$\begin{aligned}\mathscr{P}_{\mathrm{sw}} \hat{=} \Pr(\Delta t \leq t_{\mathrm{p}}) &= \int_0^{t_{\mathrm{p}}} f_{\mathrm{sw}}(t)\,\mathrm{d}t \\ &= 1 - \exp\left(-\frac{t_{\mathrm{p}}}{<\Delta t>}\right).\end{aligned} \tag{2.17}$$

Dans ces conditions, en jouant sur la durée t_{p} de l'impulsion de programmation, il est possible d'ajuster la probabilité de commutation de la jonction selon nos souhaits, d'une faible valeur ($\mathscr{P}_{\mathrm{sw}} \ll 1$ lorsque $t_{\mathrm{p}} \ll <\Delta t>$) à une forte valeur ($\mathscr{P}_{\mathrm{sw}} \approx 1$ pour $t_{\mathrm{p}} \gg <\Delta t>$).

Remarque

Une différence notable des jonctions tunnel magnétiques à transfert de spin avec la majorité des autre composants mémristifs est que les premières ne possèdent par de seuil abrupt, avec une discontinuité de comportement de part à d'autre de ce seuil. Même une valeur modeste de courant présente une chance non nulle de provoquer la commutation d'une jonction tunnel magnétique vers son autre état stable.

Comportement dans les régimes de courant plus élevé. Lorsque la densité de courant injectée J_{s} devient plus élevée, la loi aléatoire suivie par le délai avant commutation Δt n'est manifestement plus de type exponentiel, comme cela peut être constaté sur les résultats de

I. De l'anglais *Probability Density Function*. De façon analogue, nous pourrons utiliser l'acronyme CDF (*Cumulative Distribution Function*) pour désigner la fonction de répartition.

simulation présentés aux figures 2.15 (b–d). Nous pouvons observer qu'au fur et à mesure que la densité de courant J_s augmente, la fonction de densité de probabilité empirique[I] tend vers une distribution certes toujours désaxée vers la droite et unimodale mais dont le mode est désormais de valeur non nulle (contrairement à la loi exponentielle). Différentes distributions de probabilité ont été proposées dans la littérature pour décrire le comportement statistique du délai avant commutation Δt dans le régime de plus fort courant[197,225–227], sans qu'aucune ne s'impose indubitablement. Pour des raisons pratiques, nous privilégierons une distribution qui puisse être indifféremment employée dans n'importe quel régime de courant, ce qui n'est pas le cas des propositions de la littérature. Nous avons découvert que l'ajustement d'une loi gamma $\Gamma(k,\theta)$, dont la fonction de densité de probabilité est

$$f_{sw}(t;k,\theta) = \frac{t^{k-1}\exp\left(-\frac{t}{\theta}\right)}{\Gamma(k)\theta^k}, \tag{2.18}$$

est en accord satisfaisant avec les résultats issus de nos simulations physiques, et ce quel que soit le régime de courant considéré, comme cela peut être constaté sur la figure 2.15.

Les paramètres k et θ sont deux nombres réels strictement positifs et $\Gamma(k)$ désigne la *fonction* gamma évaluée en k. Il est à noter qu'une autre paramétrisation de la loi gamma peut être rencontrée, $\Gamma(\alpha,\beta)$, pour laquelle $\alpha = k$ tandis que $\beta = \theta^{-1}$. Nous veillerons à n'utiliser que la première dans le présent document.

La distribution statistique du délai avant commutation Δt est parfois qualifiée de « pseudo-log-normale »[210] pour les densités de courant supérieures à la valeur critique J_{c0}. Il s'avère que dans le cas d'une distribution désaxée vers la droite, les lois gamma et log-normale sont dans la majorité des situations d'allure proche, voire interchangeables[228–230]. Dans le cas présent, l'un des avantages d'une description statistique basée sur une loi gamma plutôt que log-normale tient au fait que la loi gamma puisse tendre vers une loi exponentielle. Si le délai avant commutation Δt suit une loi gamma $\Gamma(k=1,\theta)$, il est en effet possible de constater à l'aide de l'équation (2.18) que le délai avant commutation Δt suit bien loi exponentielle, de moyenne θ^{-1}. Il est ainsi possible avec cette seule loi gamma de passer d'une distribution de mode non nul et désaxée à droite dans les régimes de plus fort courant à une distribution exponentielle dans le régime de plus faible courant.

Le paramètre de forme k et le paramètre d'échelle θ peuvent s'exprimer simplement en fonction de la valeur moyenne et de l'écart-type du délai avant commutation Δt à travers les

I. Dans le cas présent réalisée à partir de l'histogramme des délais avant commutation simulés.

225. H. Tomita et coll., IEEE Transactions on Magnetics, 2011.
226. H. Zhao et coll., IEEE Transactions on Magnetics, 2012.
227. W. S. Zhao et coll., Microelectronics Reliability, 2012.
228. H.-K. Cho, K. P. Bowman et G. R. North, Journal of Applied Meteorology, 2004.
229. A. Abdi et M. Kaveh, IEEE Vehicular Technology Conference, 1999.
230. A. B. Z. Salem et T. D. Mount, Econometrica : Journal of the Econometric Society, 1974.

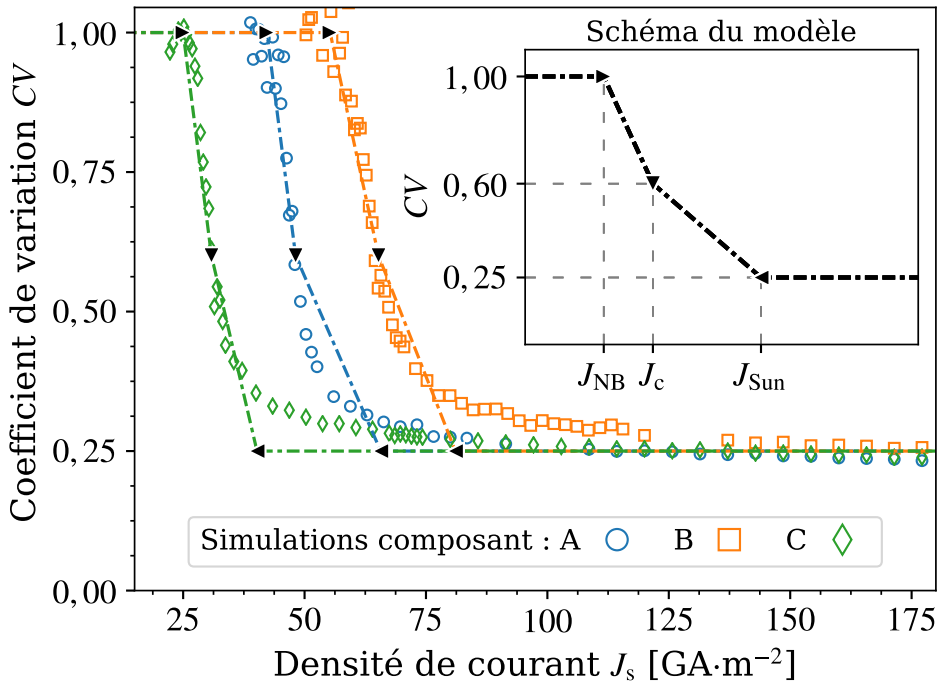


FIGURE 2.14 – Coefficient de variation CV du délai avant commutation Δt en fonction de la densité de courant injectée J_s. Les commutations sont de type antiparallèle vers parallèle (AP → P) et le détail des géométries A, B et C est présenté dans la table 2.1. *Symboles* : résultats des simulations physiques. *Lignes en pointillé* : prédictions de notre modèle empirique. *Insert* : schéma de construction du modèle empirique (*N.B.* : l'échelle n'est pas respectée).

deux relations

$$<\Delta t> = \theta \times k \tag{2.19}$$

$$\text{std}[\Delta t] = \theta \times \sqrt{k} \tag{2.20}$$

Le coefficient de variation CV du délai avant commutation Δt est alors entièrement défini par le paramètre de forme k puisque

$$CV \hat{=} \frac{\text{std}[\Delta t]}{<\Delta t>} = \frac{1}{\sqrt{k}} \tag{2.21}$$

où std$[\Delta t]$ désigne l'écart-type du délai moyen avant commutation. Ce coefficient de variation CV est tracé sur la figure 2.14. Nous pouvons constater que quelle soit la géométrie, différentes zones relativement invariantes semblent exister. Ainsi des plus faibles valeurs de densité de courant J_s aux plus fortes, nous pouvons identifier :

1. un plateau à $CV = 1$ lorsque $J_s \leq J_{NB}$, ce qui correspond au résultat théorique attendu dans le cas d'une loi exponentielle ;
2. un segment de forte pente négative lorsque $J_{NB} < J_s \lesssim J_{c0}$;
3. une portion coudée dans la région $J_s \sim J_{Sun}$;
4. un plateau autour de $CV = 0{,}25$ (avec en réalité une très légère pente négative) pour les valeurs les plus élevées de J_s.

De façon à garder une description simple du paramètre de forme k, en nous basant sur les observations précédentes, nous proposons la modélisation empirique présentée en insert de la figure 2.14 pour le coefficient de variation CV. Ce modèle est constitué de deux plateaux, $CV = 1$ pour $J_s \leq J_{NB}$ et $CV = 0{,}25$ lorsque $J_s \geq J_{Sun}$, reliés entre eux par deux segments de droite se rejoignant en le point de coordonnées $(J_{c0},\ 0{,}6)$. La valeur du plus faible plateau ainsi que le choix du point $(J_{c0},\ 0{,}6)$ sont basés sur des constatations empiriques : pour chacune des géométries de jonction tunnel magnétique simulées, la courbe du coefficient de variation CV s'approche de ces valeurs.

De ce modèle analytique empirique simple, il est aisé d'obtenir la valeur du paramètre de forme k de la loi gamma régissant le délai avant commutation Δt. De l'équation (2.19) et du modèle du délai moyen avant commutation $<\Delta t>$ de la partie 2.2.4.1 , il est ensuite possible de déduire la valeur du paramètre d'échelle θ. Nous sommes donc désormais en mesure, connaissant uniquement les paramètres de la géométrie et des matériaux de la jonction tunnel magnétique considérée, de calculer analytiquement les paramètres définissant la distribution de probabilité du délai moyen avant retournement Δt, et ce quelle que soit la valeur de la densité de courant injectée J_s. Il est à noter qu'un modèle similaire avec des courants de polarité opposée est à employer dans le cas d'une transition de la configuration magnétique parallèle vers la configuration antiparallèle (P$\rightarrow$AP).

La figure 2.15 montre des exemples de notre estimateur de la fonction de densité de probabilité, dans le cas de la géométrie A de la table 2.1 et pour des valeurs de densité de courant J_s

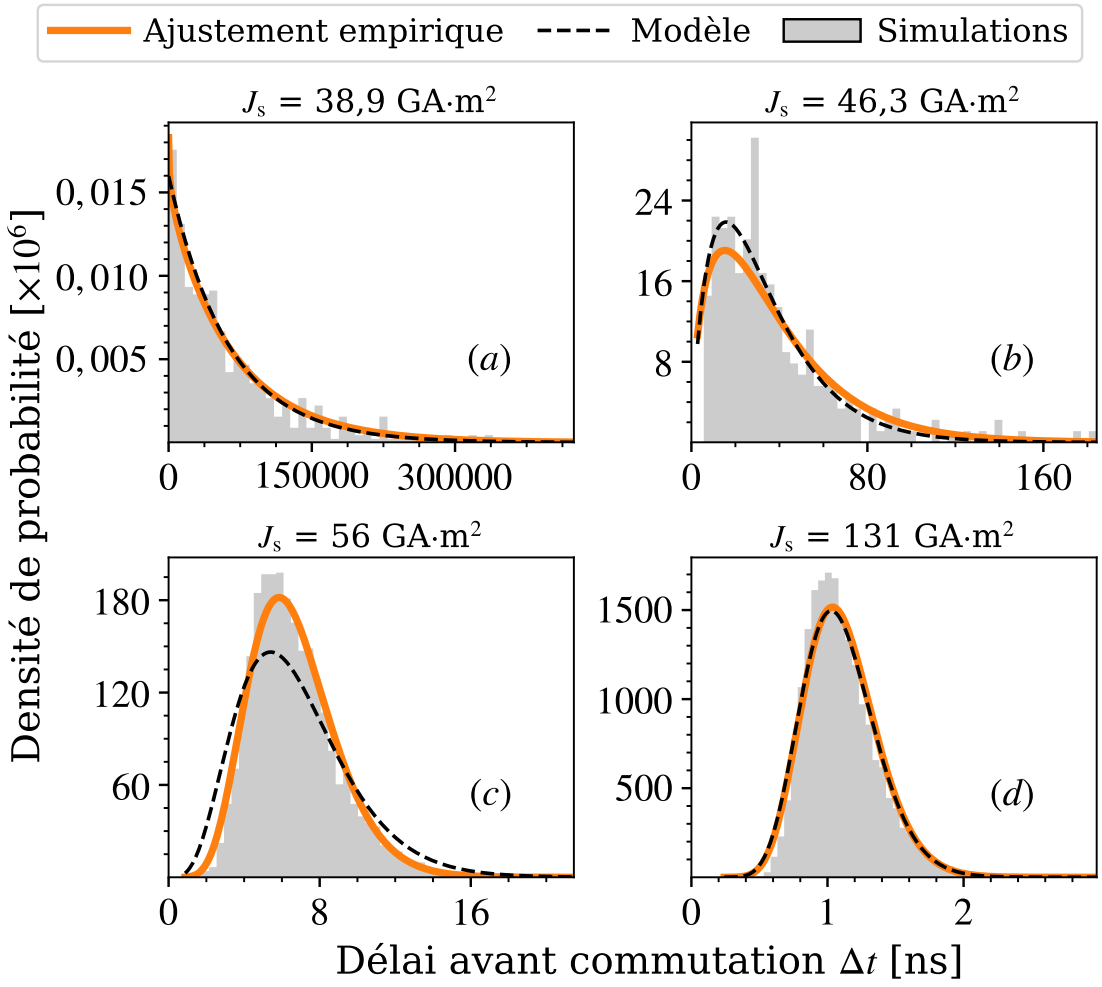


FIGURE 2.15 – Fonctions de densité de probabilité (PDF) pour des valeurs de densité de courant injectée J_s dans chacune des portions du modèle empirique du coefficient de variation CV du délai avant commutation Δt, dans le cas de la géométrie A de la table 2.1. Les commutations sont de type antiparallèle vers parallèle (AP→P).
Aires grisées : les histogrammes des résultats issus des simulations physiques. Leur aire est normée à l'unité.
Lignes en trait plein : les lois gamma dont les paramètres sont déduits de la moyenne et de l'écart-type des résultats de simulations physiques.
Lignes en pointillé : les lois gamma prédites par notre modèle empirique.

situées successivement dans chacune des portions du modèle adopté pour décrire le coefficient de variation CV du délai avant commutation Δt. Dans les cas (a) et (d), les prédictions de notre modèle sont particulièrement proches des résultats des simulations physiques, ce qui s'explique par le fait qu'à la fois la moyenne et le coefficient de variation du délai avant commutation Δt sont estimés fidèlement. Dans les cas (b) et (c), de petites déviations apparaissent, mais la distribution de probabilité estimée possède néanmoins une forme qualitativement correcte ainsi que la bonne valeur moyenne. Des comportements similaires ont été observés sur les autres géométries simulées.

Sur la figure 2.7 page 70, nous pouvons observer que la fonction de répartition (CDF) expérimentale du délai avant commutation Δt peut être ajustée de façon satisfaisante avec notre modèle, et que l'ajustement reste valable pour différentes tensions de programmation situées dans la zone intermédiaire. Ce dernier point illustre par ailleurs la possibilité d'ajuster numériquement le modèle analytique à d'éventuelles données expérimentales. Ceci est notamment intéressant dans les cas où des écarts quantitatifs sont observés entre ces dernières et les prédictions du modèle, sans que les phénomènes physiques qui régissent le comportement de la commutation n'aient été profondément modifiés. Ces différences peuvent par exemple provenir des effets de sous-volumes magnétiques[208], qui limitent la hauteur de barrière énergétique, ou encore des imperfections des procédés de fabrication.

2.2.4.3 Exemple d'usage simple de ce modèle stochastique

Il est possible, à l'aide du modèle présenté précédemment, d'estimer la vitesse de programmation et la consommation énergétique d'une jonction tunnel magnétique lorsqu'un taux d'erreur de programmation a été fixé. Connaissant la densité de courant injectée J_s (donc le courant I), le modèle présenté auparavant permet de calculer la durée t_p de l'impulsion de programmation qui garantisse le taux d'erreur de programmation choisi. Ceci permet ensuite d'en déduire une *estimation* de l'énergie de programmation $E_w = R(I)\,I^2\,t_p$, où $R(I)$ est la résistance électrique de la jonction dans les conditions de programmation adoptées. La figure 2.16 présente, pour différentes valeurs de taux d'erreur, l'évolution de la durée de programmation requise et de l'énergie de programmation associée avec la densité de courant injectée J_s. Nous pouvons voir que les prédictions de notre modèle analytique sont en assez bon accord avec les estimations réalisées à partir des résultats des simulations physiques, bien que les premières soient presque instantanées à calculer tandis que les secondes ont nécessité plusieurs jours de calcul pour collecter suffisamment de points de données.

Nous pouvons observer que le comportement de la durée de programmation t_p comme celui de l'énergie associée E_w dépend du régime de courant employé. Ces deux quantités varient en outre de presque une décade selon le taux d'erreur désiré pour la jonction tunnel magnétique. Par exemple, nous pouvons constater qu'accroître la densité de courant J_s accroît toujours la vitesse de programmation. Cependant, puisque cette dépendance se fait plus faible dans le régime des forts courants, l'énergie de programmation présente un minimum, qui est atteint

pour une certaine valeur de densité de courant J_s qui dépend du taux d'erreur de programmation souhaité. Ce minimum a lieu dans la zone intermédiaire de densité de courant J_s, située entre le modèle de Néel-Brown et le modèle de Sun.

Une telle analyse permet également de fournir des pistes pour optimiser la consommation énergétique des puces exploitant des jonctions tunnel magnétiques. Ainsi, les applications mémoires conventionnelles utilisant des jonctions tunnel magnétique pourraient se focaliser sur la maximisation de la vitesse de programmation. D'un autre côté, dans certaines applications émergentes et fortement parallèles telles que le calcul neuromorphique [231,232], la consommation énergétique est un sujet plus important que la vitesse de programmation : il pourrait être bénéfique de programmer la jonction à la densité de courant J_s nécessitant le moins d'énergie possible.

Si la suite du travail présenté dans ce document ciblera des applications neuromorphiques privilégiant à priori la consommation énergétique à la vitesse de programmation, les usages mémoires plus conventionnels peuvent préférer se placer à un régime de courant un peu plus élevé pour favoriser la vitesse de programmation. Néanmoins même dans ce second cas, il est envisageable de réduire la consommation en acceptant un taux d'erreur plus élevé, ce qui est possible pour certains problèmes (tels que le traitement d'image et de vidéo). Ce type de problématique a été étudiée par Nicolas LOCATELLI, membre de l'équipe NANOARCHI, qui a porté vers un modèle Verilog-A les équations analytiques présentées dans ce chapitre, permettant ainsi la simulation SPICE[I] de circuits électroniques qui intégrent des jonctions tunnel magnétiques et les programment de façon probabiliste[233].

231. M. SHARAD, D. FAN et K. ROY, Proceedings of the Annual Design Automation Conference, 2013.
232. W. S. ZHAO et coll., IFIP/IEEE International Conference on Very Large Scale Integration, 2013.
I. De l'anglais *Simulation Program with Integrated Circuit Emphasis*.
233. N. LOCATELLI et coll., International Conference on Memristive Systems, 2015.

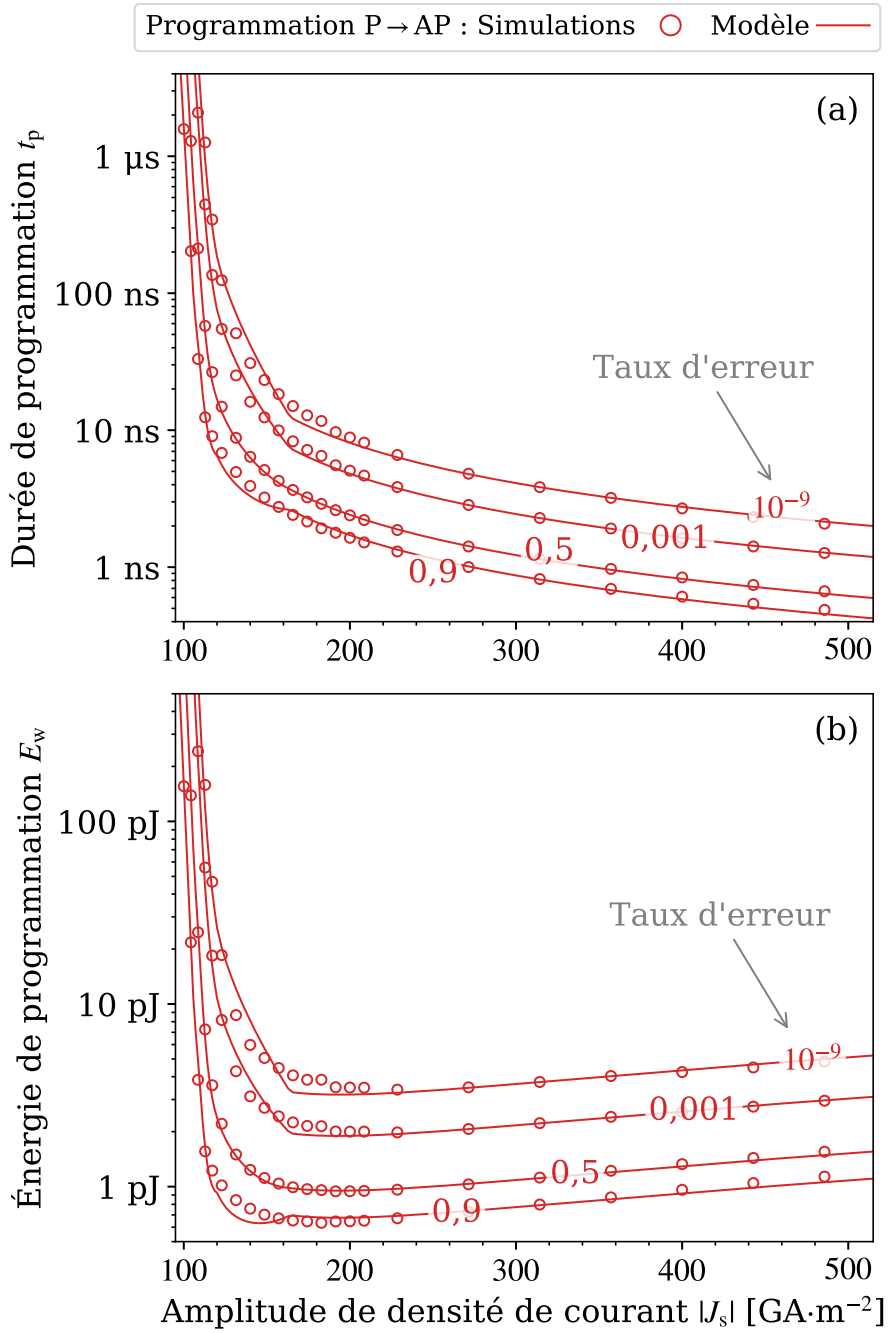


FIGURE 2.16 – *(a)* : la durée de programmation t_p nécessaire, en fonction de la densité de courant injectée J_s, pour garantir un certain taux d'erreur (c'est-à-dire une certaine probabilité d'échec du retournement de la couche libre).
(b) : estimation de l'énergie de programmation E_w en fonction de la densité de courant J_s.
Les commutations sont de type parallèle vers antiparallèle (P → AP). Les taux d'erreurs considérés sont indiqués par les nombres sur les courbes. Les estimations tirées de notre modèle analytique sont tracées en trait plein, tandis que celles basées sur les résultats des simulations physiques sont indiquées par les symboles.

2.3 Conclusions

Après avoir dans un premier temps présenté succinctement les phénomènes physiques majeurs régissant le comportement des jonctions tunnel magnétiques à transfert de spin ainsi que quelques éléments de technologie, nous avons dans un second temps porté notre attention sur leur modélisation :

1. physique, par le biais de simulations Monte-Carlo des équations décrivant la dynamique de l'aimantation de la couche libre ;
2. à plus haut niveau, en développant à partir de la littérature et des résultats des simulations Monte-Carlo, un modèle analytique de la statistique du délai avant commutation d'une jonction tunnel magnétique.

Ce nouveau modèle haut niveau est en accord satisfaisant avec les simulations physiques Monte-Carlo [I] tout en offrant la possibilité d'être ajusté numériquement à des mesures expérimentales si de telles données sont disponibles. Par ailleurs, il est beaucoup plus rapide à simuler que les simulations de la dynamique magnétique, ce qui nous permettra d'étudier dans le chapitre 3 un système neuromorphique ayant recours à un nombre important de ces nanocomposants mémoires innovants comme synapses artificielles. Au terme de ce chapitre, nous disposons en effet des bases permettant d'envisager utiliser les jonctions tunnel magnétiques à transfert de spin comme des éléments synaptiques binaires naturellement probabilistes, par l'exploitation du comportement *intrinsèque* de ces nanocomposants.

Nous terminons ce chapitre avec un exemple simple d'application de ce nouveau modèle, qui illustre l'intérêt de ces travaux de modélisation en suggérant que le régime de programmation optimal se situe dans la zone précédemment mal gérée par les modèles de la littérature.

I. Qui résolvent l'équation différentielle décrivant la dynamique de l'aimantation magnétique de la couche libre.

Chapitre 3

Étude de l'apprentissage par un ensemble de jonctions tunnel magnétiques utilisées comme synapses stochastiques binaires

"We are such stuff as dreams are made on; and our little life is rounded with a sleep."

PROSPERO

"CE CHAPITRE *évalue la possibilité d'utiliser des jonctions tunnel magnétiques à transfert de spin comme synapses artificielles binaires et stochastiques au sein d'une architecture neuromorphique. Par le biais des simulations informatiques à l'échelle d'un système comportant plusieurs centaines de milliers de synapses, nous mettons en évidence l'apprentissage d'une tâche dynamique pratique, en exploitant une règle d'apprentissage qui s'appuie sur le comportement intrinsèque des jonctions tunnel magnétiques étudié dans le chapitre précédent. Ces résultats sont obtenus grâce à des outils logiciels spécifiquement développés durant la thèse, permettant de décrire de façon fonctionnelle une telle architecture tout en conservant une description du comportement synaptique proche de la physique des composants employés. Un intérêt majeur de cette approche est de pouvoir étudier l'influence du régime de programmation des jonctions tunnel magnétiques sur leurs performances en tant que synapses artificielles et sur la consommation électrique associée. "*

Introduction

Dans ce chapitre, nous présentons des simulations informatiques d'un système neuromorphique simple qui utilise un grand nombre de jonctions tunnel magnétiques à transfert de spin comme synapses artificielles binaires. Nous exploitons le comportement stochastique naturel de ces nanocomposants pour mettre en œuvre une règle d'apprentissage *probabiliste* qui s'affranchit de la nécessité de circuits externes de génération de nombres aléatoires. Contrairement aux composants mémoires filamentaires présentés dans les parties 1.4.1 et 1.4.2, ce comportement aléatoire intrinsèque, dont l'origine physique est bien comprise, peut être contrôlé, ce qui permet d'envisager de façon réaliste des utilisations allant au-delà de simples mémoires non volatiles conventionnelles.

Les travaux du présent chapitre capitalisent sur le modèle analytique haut niveau du délai aléatoire avant commutation d'une jonction tunnel magnétique à transfert de spin développé dans le chapitre 2. Ce modèle présente en effet l'avantage d'offrir une bonne précision de prédiction de la distribution statistique du délai en question, tout en étant plus rapide à simuler que des modèles plus proches des équations de la physique de ces nanocomposants. Les simulations mises en place nous permettent d'étudier l'influence des imperfections de telles synapses sur l'apprentissage du système. Disposer d'un modèle désormais capable de décrire le comportement stochastique des jonctions tunnel magnétiques à transfert de spin dans tous les régimes de programmation possibles nous offre également la possibilité de nous intéresser à l'effet du régime de programmation sur l'apprentissage et la consommation énergétique du système.

3.1 Présentation des simulations réalisées à l'échelle d'un système

3.1.1 Architecture du système

Dans ce chapitre, nous allons montrer par des simulations à l'échelle d'un système, l'intérêt des jonctions tunnel magnétiques à transfert de spin employées comme éléments synaptiques binaires, programmés selon une règle stochastique. Dans ce but, nous avons adapté un schéma proposé auparavant dans la littérature pour les mémoires à changement de phase[137,138] et les mémoires électrochimiques métalliques à pont conducteur[88,132]. Le système en question implémente un réseau de neurones impulsionnels capable de réaliser un apprentissage non supervisé en s'appuyant sur une règle de type *Spike-Timing Dependent Plasticity*, simplifiée par rapport à la version biologique usuelle (figure 1.10 page 24) et adaptée aux caractéristiques des nanocomposants étudiés. L'architecture de ce système est présentée dans ses grandes lignes à la figure 3.1.

137. O. BICHLER et coll., IEEE Transactions on Electron Devices, 2012.
138. M. SURI et coll., International Electron Devices Meeting, 2011.
88. M. SURI et coll., IEEE Transactions on Electron Devices, 2013.
132. M. SURI et coll., IEEE International Electron Devices Meeting, 2012.

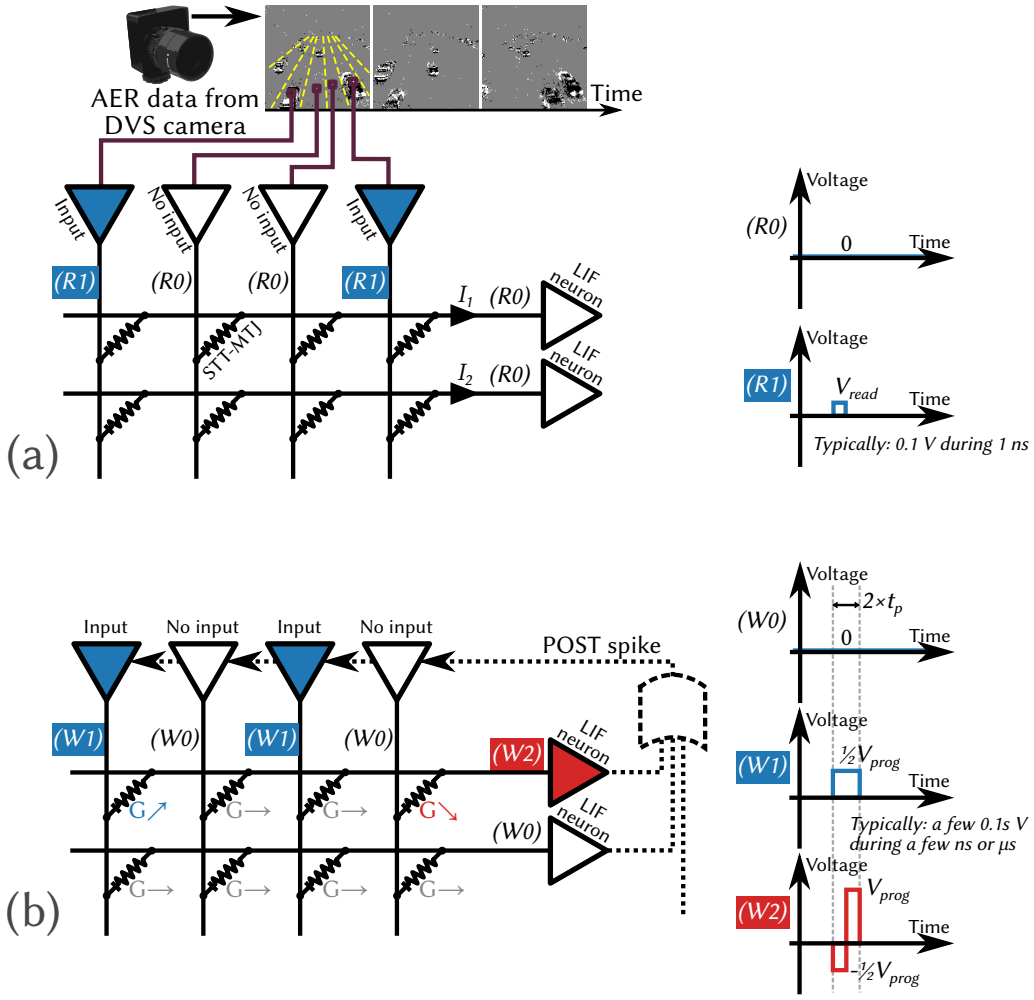


FIGURE 3.1 – Schéma général de l'architecture employée pour l'apprentissage à l'échelle d'un système. Les synapses sont organisées en une matrice sans transistors de sélection (structure 1R), chacune étant constituée d'une unique jonction tunnel magnétique à transfert de spin (STT-MTJ). Les entrées présynaptiques (▽) sont alimentées par une séquence d'événements enregistrée par une rétine neuromorphique de type *Dynamic Vision Sensor*[62] (DVS) dans un format *Address-Event Representation* (AER). Les trois vignettes supérieures en niveaux de gris constituent un exemple de représentation de ces signaux d'entrée. Les neurones de sortie (▷) sont de type « intègre-et-tire avec fuites » (LIF). Le caractère coloré des symboles triangulaires indique l'activité récente des entrées et sorties correspondantes, dans une fenêtre temporelle définissant le caractère causal ou non d'une paire d'impulsions.
(a) Phase de lecture, ayant lieu pour chaque tir d'un neurone d'entrée (en dehors d'une phase d'écriture) durant la phase d'inférence du système.
(b) Phase de programmation, qui a lieu lorsqu'un neurone de sortie tire et réalise un apprentissage de type *Spike-Timing Dependent Plasticity* probabiliste. Les formes d'onde de tension de programmation *(W1)* and *(W2)* sont appliquées simultanément. Du fait de la nature stochastique de la commutation, seules deux jonctions tunnel magnétiques à transfert de spin changent d'état dans cet exemple ($G\nearrow$ et $G\searrow$).

Les neurones d'entrée présentent les signaux d'entrée sous la forme d'impulsions asynchrones, qui peuvent par exemple provenir d'un capteur neuromorphique tel que la rétine artificielle de type *Dynamic Vision Sensor*[62] présentée dans la partie 1.1.3. Les jonctions tunnel magnétiques à transfert de spin sont organisées en une matrice connectant les neurones d'entrée aux neurones de sortie. La configuration architecturale la plus simple est une matrice passive, qualifiée de « 1R », proposée également dans le cadre d'applications mémoires[234]. Cette configuration est extrêmement efficace en termes de densité d'intégration et permet de connecter chaque neurone d'entrée à chaque neurone de sortie par le biais d'une unique jonction tunnel magnétique à transfert de spin, à la manière de la figure 1.6a. Contrairement au cas d'une puce mémoire, il est envisageable avec cette configuration de recourir à une phase de lecture hautement parallèle (voir la figure 1.11 page 29). Ceci permet en principe d'amoindrir les problèmes des courants parasites de fuite (*sneak path currents* en anglais) durant les phases d'inférence. Toutefois, des problèmes demeurent pour les phases d'écriture (c'est-à-dire l'apprentissage) et de lecture (autrement dit l'inférence), ce qui requiert une énergie inutile et sera discuté ultérieurement dans le chapitre 5. Pour s'affranchir de ce problème, il est possible de modifier cette structure en ajoutant un transistor de sélection en série avec chaque cellule mémoire (configuration dite « 1T-1R »). Cette modification a cependant pour effet de réduire la compacité de l'architecture par rapport à une configuration 1R. Des comparaisons plus détaillées sur les avantages et les inconvénients des configurations 1R et 1T-1R seront apportées dans le chapitre 5, portant sur les problématiques soulevées à l'échelle du circuit par l'utilisation synaptique de jonctions tunnel magnétiques à transfert de spin.

Dans le système que nous considérons, un neurone d'entrée ($\triangledown$) qui tire applique une impulsion de tension de courte durée et de faible amplitude sur la matrice d'éléments synaptiques, comme cela est illustré sur la figure 3.1a. Ceci a pour effet de faire circuler à travers les éléments de la matrice synaptique des courants qui atteignent de façon parallèle les différents neurones de sortie. Ces neurones de sortie, schématisés par des triangles ($\triangleright$) sur la figure 3.1, lisent ces courants tout en maintenant un potentiel constant à leur borne d'entrée. Ceci permet dans le cas d'une matrice sans sélecteurs (structure 1R) de limiter les courants parasites de fuite à travers les autres synapses et peut être réalisé en ayant par exemple recours à des circuits convoyeurs de courant de seconde génération[235].

Lors de chaque événement de lecture, le courant reçu par un neurone de sortie dépend de l'état (parallèle ou antiparallèle) de la synapse le reliant à l'entrée activée. Les neurones de sorties présentent deux caractéristiques :

1. Durant la lecture, à l'aide d'un amplificateur de détection (*sense-amplifier* en anglais), un neurone de sortie est capable de déterminer si l'impulsion qu'il reçoit de la part d'un neurone d'entrée a traversé une synapse dans l'état parallèle ou antiparallèle. Cette approche suppose que durant la durée d'une impulsion de lecture, plusieurs neurones d'en-

62. P. LICHTSTEINER, C. POSCH et T. DELBRÜCK, IEEE Journal of Solid-State Circuits, 2008.
234. W. S. ZHAO et coll., IEEE Transactions on Nanotechnology, 2012.
235. G. LECERF, J. TOMAS et S. SAÏGHI, IEEE International Symposium on Circuits and Systems, 2013.

trée ne peuvent pas être activés simultanément. Il s'agit de la situation rencontrée dans l'application que nous allons utiliser (traitement de données vidéo issues d'une rétine neuromorphique de type *Dynamic Vision Sensor*).

2. Les neurones de sortie intègrent l'information reçue de la part de leur amplificateur de détection. De façon fonctionnelle, ces neurones impulsionnels sont de type « intègre-et-tire avec fuites » (LIF). Comme nous l'avons vu dans le chapitre 1, ces derniers peuvent être réalisés par le biais de circuits analogiques ou numériques [13,16,17,236,237]. Il est par ailleurs possible de faire l'hypothèse d'un choix de conception selon lequel seules les impulsions d'entrée ayant traversé des synapses dans leur état parallèle sont intégrées par les neurones de sortie. De cette manière, les jonctions tunnel magnétiques dans leur état antiparallèle sont interprétées comme des synapses de poids nul, tandis que les jonctions dans leur état parallèle se comportent comme des synapses de poids unitaire. Une telle interprétation binaire des poids synaptiques permet notamment de lutter contre les effets de la variabilité entre composants ou du bruit.

Lorsqu'un neurone de sortie tire, ce dernier inhibe les autres neurones de sortie en remettant à zéro leur variable interne. En pratique, ce comportement peut être réalisé par une approche de type « plus proche voisin » telle qu'un réseau diffusif[238]. Pourvu de ce mécanisme d'inhibition latérale, notre architecture n'est pas sans rappeler un réseau de neurones de type *Winner-Takes-All*. Cette approche permet d'éviter qu'un grand nombre de neurones se spécialisent dans la détection d'un même motif, par exemple en raison d'une fréquence d'apparition supérieure de ce dernier.

3.1.2 Une règle d'apprentissage de type *Spike-Timing Dependent Plasticity* adaptée aux spécificités des composants

Lorsqu'un neurone de sortie émet un tir, le système bascule dans une phase d'écriture, durant laquelle sont appliquées sur la matrice d'éléments synaptiques des impulsions de tension permettant d'implémenter une règle d'apprentissage *adaptée* du modèle usuel[57,239] de la *Spike-Timing Dependent Plasticity*, observée en biologie par Bi et Poo (figure 1.10 page 24). Comme nous l'avons vu dans le chapitre 1, de nombreux travaux portant sur l'utilisation de composants mémoires innovants au sein de réseaux de neurones artificiels pour implémenter une règle d'apprentissage de type *Spike-Timing Dependent Plasticity* ont été proposés dans la

13. P. A. MEROLLA et coll., Science, 2014.
16. S. SAÏGHI et coll., IEEE Transactions on Biomedical Circuits and Systems, 2011.
17. G. INDIVERI et coll., Frontiers in Neuroscience, 2011.
236. S. MITRA, S. FUSI et G. INDIVERI, IEEE Transactions on Biomedical Circuits and Systems, 2009.
237. Y. WANG et S.-C. LIU, IEEE Transactions on Biomedical Circuits and Systems, 2013.
238. J. ARTHUR et K. BOAHEN, Advances in neural information processing systems, 2006.
57. G.-Q. BI et M.-M. POO, The Journal of neuroscience, 1998.
239. H. MARKRAM et coll., Science, 1997.

littérature[124,130,138,235,240–246], généralement en ayant recours à des stratégies de programmation relativement sophistiquées. Dans le cas présent, nous proposons au contraire une version simplifiée de la règle usuellement adoptée, dans une interprétation probabiliste en raison du caractère binaire des synapses que nous comptons employer[87]. Si la mise en œuvre de la *Spike-Timing Dependent Plasticity* n'est pas intrinsèquement réalisée par les jonctions tunnel magnétiques à transfert de spin, il s'agit néanmoins de mettre en place une règle d'apprentissage qui tire partie du comportement naturellement probabiliste de ces composants.

Notre réinterprétation de la *Spike-Timing Dependent Plasticity* procède à deux grandes hypothèses simplificatrices. Tout d'abord, une phase d'apprentissage a lieu uniquement lorsqu'un neurone de sortie tire, ce qui diffère de la plupart des modèles de *Spike-Timing Dependent Plasticity* mais s'avère plus aisé à mettre en place à l'aide de nanocomposants mémoires [23,88,138,246]. Dans le domaine des neurosciences, un choix similaire a été fait dans les travaux de B. NESSLER et coll.[247]. Ces travaux ont notamment montré que sous cette hypothèse, la *Spike-Timing Dependent Plasticity* peut être réinterprétée comme un algorithme d'apprentissage automatique de type espérance-maximisation [I], utilisé dans le domaine de l'apprentissage automatique. Partant de ce constat, D. QUERLIOZ a récemment montré la pertinence de ce type de règle d'apprentissage pour réaliser un système d'inférence se basant sur des composants mémoires innovants[248]. Par ailleurs, le mécanisme de *Spike-Timing Dependent Plasticity* que nous proposons est probabiliste (et non déterministe) : les nanocomposants synaptiques ont seulement une certaine *probabilité* d'évolution lors d'une phase d'apprentissage.

Dans la pratique, les phases de *Spike-Timing Dependent Plasticity* agissent comme des opérations de programmation du réseau synaptique, présentées à la figure 3.1b. Lorsqu'un neurone de sortie tire, le système entier entre dans une phase de programmation. Le neurone de sortie activé applique la forme d'onde de tension *(W2)* le long de sa ligne de la matrice d'éléments synaptiques, tandis que les entrées actives au cours d'une fenêtre de temps « récente » appliquent la forme d'onde de tension *(W1)* le long de leur colonne, et les autre entrées le signal *(W0)*. La combinaison de ces formes d'onde permet de mettre en œuvre notre interprétation de la *Spike-Timing Dependent Plasticity* : lorsque qu'un neurone de sortie tire, une jonction tunnel magnétique à transfert de spin qui lui est reliée présente (figure 3.2) :

124. S. YU et coll., IEEE Transactions on Electron Devices, 2011.
130. S. H. JO et coll., Nano Letters, 2010.
240. G. S. SNIDER, Nanotechnology, 2007.
241. G. S. SNIDER, Proceedings of the IEEE International Symposium on Nanoscale Architectures, 2008.
242. B. LINARES-BARRANCO et T. SERRANO-GOTARREDONA, Proceedings of the IEEE Conference on Nanotechnology, 2009.
243. F. ALIBART et coll., Advanced Functional Materials, 2012.
244. K. SEO et coll., Nanotechnology, 2011.
245. A. AFIFI, A. AYATOLLAHI et F. RAISSI, European Conference on Circuit Theory and Design, 2009.
246. D. QUERLIOZ, O. BICHLER et C. GAMRAT, International Joint Conference on Neural Networks, 2011.
87. W. SENN et S. FUSI, Physical Review E, 2005.
23. D. QUERLIOZ et coll., IEEE Transactions on Nanotechnology, 2013.
247. B. NESSLER et coll., PLoS Computational Biology, 2013.
I. *Expectation-Maximization* en anglais.
248. D. QUERLIOZ et coll., Proceedings of the IEEE, 2015.

– une certaine probabilité de basculer dans l'état parallèle si le neurone d'entrée connecté à son autre borne a été actif au cours d'une fenêtre de temps « récente » ;
– une certaine probabilité de basculer dans l'état antiparallèle si au contraire le neurone d'entrée connecté à son autre borne n'a pas été activé au cours de cette même fenêtre temporelle.

Dans le même temps, les jonctions tunnel magnétiques à transfert de spin connectées aux autres neurones de sortie sont soit non sélectionnées, soit sélectionnées à moitié, c'est-à-dire que la tension appliquée à leurs bornes est respectivement nulle ou $V_{prog}/2$[I]. En se basant sur les conclusions du modèle présenté à la partie 2.2.4.2 et en choisissant bien V_{prog}, il est possible d'obtenir une probabilité de commutation négligeable de ces éléments synaptiques non entièrement sélectionnés. La simplicité de la règle de *Spike-Timing Dependent Plasticity* proposée ci-dessus permet notamment d'utiliser des formes d'onde de tension plus simples que celles généralement employées avec d'autres types de familles de nanocomposants mémoires innovants[120].

Cette règle d'apprentissage probabiliste est similaire à celle proposée pour des synapses électrochimiques métalliques dans les travaux de SURI et coll.[88,132]. Dans notre cas, la forme des ondes de tension employées est légèrement plus simple, notamment grâce à la symétrie en tension des jonctions tunnel magnétiques à transfert de spin pour obtenir des commutations de l'état parallèle vers l'état antiparallèle (P→AP) ou inversement de l'état antiparallèle vers l'état parallèle (AP→P). Cette règle d'apprentissage n'est également pas sans rappeler la *Spike-Timing Dependent Plasticity* simplifiée proposée par NESSLER pour des études de neurosciences computationnelles[247], ici sous une forme binaire et probabiliste.

Nous verrons par la suite que même une règle de *Spike-Timing Dependent Plasticity* simple comme celle que nous proposons permet de réaliser un apprentissage non trivial de type non supervisé.

3.1.3 Méthodologie des simulations à l'échelle d'un système

Afin de démontrer l'intérêt des jonctions tunnel magnétiques à transfert de spin en tant que synapses artificielles, nous procédons à des simulations à l'échelle d'un système, dans lesquelles ces nanocomposants, et notamment leur comportement stochastique, sont modélisés selon les résultats de la partie 2.2.4.2.

3.1.3.1 Description générale et intérêt des outils de simulations employés

Nous utilisons un simulateur de réseau de neurones artificiels à impulsions, fonctionnant en pas de temps et spécialement développé au cours de la thèse. Le cœur de calcul de cet outil est écrit en C++ selon une approche par objets particulièrement adaptée à la description de l'architecture désirée. L'analyse des résultats est quant à elle effectuée à l'aide de scripts écrits

I. V_{prog} est la valeur maximale de la forme d'onde *(W2)* présentée à la figure 3.1b.
120. T. SERRANO-GOTARREDONA et coll., Frontiers in Neuroscience, 2013.

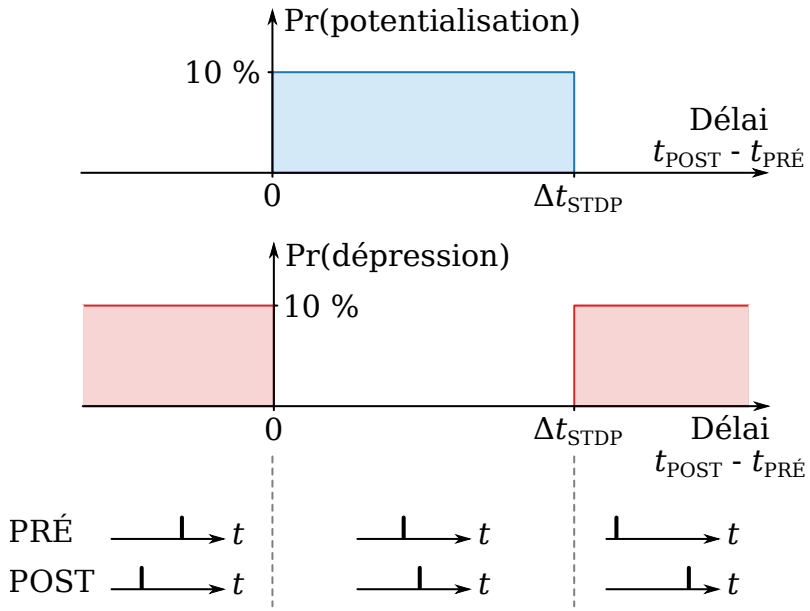


FIGURE 3.2 – Réinterprétation probabiliste de la règle d'apprentissage de type *Spike-Timing Dependent Plasticity* employée. Lorsqu'un délai $t_{POST} - t_{PRÉ}$ entre impulsions pré- et postsynaptiques positif et inférieur à la durée Δt_{STDP} est appliqué sur une synapse déprimée, cette dernière présente une probabilité (constante sur cet intervalle de temps) de se potentialiser (*panneau du haut, aire bleue*). De façon analogue, l'application d'un couple d'impulsions pré- et postsynaptiques de délai $t_{POST} - t_{PRÉ}$ situé hors de la fenêtre temporelle $[0, \Delta t_{STDP}]$ sur une synapse potentialisée entraîne la dépression de cette dernière selon une probabilité dont la valeur ne dépend pas davantage de celle du délai (*panneau du centre, aires rouges*). Les différentes situations de paires d'impulsions sont esquissées sur le panneau du bas.

en Python. Ce flot de travail permet d'associer la vitesse d'exécution d'un langage compilé pour le code de calcul, à la flexibilité d'un langage interprété et expressif pour le code servant à l'analyse des données produites. Le protocole de type *Address-Event Representation* utilisé pour les fichiers d'événements d'entrée et de sortie a été légèrement modifié par rapport à la version usuelle[61], principalement pour étendre la durée maximale des simulations réalisables sans nécessiter de dégrader la résolution temporelle.

Un intérêt majeur de ce simulateur est la possibilité d'utiliser des modèles proches de la physique des nanocomposants synaptiques employés, permettant notamment de prendre en compte leurs imperfections. Les neurones d'entrée et de sortie du système sont quant à eux décrits de façon purement *fonctionnelle* dans le simulateur et supposés réalisés en technologie CMOS le cas échéant. Ce type de simulations est significativement plus rapide que des simulations électriques SPICE[I]. Dans le cas de l'application que nous allons détailler, les simulations sur un processeur généraliste actuel sont généralement 25 à 75 fois plus lentes que le temps réel, selon le nombre d'éléments synaptiques simulés et le pas de temps employé. Ceci est suffisamment rapide pour rendre possible une étude statistique par le biais de simulations Monte-Carlo d'une tâche pratique.

3.1.3.2 Description des synapses artificielles et des impulsions associées

Les jonctions tunnel magnétiques à transfert de spin étudiées dans les simulations présentées ci-après sont basées sur une technologie 45 nm. Il s'agit des composants étiquetés « A » dans la table 2.1 page 79, à aimantation dans le plan, de section elliptique d'une largeur (totale) de 40 nm pour une longueur (totale) de 100 nm. L'épaisseur de la couche mince est de 2 nm, tandis que le coefficient d'amortissement de Gilbert $\alpha = 0{,}01$ et l'aimantation à saturation M_s est telle que $\mu_0 \times M_s = 1{,}5\,\mathrm{T}$. Enfin, la magnétorésistance tunnel choisie est de 150 % (c'est-à-dire que $R_{AP}/R_P = 2{,}5$).

La tension de programmation V_{prog} introduite à la figure 3.1b varie de 0,3 V à 0,6 V, de façon à étudier son influence sur le comportement du système. Dans le cas des jonctions tunnel magnétiques étudiées, ceci revient à faire évoluer l'amplitude de la densité de courant $|J_s|$ entre 95 et 190 $\mathrm{GA{\cdot}m^{-2}}$. Si l'on souhaite maintenir des probabilités de commutation identiques dans les différents régimes de programmation, il est nécessaire de modifier la durée t_p des impulsions de programmation selon la valeur V_{prog}, éventuellement sur plusieurs décades comme nous avons pu le voir à travers certains exemples du chapitre 2. Pour obtenir une probabilité de commutation de 10 % (identique pour les deux types de commutations, voir figure 3.2), la durée de programmation t_p s'étend ainsi de 1,88 ns (pour $V_{prog} = 0{,}6\,\mathrm{V}$) à 16,15 µs (pour $V_{prog} = 0{,}3\,\mathrm{V}$) dans le cas des jonctions tunnel magnétiques présentées juste avant.

Enfin, les impulsions de lecture employées ont une durée de 1 ns et une amplitude de 0,1 V,

61. K. A. BOAHEN, IEEE Transactions on Circuits and Systems II : Analog and Digital Signal Processing, 2000.
I. De l'anglais *Simulation Program with Integrated Circuit Emphasis*.

ce qui assure une probabilité négligeable de commutation parasite [I], inférieure à 10^{-18} dans le cas des jonctions tunnel magnétiques étudiées.

3.1.3.3 Dernières précisions sur les simulations réalisées

Notre tâche de démonstration s'appuie sur un enregistrement de 82 s issu d'une rétine neuromorphique de type *Dynamic Vision Sensor*[62], similaire à celle présentée dans la partie 1.1.3, observant le passage de véhicules sur une autoroute à six voies de circulation, près de Pasadena aux États-Unis. Cet enregistrement « vidéo » est librement accessible sur Internet[249]. Comme mentionné dans la partie 1.1.3, cette rétine artificielle présente un comportement inspiré de celui d'une rétine biologique : elle fonctionne de manière asynchrone (et non pas image par image comme une caméra conventionnelle), émettant des impulsions lorsque l'intensité d'un pixel change suffisamment. Chaque pixel de cette rétine est en réalité dédoublé en deux versions, l'une sensible à une augmentation de l'intensité lumineuse, l'autre à une diminution.

Chaque neurone d'entrée de notre système correspond à l'un des 32 768 pixels [II] de la rétine neuromorphique. L'architecture simulée possède 20 neurones de sortie, chacun d'entre eux étant connecté à chaque neurone d'entrée par le biais d'une (unique) jonction tunnel magnétique à transfert de spin. Un neurone de sortie qui vient d'émettre une impulsion de sortie est remis à zéro durant une période réfractaire de 0,5 s, tandis que les autres neurones de sortie le sont également, durant une période d'inhibition de 10 ms.

L'enregistrement issu de la rétine neuromorphique est présenté plusieurs fois [III], à vitesse réelle et sans temps mort entre les itérations. Du fait de la règle de *Spike-Timing Dependent Plasticity* simplifiée, chacun des neurones de sortie se spécialise naturellement dans la détection des véhicules circulant sur une voie particulière, comme nous pouvons le constater sur la figure 3.3 qui présente les impulsions postsynaptiques (traits pleins rouges) à la fin d'une simulation en comparaison des instants de passage des véhicules (pointillés bleus).

I. Dans la pratique, il est nécessaire que ces valeurs soient également compatibles avec la vitesse et la résolution du circuit de lecture.

249. jAER datasets, URL : https://sourceforge.net/p/jaer/wiki/AER%20data/.

II. $2 \times (128 \times 128) = 32\,768$

III. Pour le reste de ce chapitre, sauf en cas de mention contraire, cinq présentations successives des mêmes entrées sont utilisées. Les performances sont mesurées durant la dernière de ces présentations.

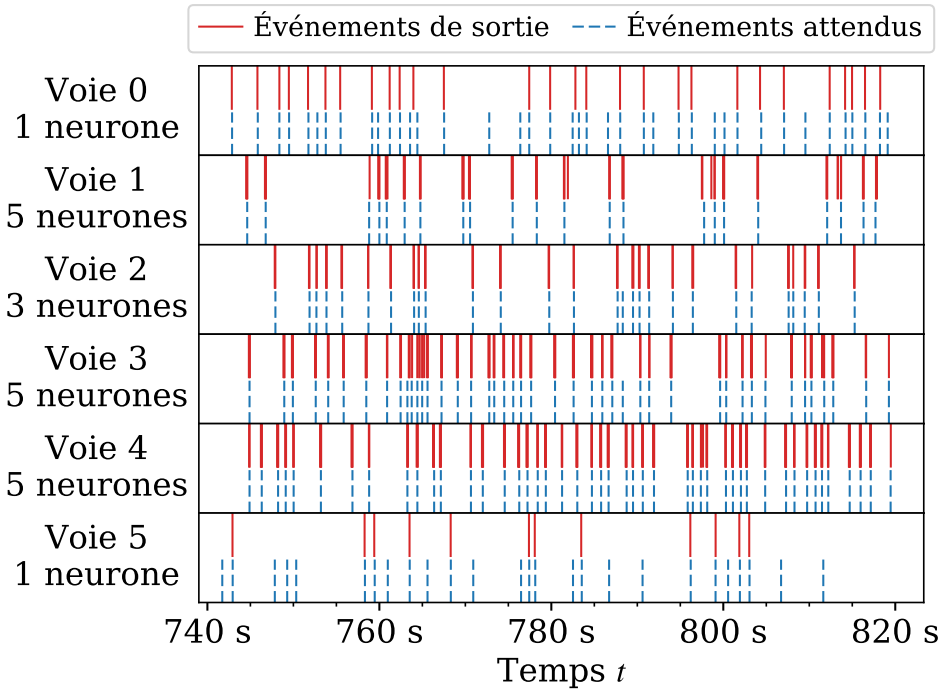


FIGURE 3.3 – Impulsions émises par les 20 neurones postsynaptiques (traits pleins rouges), lors de la dixième présentation des signaux d'entrées d'une simulation. Les traits bleus en pointillé indiquent les instants de passage des véhicules attendus voie par voie. Les neurones sont regroupés selon la voie de circulation dans laquelle ils semblent s'être spécialisés et leur nombre est indiqué à gauche. *NB* : les barres verticales rouges plus épaisses que les autres sont dues à la présence de plusieurs impulsions postsynaptiques temporellement proches.

3.2 Étude d'un système idéal

Dans un premier temps, nous nous plaçons dans une situation où les imperfections des différents éléments du système sont négligées. Exceptées les estimations de consommation électrique, dans l'hypothèse d'un tel système idéal sans variabilité ni synaptique ni neuronale et dont les signaux électriques ne sont sujets à aucun phénomène de bruit, les résultats seront les mêmes quel que soit le régime de courant adopté pour la programmation des jonctions tunnel synaptiques à transfert de spin, à condition de moduler la durée des impulsions selon le régime de programmation pour maintenir les mêmes probabilités de commutation. Dans cette partie, nous adoptons une tension de programmation $V_{\text{prog}} = 0{,}46\,\text{V}$ correspondant au régime de programmation intermédiaire, et la durée t_{p} des impulsions de programmation associées à chacun des deux types de commutations possibles est ajustée afin que les deux aient une probabilité de succès de 10 % (cas présenté à la figure 3.2).

3.2.1 Mise en évidence d'une spécialisation des neurones postsynaptiques

La spécialisation des neurones de sortie après apprentissage est visible sur la figure 3.3, mais également sur les cartes des états de conductance de la figure 3.4, qui montre la configuration de chacune des jonctions tunnel magnétiques à transfert de spin à l'issue d'une simulation. L'image du haut est une visualisation des données d'entrée (à un moment arbitraire de l'enregistrement) : toute entrée ayant été active durant une fenêtre (arbitraire) de 30 ms est représentée en teinte claire et dans une teinte sombre sinon. Les lignes jaunes en pointillé matérialisent les différentes voies de circulation de l'autoroute. Chacune des autres vignettes représente l'état à l'issue d'une simulation des jonctions tunnel magnétiques à transfert de spin connectées à un neurone de sortie donné (en blanc les synapses dans leur état parallèle et en noir dans leur état antiparallèle). Les neurones de sortie sont regroupés selon la voie de circulation dans laquelle ils se sont spécialisés (numérotées de 0 à 5 de la gauche vers la droite). Le nombre de neurones de sortie spécialisés dans une voie de circulation donnée n'est pas uniforme : dans l'exemple présenté à la figure 3.4, les voies 1, 3 et 4 comptent chacune cinq neurones de sortie spécialisés, la voie 2 trois et les voies 0 et 5 seulement un. Ceci dépend du nombre de véhicules en circulation sur chacune des voies, ainsi que de la quantité de pixels que ces derniers activent de leur passage. En particulier, moins de véhicules passent sur la voie 5 que sur les quatre voies centrales, et moins de pixels sont activés par les motifs de la voie 0.

La figure 3.5 montre, pour un neurone de sortie donné, comment un motif globalement stable émerge de la distribution aléatoire uniforme des états synaptiques initiaux. Au début de la simulation, les états des jonctions tunnel magnétiques à transfert de spin sont aléatoires, après 33 s un début de motif indique que le neurone de sortie considéré a commencé à se spécialiser dans la voie 4, tandis qu'au-delà de 114 s, le motif en question n'évolue plus de manière globale. Une signature indirecte de cette stabilisation d'ensemble des motifs est visible sur la figure 3.6, qui présente l'évolution au cours d'une simulation de la somme cumulée des commutations de

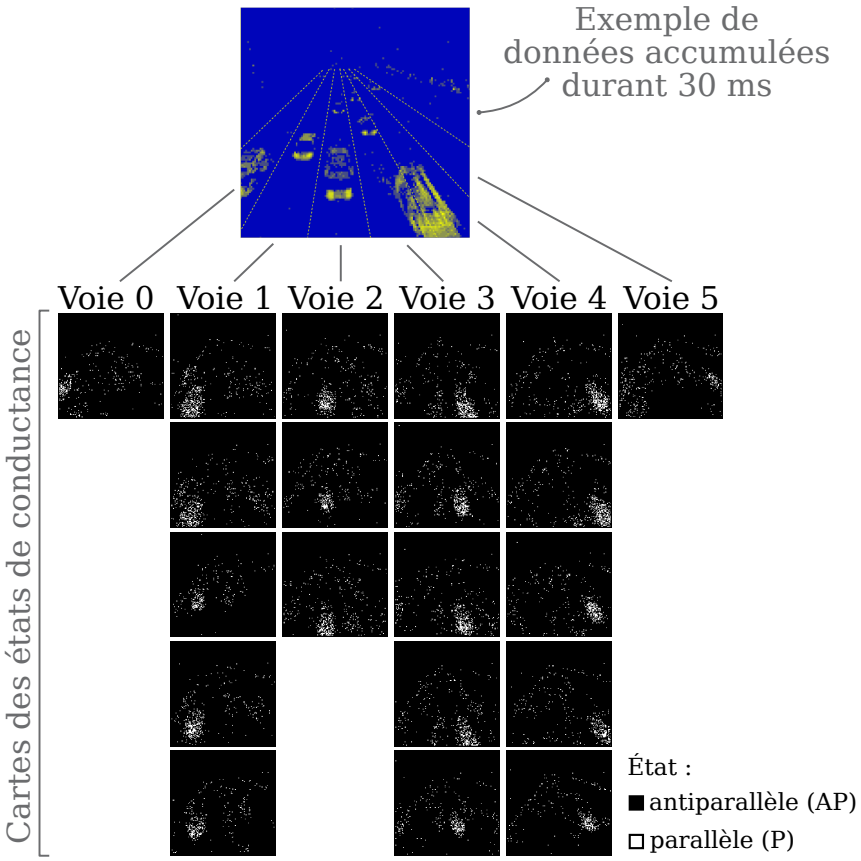


FIGURE 3.4 – *(En haut)* Image construite en accumulant en ton clair un type d'entrées présynaptiques durant 30 ms. Les lignes jaunes en pointillé ont été rajoutées pour illustrer les différentes voies de circulation. *(En bas)* Visualisation de l'état des jonctions tunnel magnétiques après apprentissage. À chaque neurone de sortie correspond une vignette particulière. Pour chacune de ces vignettes, la couleur d'une case indique l'état de la jonction tunnel magnétique connectée au pixel de mêmes coordonnées au sein de la rétine artificielle : blanche pour l'état parallèle (P), noire pour l'état antiparallèle (AP).

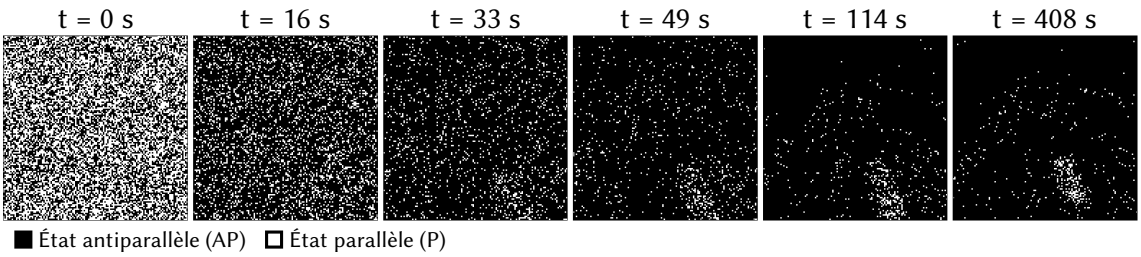


FIGURE 3.5 – Évolution durant la phase d'apprentissage de l'état de jonctions tunnel magnétiques connectées à l'un des neurones de sortie. Les conventions utilisées pour la représentation des cartes de conductance sont similaires à celles de la figure 3.4.

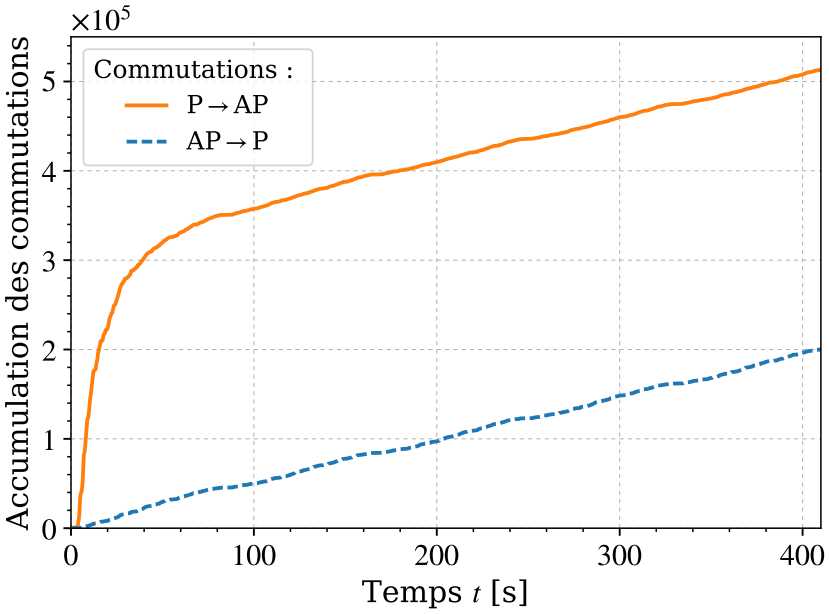


FIGURE 3.6 – Évolution des commutations synaptiques au cours d'une simulation. Le même fichier d'impulsions présynaptiques, d'une durée de 82 s, est présenté successivement 5 fois. L'accumulation au cours du temps du nombre de commutations de l'état antiparallèle vers l'état parallèle (AP → P, cas de potentialisation synaptique) est représentée en pointillé bleu, tandis que celle du nombre de commutations opposées, de l'état parallèle vers l'état antiparallèle (P → AP, cas de dépression synaptique), est quant à elle illustrée en trait plein orange.

chaque type possible. Dans un premier temps, jusqu'à environ 60 s, bien plus de dépressions synaptiques se produisent que d'événements de potentialisation, comme cela est également visible sur les quatre premières vignettes de la figure 3.5. Ensuite, les pentes moyennes des deux courbes de la figure 3.6 se stabilisent jusqu'à l'obtention de tendances parallèles, signe que des nombres similaires de chaque type de commutation ont lieu : à de petites fluctuations près, la conductance moyenne de la matrice synaptique n'évolue plus.

3.2.2 Quelques considérations pratiques

Une fois que l'apprentissage est achevé, il est possible de désactiver les mécanismes d'inhibition latérale et de programmation (donc l'apprentissage), ce qui peut constituer une stratégie de réduction de la consommation énergétique. La décision de désactiver l'inhibition latérale et l'apprentissage est essentiellement motivée par des considérations pragmatiques d'ingénierie. Toutefois, d'une certaine manière, cela pourrait également être rapproché de certains modèles biologiques tels que les mécanismes de modulation de type *attention-driven* ou *top-down*, qui affectent la plasticité synaptique et le couplage neuronal au cours du temps. Le fait de désactiver le mécanisme d'inhibition latérale augmente légèrement le taux de détection du système, notamment parce le système peut alors réagir à un passage presque simultané de deux véhicules

sur deux voies différentes, tandis qu'auparavant le passage du second véhicule était masqué par la période d'inhibition consécutive à l'activation du neurone de sortie ayant détecté le premier véhicule. En outre, désactiver l'apprentissage continu diminue la consommation énergétique du système, au détriment toutefois de la capacité à s'adapter à un changement des motifs récurrents dans les stimulus d'entrée.

Une propriété intéressante supplémentaire est le fait qu'une fois que le système a appris sa tâche, il est possible de l'éteindre et le rallumer sans qu'il n'oublie sa fonction, puisque les jonctions tunnel magnétiques à transfert de spin sont des cellules mémoires non volatiles.

3.2.3 Performances obtenues après apprentissage

Afin d'estimer les performances du système quant à sa tâche de détection de véhicules, nous avons choisi de retenir pour chaque voie uniquement le neurone de sortie présentant le meilleur taux de détection. Cette opération pourrait à priori être réalisée automatiquement, en ajoutant une seconde couche au système[250].

En interprétant le système (à l'issue de l'apprentissage) comme un détecteur de véhicules, le taux de détection est 97,3 % en moyenne pour les quatre voies de circulation intérieures. Les taux de détection des deux voies de circulation extérieures restantes sont quant à eux de 62 % et 28 %. La proportion d'événements de type faux positifs parmi les impulsions de sortie est de 4,7 %. Si l'apprentissage et le mécanisme d'inhibition sont conservés tout le temps actifs, alors le taux de détection moyen sur les quatre voies de circulation intérieures est légèrement réduit à 94,6 % tandis que la proportion de faux positifs monte à 5 %.

Le meilleur résultat publié à notre connaissance sur le même jeu de données[250] fait état d'un taux de détection de 98,1 % pour une proportion de faux positifs de 4,3 %, en ayant recours à un réseau de neurones utilisant des synapses analogiques en double précision. Les résultats obtenus par le système que nous avons présenté, qui exploite quant à lui des synapses binaires et stochastiques, apparaissent donc particulièrement encourageants.

3.2.4 Estimation de la consommation énergétique nécessaire à l'apprentissage

Durant le processus l'apprentissage de la matrice synaptique sans éléments de sélection (structure 1R) étudiée, la consommation moyenne de puissance requise pour programmer *uniquement* les jonctions tunnel magnétiques à transfert de spin [I] est estimée à 185 nW : un fonctionnement faible puissance semble donc envisageable. Nous montrerons dans la partie 3.3.3 que ce nombre dépend en réalité significativement du régime de programmation adopté.

Les courants parasites de fuite représentent seulement une faible proportion de la puissance consommée, à savoir 8,4 %. Ceci est dû au caractère hautement parallèle de l'opération d'écriture implémentant la règle d'apprentissage. Toutefois, il est à noter que la proportion de

250. O. BICHLER et coll., Neural Networks, 2012.

I. C'est-à-dire en ne tenant compte ni de la consommation des neurones ni de celle des autres circuits de contrôle extérieurs.

la puissance perdue du fait des courants parasites de fuite croît linéairement avec le nombre de neurones de sortie (cf. partie 5.3.2), ce qui pourrait être une limite possible dans le cas d'un système de très grande taille. La puissance moyenne totale[I] consommée est plus faible que celle d'une matrice de synapses à pont conducteur utilisée pour résoudre une tâche similaire[88]. Ceci s'explique notamment en raison de la faible tension requise par les jonctions tunnel magnétiques à transfert de spin ainsi que de la grande vitesse de programmation dans le régime probabiliste.

3.2.5 Une sensibilité limitée au rapport des probabilités de commutation

Il est intéressant de noter que les performances du système semblent seulement faiblement dépendre de la valeur précise de la probabilité de succès des commutations. Si par exemple nous ajustons la durée de programmation t_p pour que les commutations de l'état parallèle vers l'état antiparallèle (P → AP) aient une probabilité de réussite de 5 % (cette valeur demeurant égale à 10 % pour l'autre type de commutations), alors le taux de détection moyen pour les quatre voies de circulations intérieures, s'il chute à 86 %, ne tombe pas non plus très fortement.

3.3 Étude d'un système présentant de la variabilité

3.3.1 Influence des variations des composants synaptiques

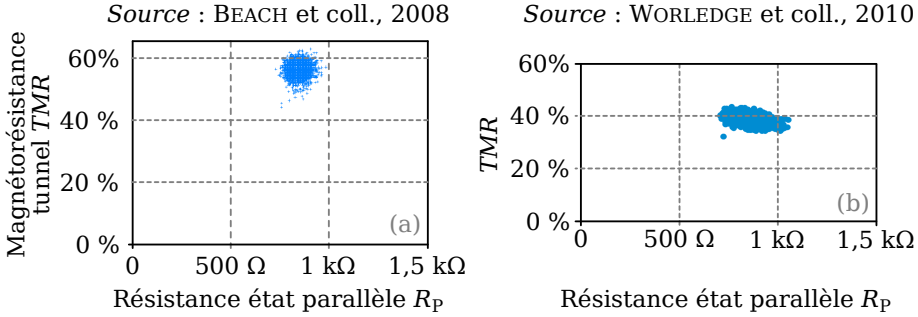


FIGURE 3.7 – Nuages de points expérimentaux de la magnétorésistance tunnel *TMR* et de la résistance dans l'état parallèle R_P, tirés des références [182] *(a)* et [184] *(b)* de la littérature. L'écart-type de chaque variable aléatoire est d'environ 5 % relativement à sa valeur moyenne et l'absence de tendance oblique est une signature de non corrélation entre ces deux grandeurs.

En pratique, parce qu'une certaine variabilité touche les nanocomposants, des jonctions tunnel magnétiques à transfert de spin différentes ont dans probabilités de commutation différentes, quand bien même elles sont soumises à des impulsions de programmation identiques. Des simulations Monte-Carlo ont été réalisées de façon à évaluer la robustesse de notre approche à ce problème.

I. Les valeurs données regroupent la puissance effectivement liée à la programmation synaptique mais également celle due aux courants parasitesde fuite.

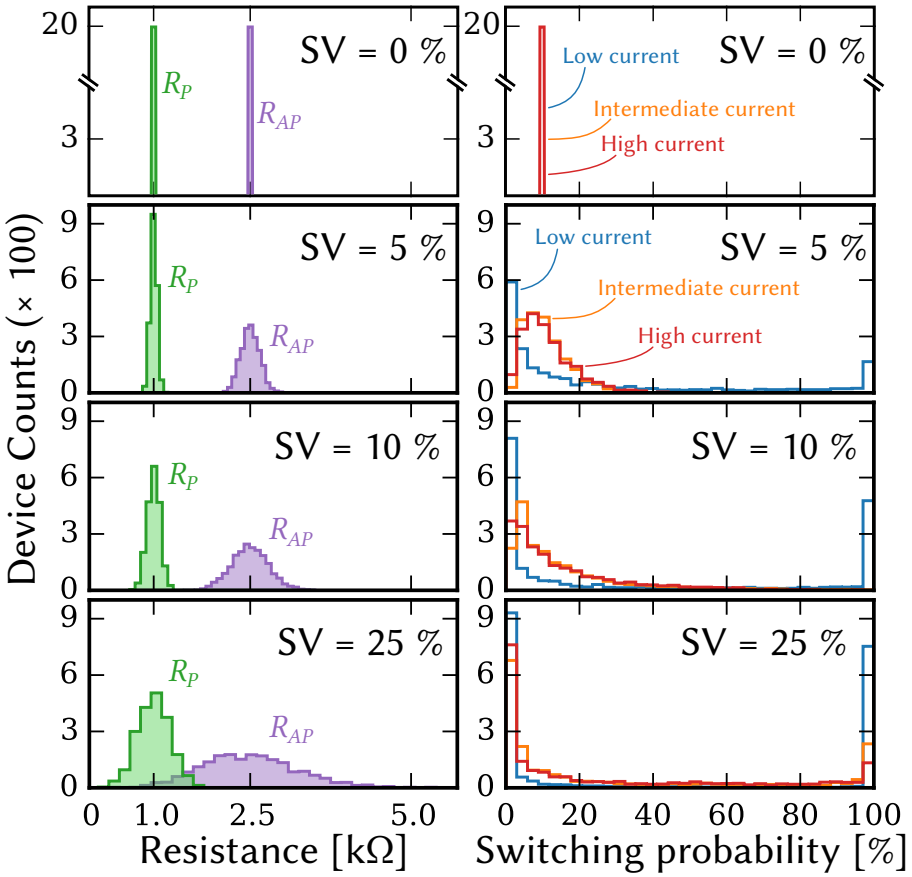


FIGURE 3.8 – Histogrammes (sur 2000 jonctions tunnel magnétiques) des valeurs de résistance dans l'état parallèle R_P et antiparallèle R_{AP} (*panneau de gauche*) en présence de variabilité synaptique SV, ainsi que des probabilités de commutation des ces mêmes composants (*panneau de droite*). Les probabilités de commutation sont tracées dans les trois régimes de courant, faible, intermédiaire et fort, obtenus avec $V_{prog} = 0{,}3\,V$, $0{,}4\,V$ et $0{,}6\,V$ respectivement. La durée des impulsions de programmation est ajustée dans chaque régime pour obtenir une probabilité de commutation de 10 % lorsque la variabilité synaptique est nulle ($SV = 0$). De haut en bas, la variabilité synaptique SV est égale à un coefficient de variation nul puis de 5 %, 10 % et 25 % des valeurs de résistance dans l'état parallèle R_P et de magnétorésistance tunnel TMR (tirées de façon indépendante).

Nous avons considéré des variations portant sur les valeurs de résistance de jonctions tunnel magnétiques à transfert de spin dans les états parallèle (R_P) et antiparallèle (R_{AP}). Pour ce faire, de la dispersion gaussienne est ajoutée (de façon indépendante) sur les valeurs de résistance dans l'état parallèle R_P et de magnétorésistance tunnel TMR. Nous nommons « variabilité synaptique » (SV) le coefficient de variation[I] des valeurs de résistance dans l'état parallèle R_P et de la magnétorésistance tunnel TMR. Cette façon d'introduire de la variabilité sur les composants se base sur des observations expérimentales rapportées dans la littérature et reproduites à la figure 3.7, qui suggèrent que les variations affectant la résistance dans l'état parallèle R_P et la magnétorésistance tunnel TMR ne sont pas corrélées et présentent des coefficients de variations similaires[182,184].

Étant donné qu'elles modifient la densité de courant circulant à travers les jonctions, les variations considérées ont une influence très importante sur les probabilités de commutation lorsque ces composants sont soumis à des impulsions de programmation, comme nous pouvons le constater avec la figure 3.8. Les vignettes de gauche de cette figure présentent les histogrammes des valeurs de résistance dans les états parallèle et antiparallèle, respectivement R_P et R_{AP}. Les vignettes de droite présentent les histogrammes des probabilités de commutation de ces jonctions tunnel magnétiques à transfert de spin, lorsque ces dernières sont soumises à des impulsions de programmation conçues pour assurer 10 % de chances de succès à variabilité synaptique nulle ($SV = 0$). Ces probabilités sont calculées à l'aide du modèle analytique de la partie 2.2.4.2. Trois histogrammes sont superposés, présentant les comportements dans les différents régimes de courant : faible, intermédiaire et fort. Nous pouvons observer sur cette figure que la variabilité dont souffrent les probabilités de commutations est sensiblement plus importante que celle qui a été ajoutée aux états de résistance. Par ailleurs, la variabilité des probabilités de commutation est considérablement plus forte dans le cas d'une programmation à courant faible que dans les cas des régimes de courants plus élevés (intermédiaire et fort).

3.3.2 Performances

Lors de simulations à l'échelle d'un système complet, nous observons une tolérance remarquable de ce dernier aux variations affectant les éléments synaptiques.

Considérons dans un premier temps le cas de jonctions tunnel magnétiques à transfert de spin programmées dans leur régime intermédiaire de courant. La figure 3.9a trace le taux de détection des véhicules en fonction de la variabilité synaptique SV qui affecte les jonctions tunnel magnétiques dans deux situations distinctes : lorsque les mécanismes de programmation de type *Spike-Timing Dependent Plasticity* et d'inhibition latérale sont éteints à l'issue de la phase d'apprentissage, et lorsqu'ils ne le sont pas. En l'absence de variabilité synaptique ($SV = 0$), la situation pour laquelle les mécanismes précités sont tout le temps actifs présente de meilleures

I. Pour rappel, le coefficient de variation d'une grandeur aléatoire est le rapport de son écart-type par sa valeur moyenne.

182. R. BEACH et coll., IEEE International Electron Devices Meeting, 2008.

184. D. C. WORLEDGE et coll., IEEE International Electron Devices Meeting, 2010.

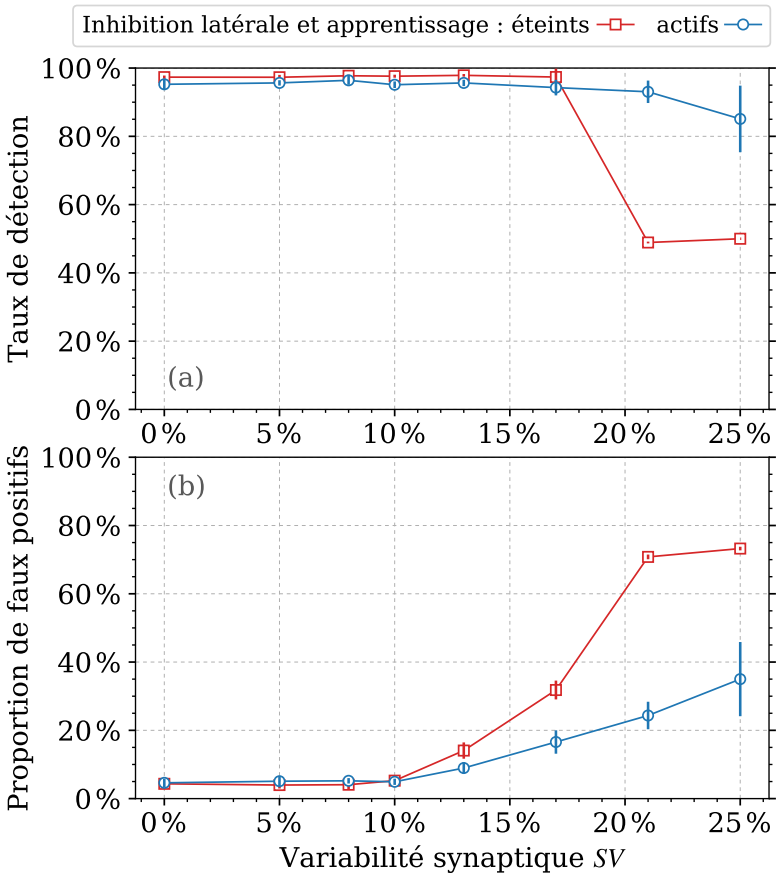


FIGURE 3.9 – Taux de détection des véhicules (a) et proportion d'événements faux positifs (b) en fonction de la variabilité synaptique SV, dans une situation pour laquelle l'apprentissage et le mécanisme d'inhibition latérale sont désactivés à l'issue du processus d'apprentissage (□) ou non (○). Chaque simulation est répétée dix fois et les barres d'erreur représentent une fois l'écart-type obtenu.

performances que la situation où ces derniers sont arrêtés après la phase d'apprentissage. Dans les deux situations, une variabilité synaptique SV de 17 % peut être atteinte sans effet notable sur les performances de détection du système, quand bien même la variabilité associée aux probabilités de commutation est alors considérable (comme présenté à la figure 3.8). La robustesse du système est ainsi particulièrement notable. Lorsque la variabilité synaptique SV dépasse les 17 %, le taux de détection diminue abruptement dans la situation où la mise en œuvre de l'apprentissage et l'inhibition latérale ont été désactivées à l'issue de l'apprentissage, alors que le déclin est moins rapide lorsque ces deux mécanismes sont gardés actifs. La situation pour laquelle l'inhibition latérale n'est pas désactivée est plus robuste car dans le cas d'une forte variabilité des éléments synaptiques, certains neurones de sortie sont activés préférentiellement, au détriment des autres neurones de sortie dont le mécanisme d'inhibition latérale limite l'activité.

La figure 3.9b présente l'évolution de la proportion d'événements de type faux positifs avec la variabilité synaptique SV, dans les mêmes conditions que celle de la figure 3.9a. Des tendances similaires peuvent être observées, bien que le nombre d'événements de type faux positifs commence à augmenter significativement au-delà d'une variabilité synaptique SV de 13 %, une valeur un peu plus faible que pour le taux de détection.

Il est à noter que les références [182, 184] font état d'une variabilité synaptique d'environ 5 %. Les degrés de variabilité que nous considérons sont donc dans l'ensemble sensiblement plus élevés que ceux rencontrés dans une technologie réaliste.

Nous avons également étudié le cas de fluctuations transitoires des propriétés des jonctions tunnel magnétiques à transfert de spin, comme peuvent par exemple en entraîner des variations de température éphémères. Pour ce faire, nous avons réalisé des simulations au cours desquelles, *lors de chaque tentative de programmation d'une jonction tunnel magnétique*, la résistance dans l'état parallèle R_P et la magnétorésistance tunnel TMR de la jonction en question sont tirées aléatoirement et indépendamment, chacune avec une dispersion gaussienne de coefficient de variation égal à 10 %. Ceci revient à introduire un degré artificiellement élevé de variations puisque les caractéristiques des jonctions tunnel magnétiques s'avèrent expérimentalement stables d'un cycle de programmation à l'autre. Néanmoins, il se trouve que même dans de telles conditions, le système considéré présente un taux de détection des véhicules et une proportion d'événements de type faux positifs analogues à la situation où les caractéristiques des jonctions tunnel magnétiques sont invariantes dans le temps.

3.3.3 Influence du régime de programmation

Nous avons vu dans la partie 2.2.4.2 qu'il possible d'utiliser les jonctions tunnel magnétiques à transfert de spin dans différents régimes de programmation, à savoir à courant faible, intermédiaire ou fort. Dans cette partie, les simulations à l'échelle d'un système nous permettent de mieux comprendre les avantages et les inconvénients de ces différents régimes quant à l'utilisation de jonctions tunnel magnétiques à transfert de spin comme synapses artificielles.

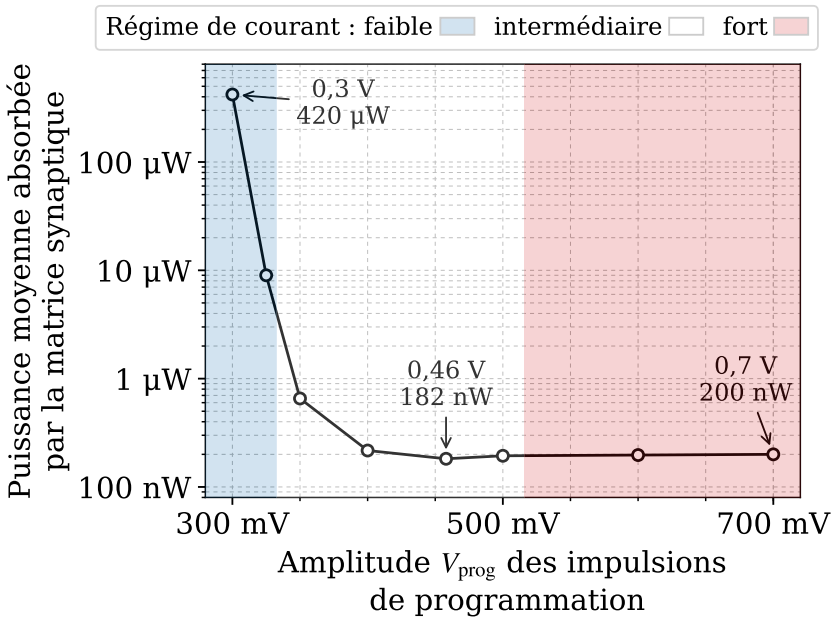


FIGURE 3.10 – Évolution, en fonction de l'amplitude de la tension de programmation V_{prog} employée, de la puissance moyenne consommée par l'ensemble des 655 360 synapses pour leur programmation durant l'apprentissage. Les différentes aires colorées indiquent la nature du régime de courant correspondant : faible (bleu), intermédiaire (blanc) et fort (rouge).

Rappelons tout d'abord que sans variabilité synaptique ($SV = 0$), les trois régimes de programmation sont analogues puisqu'il n'y a alors pas de dispersion des probabilités de commutation : le taux de détection des véhicules et la proportion d'événements de type faux positifs sont identiques d'un régime à l'autre. Toutefois, il n'en est pas de même en ce qui concerne l'énergie requise pour programmer les jonctions tunnel magnétiques à transfert de spin, qui dépend significativement du régime de programmation comme l'illustre la figure 3.10. La puissance moyenne consommée par la matrice d'éléments synaptiques durant la phase d'apprentissage est minimale dans le régime intermédiaire de courant, avec moins de 185 nW lorsque $V_{prog} = 0{,}46\,V$. Pour des valeurs de tension de programmation plus élevées, la puissance moyenne consommée augmente légèrement avec la tension, atteignant par exemple 200 nW quand $V_{prog} = 0{,}6\,V$ (régime de fort courant). Lorsque la tension de programmation diminue, la puissance moyenne consommée croît également, notamment de façon très abrupte dans le régime de faible courant : dans le cas étudié, la matrice d'éléments synaptique consomme en moyenne 0,42 mW lorsque $V_{prog} = 0{,}3\,V$! En effet dans ce régime de programmation, bien que la diminution de l'amplitude des impulsions tende à réduire de façon quadratique la consommation énergétique, la durée nécessaire au maintien de la probabilité de commutation croît quant à elle exponentiellement.

Par ailleurs, en présence d'une variabilité synaptique SV non nulle, les trois régimes de programmation ne sont plus équivalents, comme nous pouvons le constater avec les simula-

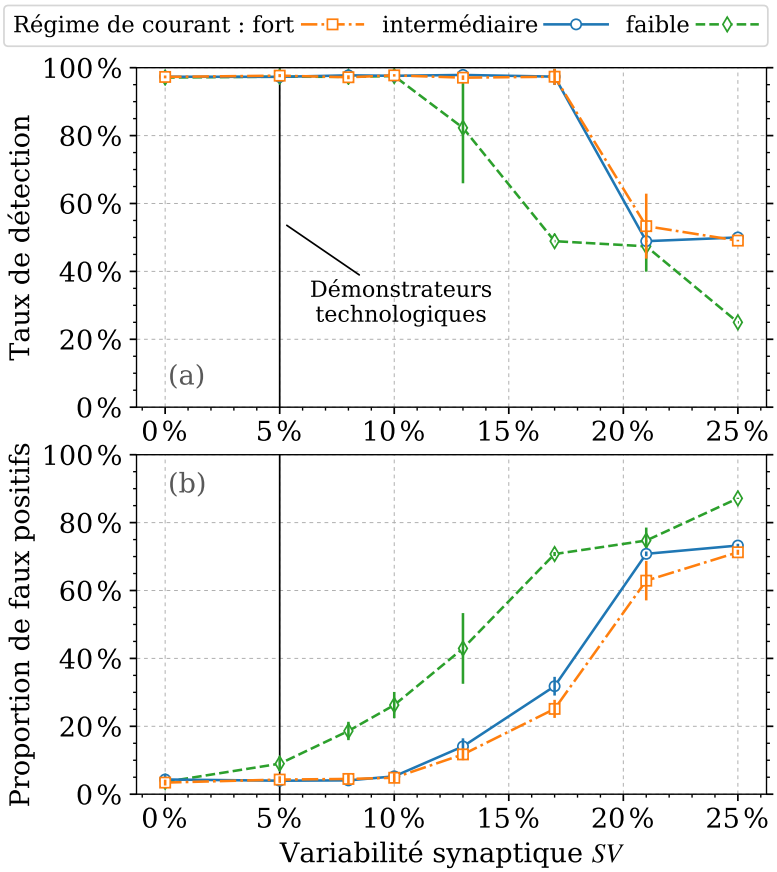


FIGURE 3.11 – Taux de détection des véhicules (a) et proportion d'événements de type faux positifs (b) en fonction de la variabilité synaptique SV, lorsque les jonctions tunnel magnétiques à transfert de spin sont programmées dans leur régime de courant faible (◇), intermédiaire (○) ou fort (□). La programmation de type *Spike-Timing Dependent Plasticity* et le mécanisme d'inhibition latérale sont désactivés à l'issue de la phase d'apprentissage. Chaque simulation est répétée dix fois et les barres d'erreur représentent une fois l'écart-type obtenu. Le trait plein vertical noir indique la variabilité touchant les démonstrateurs technologiques de la littérature[182,184].

tions Monte-Carlo de la figure 3.11. Ainsi, si dans le cas des régimes de courant intermédiaire et fort, le taux de détection des véhicules s'avère relativement robuste jusqu'à atteindre une variabilité synaptique SV de 17 %, ce dernier diminue significativement dès une variabilité synaptique SVde 10 % dans le régime de faible courant. De même, jusqu'à atteindre 10 % de variabilité synaptique SV, la proportion d'événements de type faux positif évolue peu dans les deux régimes de plus fort courant, alors que dans le cas du régime de faible courant, cette proportion augmente dès que la variabilité synaptique SV s'accroît.

Ce contraste peut s'expliquer par la physique des jonctions tunnel magnétiques à transfert de spin. Comme nous l'avons vu dans la partie 2.2.4.2, dans le régime de faible courant, le délai moyen avant commutation d'une jonction tunnel magnétique à transfert de spin présente une dépendance exponentielle à la densité de courant qui traverse cette dernière. En présence de variabilité entre les composants, chaque jonction tunnel magnétique est alors programmée avec une densité de courant différente, ce qui a pour effet de disperser les probabilités de commutation de façon importante, contrairement aux régimes de programmation à courant intermédiaire ou fort, dans lesquels la dépendance du délai moyen avant commutation avec la densité de courant est moins abrupte.

Pour résumer, les régimes de programmation à courant intermédiaire ou fort nécessitent une puissance plus faible lors de la phase d'apprentissage et tolèrent mieux la variabilité entre les composants que le régime de programmation à faible courant. Étant donné que le régime de programmation intermédiaire requiert des tensions plus faibles que le régime à fort courant, les jonctions tunnel magnétiques à transfert de spin devraient présenter une meilleure fiabilité dans ce régime. Le régime de programmation à courant intermédiaire apparaît ainsi être le régime optimal pour utiliser des jonctions tunnel magnétiques à transfert de spin comme synapses artificielles. Ceci diffère des applications mémoires plus traditionnelles des jonctions tunnel magnétiques à transfert de spin, pour lesquelles la vitesse de programmation est le premier critère d'intérêt et un courant de programmation plus élevé alors un choix préférable.

3.4 Conclusions

Possibilité d'apprentissage à l'échelle d'un système. Dans ce chapitre, nous avons présenté des simulations informatiques fonctionnelles d'un système neuromorphique de grande échelle qui utilise des jonctions tunnel magnétiques à transfert de spin comme synapses artificielles binaires. Ces travaux ont montré la possibilité de remplir une tâche cognitive réaliste en utilisant une règle d'apprentissage de type *Spike-Timing Dependent Plasticity*, adaptée selon une interprétation probabiliste afin de tirer profit du comportement intrinsèquement stochastique de ces nanocomposants mémoires innovants.

Les synapses étudiées, constituées chacune d'une unique jonction tunnel magnétique, semblent particulièrement intéressantes pour s'attaquer à des tâches dynamiques, comme la détection de véhicules présentée, puisque les performances obtenues approchent celles de systèmes

se basant sur des synapses analogique cumulatives. Dans le cas de tâches plus statiques, telles que la classification de chiffres manuscrits, des travaux conceptuels récents[248] suggèrent la nécessité d'une certaine redondance des nanocomposants stochastiques constituant les synapses pour approcher les performances de synapses cumulatives. Toutefois, le niveau de redondance reste modéré pour la tâche présentée : il semble pertinent d'envisager étudier l'apprentissage de ce type de tâches en utilisant des synapses stochastiques constituées chacune d'un faible nombre de jonctions tunnel magnétiques à transfert de spin.

Des résultats encourageants pour la conception de futurs systèmes neuromorphiques. Les simulations présentées dans ce chapitre suggèrent également qu'un système neuromorphique exploitant des jonctions tunnel magnétiques à transfert de spin comme synapses puisse être tolérant à plusieurs catégories d'imprécisions touchant ces dernières. Nous avons ainsi pu observer le maintien de performances raisonnables voire analogues à celle d'un système idéal dans des situations de variabilité des caractéristiques entre composants synaptiques, de fluctuations transitoires de ces dernières ou encore de variation du rapport des chances de succès des deux types commutations. Par ailleurs, la consommation énergétique liée à la programmation des synapses peut atteindre des valeurs plus faibles que celles rapportées dans la littérature pour des architectures similaires qui traitent la même tâche cognitive mais utilisent d'autres composants mémoires innovants comme synapses artificielles.

De l'importance du régime de programmation. Disposer d'un modèle avancé du comportement stochastique que nous exploitons au sein des jonctions tunnel magnétiques à transfert de spin nous a permis d'évaluer dans quelle mesure le régime de programmation influence les performances du système. Ainsi, le régime de faible courant apparaît peu propice à un usage synaptique puisqu'il souffre rapidement de la variabilité des caractéristiques des synapses et présente une consommation énergétique qui peut devenir importante en comparaison de systèmes déjà présentés dans la littérature. En revanche, les régimes de courant intermédiaire et fort entraînent des comportements proches et intéressants. Ils maintiennent notamment un excellent niveau de performances en présence de variabilité synaptique, même pour des valeurs qui excèdent celles de démonstrateurs technologiques mémoires de la littérature. Par ailleurs, dans ces deux régimes, la programmation des synapses requiert une énergie inférieure aux valeurs rapportées dans la littérature pour un système similaire réalisant la même tâche cognitive, avec un léger avantage au régime de courant intermédiaire.

À la suite des résultats présentés précédemment, le régime de courant intermédiaire apparaît comme le régime de programmation à privilégier pour utiliser des jonctions tunnel magnétiques à transfert de spin comme synapses artificielles binaires et stochastiques dans le cadre de tâches cognitives similaires à celle employée dans ce chapitre.

Chapitre 4

Synapses électrochimiques métalliques : plasticités multiples et apprentissage

« L'expérience est une mémoire, mais l'inverse est vrai. »

Alfred CAMUS

“CE CHAPITRE *porte sur l'étude du potentiel d'usage synaptique d'une variété particulière de cellules électrochimiques métalliques, à base d'un électrolyte de sulfure d'argent (Ag_2S). Ces composants mémristifs offrent des comportements plastiques différents de ceux des jonctions tunnel magnétiques étudiées précédemment, dont notamment une forme originale et intrinsèque de plasticité hebbienne que nous avons modélisée analytiquement. Des simulations fonctionnelles d'un système exploitant des synapses décrites par ce modèle analytique sont ensuite utilisées pour réaliser une preuve de concept d'un apprentissage qui exploite les différents comportements plastiques intrinsèques offerts par ces synapses pour simplifier la gestion externe des phases d'écriture. Ces premières études sont également l'occasion d'évaluer les effets du bruit sur les entrées présynaptiques ou encore de la variabilité entre éléments synaptiques sur l'apprentissage d'un tel système.* ”

4.1 Introduction

4.1.1 Vers des synapses artificielles au comportement plastique plus riche

D'un intérêt pour des synapses artificielles de plasticité intrinsèquement riche. Les nouvelles technologies mémristives constituent potentiellement un élément charnière du développement des solutions matérielles de calcul neuro-inspiré, en offrant des comportements plastiques complexes qui permettent d'envisager en faire des éléments synaptiques compacts et capables d'apprendre. Le coût important en surface de silicium des synapses artificielles réalisées dans une technologie CMOS, que nous avons pu observer dans le cadre des réalisations matérielles présentées dans le chapitre 1, limite en effet fortement la densité d'intégration de ces composants clefs, d'autant plus si la possibilité d'un apprentissage *in situ* est requise. Nous avons alors montré dans le chapitre 3 l'intérêt d'exploiter le comportement intrinsèquement stochastique des jonctions tunnel magnétiques à transfert de spin pour réaliser de manière simple un apprentissage probabiliste sans le recours à des circuits supplémentaires de génération de nombres aléatoires.

Dans le présent chapitre, nous nous intéressons à un autre type de composants mémoires innovants, qui appartient à la famille des cellules électrochimiques métalliques et offre un comportement plastique riche (cette fois-ci déterministe). Nous allons montrer qu'il est possible d'exploiter cette richesse pour mettre en œuvre un apprentissage non supervisé. Contrairement à la situation du chapitre 3, nous verrons que ces nanocomposants sont *naturellement* capables de réaliser la dépendance temporelle d'une règle d'apprentissage de type *Spike-Timing Dependent Plasticity*. Par ailleurs, un intérêt de cette technologie académique face à celles présentées dans le chapitre 1 est l'extrême simplicité des impulsions de programmation requises, qui rend moins complexes les circuits de contrôle externes associés.

De multiples comportements plastiques. Il existe des propositions de composants mémristifs qui permettent d'obtenir une plasticité synaptique à court terme[251] (STP[I]). Cette dernière est obtenue en exploitant la volatilité d'une technologie mémoire qui, à l'issue d'une potentialisation ou d'une dépression, présente un phénomène de relaxation vers un état stable sur une échelle de temps courte (typiquement de la milliseconde à la seconde). Néanmoins, une plasticité à court terme seule est d'un intérêt limité en termes de capacités d'apprentissage et de calcul, ne pouvant être exploitée que pour des tâches simples[252,253].

Une capacité plastique plus complexe est la possibilité de transition d'une plasticité à court terme vers une plasticité à long terme (LTP[II]). Il s'agit d'un comportement observé dans

251. H. LIM et coll., Nanotechnology, 2013.
I. De l'anglais *Short-Term Plasticity*.
252. L. F. ABBOTT et coll., Science, 1997.
253. R. BERDAN et coll., IEEE Electron Device Letters, 2014.
II. De l'anglais *Long-Term Plasticity*.

plusieurs nanocomposants mémoires innovants[133,254,255] : une potentialisation faible de la cellule mémoire est suivie d'une relaxation rapide de sa conductance, tandis qu'une potentialisation plus forte permet d'obtenir une durée de rétention sensiblement plus élevée. Il est possible d'observer ce phénomène lorsque la relaxation de la conductance électrique d'une cellule mémoire est modulée par la valeur de conductance atteinte après un événement de potentialisation. Si cet effet n'est pas sans rappeler certains modèles de consolidation des souvenirs[256], son potentiel d'application reste encore une fois limité.

L'apprentissage hebbien[I], durant lequel la corrélation entre activité présynaptique et activité postsynaptique module l'évolution des synapses, est quant à lui considéré comme une explication possible au développement de capacités cognitives complexes[25]. L'une des familles de mécanismes synaptiques hebbiens les plus courantes est celle de type *Spike-Timing Dependent Plasticity* (STDP), que nous avons présentée aux chapitres 1 et 3, selon laquelle la distance temporelle entre une impulsion présynaptique et une impulsion postsynaptique définit le changement de conductance de la synapse reliant les deux neurones impliqués[239]. La plupart des stratégies utilisant des nanocomposants mémristifs pour mettre en œuvre un mécanisme de *Spike-Timing Dependent Plasticity* reposent sur la superposition temporelle d'impulsions neuronales aux formes d'onde complexes, spécialement conçues pour convertir la distance temporelle en une tension électrique[120,130,243,257,258]. Une stratégie plus intéressante est de se reposer sur une dynamique interne à la cellule mémoire, en n'ayant alors recours qu'à des impulsions neuronales de formes d'onde simples. Toutefois, seul un faible nombre de matériaux possède un tel comportement naturel de type *Spike-Timing Dependent Plasticity*[131,259] et permet de réduire la complexité des circuits de contrôle associés à la règle d'apprentissage.

Utiliser plusieurs comportements plastiques à la fois. Il a récemment été suggéré[260] que l'exploitation de plusieurs comportements plastiques différents au sein d'une même synapse pourrait améliorer les capacités d'apprentissage des systèmes neuro-inspirés.

Dans ce chapitre, nous nous intéressons aux cellules électrochimiques métalliques[128,254] (ECM). Les composants retenus utilisent un électrolyte solide de sulfure d'argent (Ag_2S) et

133. S. La Barbera, D. Vuillaume et F. Alibart, ACS Nano, 2015.
254. T. Ohno et coll., Nature materials, 2011.
255. T. Chang, S.-H. Jo et W. Lu, ACS Nano, 2011.
256. R. M. Shiffrin et R. C. Atkinson, Psychological Review, 1969.
I. C'est-à-dire suivant (au moins de façon générale) la règle de Hebbs.
25. D. O. Hebb, The organization of behavior : A neuropsychological approach, 1949.
239. H. Markram et coll., Science, 1997.
120. T. Serrano-Gotarredona et coll., Frontiers in Neuroscience, 2013.
130. S. H. Jo et coll., Nano Letters, 2010.
243. F. Alibart et coll., Advanced Functional Materials, 2012.
257. D. Kuzum et coll., Nano letters, 2011.
258. M. Suri et coll., Journal of Applied Physics, 2012.
131. C. Du et coll., Advanced Functional Materials, 2015.
259. S. Kim et coll., Nano letters, 2015.
260. F. Zenke, E. J. Agnes et W. Gerstner, Nature Communications, 2015.
128. R. Waser et coll., Advanced Materials, 2009.

présentent une transition de plasticité de court terme à plus long terme[133]. Nous avons par ailleurs découvert au cours de cette thèse l'existence d'un comportement plastique supplémentaire, de type *Spike-Timing Dependent Plasticity*. À l'aide de simulations exploitant un modèle analytique de ces cellules électrochimiques métalliques, nous présenterons une preuve de concept d'un apprentissage qui s'appuie sur les différentes plasticités offertes par ces composants mémoires innovants.

4.1.2 Contexte du projet

Les travaux présentés dans ce chapitre sont le résultat d'une collaboration avec Selina LA BARBERA et Fabien ALIBART, à l'époque respectivement doctorante et chargé de recherche à l'Institut d'Électronique, de Microélectronique et de Nanotechnologie (IEMN) de l'Université de Lille I.

La fabrication des échantillons, leur caractérisation électrique ainsi que l'élaboration d'un modèle analytique rendant compte de leur transition d'une plasticité à court terme vers une rétention plus longue ont toutes été réalisées par S. LA BARBERA. Le présent travail de thèse a quant à lui essentiellement porté sur deux volets :

1. l'élaboration, en collaboration avec S. LA BARBERA, d'un modèle analytique prenant en compte la plasticité supplémentaire de type *Spike-Timing Dependent Plasticity* ;
2. le développement de simulations numériques à l'échelle d'un système, permettant d'obtenir une preuve de concept d'un apprentissage reposant sur l'interaction de plusieurs comportements plastiques.

4.1.3 Description des composants

Constitution des cellules. Les cellules électrochimiques métalliques ont été fabriquées selon une configuration en « point de contact » (*crosspoint* en anglais), de 200 nm par 200 nm. Elles sont constituées d'une électrode inférieure en titane et platine (Ti/Pt), d'une couche d'électrolyte solide Ag_2S et d'une électrode supérieure en argent (Ag). La figure 4.1 présente une vue au microscope électronique à balayage de l'un des composants fabriqués.

Fabrication des échantillons. Les électrodes inférieures ont été dessinées par lithographie électronique puis les dépôts métalliques (5 nm de Ti et 30 nm de Pt) ont été réalisés par évaporation sous vide. Une couche de 60 nm d'électrolyte solide Ag_2S a été déposée depuis une source de matériau pur, par évaporation thermique dans un vide secondaire. L'épaisseur de cette couche a été contrôlée *in situ* par une balance à quartz. Enfin, les électrodes métalliques supérieures en Ag ont également été dessinées par lithographie électronique et déposées par évaporation sous vide.

Les mesures électriques ont été réalisées à l'aide d'un analyseur de paramètres Agilent® B1500, disposant d'une option spécifique à l'émission d'impulsions et de mesures courant-

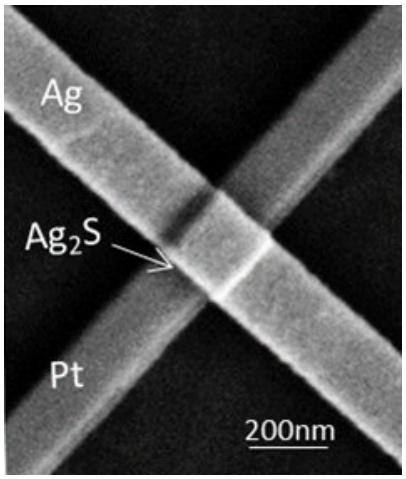


FIGURE 4.1 – Cellule électrochimique métallique de 200 nm de côté, observée au microscope électronique à balayage. *Source :* image de S. LA BARBERA.

tension rapides. L'ensemble des mesures électriques ont été effectuées sous atmosphère ambiante, avec une station sous pointes classique.

Mécanisme de commutation résistive : principe de base. Le mécanisme de base d'une transition de type SET, c'est-à-dire vers l'état de faible résistance R_{ON}, repose sur l'oxydation de l'électrode supérieure d'argent en ions Ag^+ et la réduction de ces mêmes ions Ag^+ en des filaments conducteurs d'argent, à travers la couche d'électrolyte solide. La transition de type RESET, vers l'état de résistance élevée R_{OFF}, correspond quant à elle à l'oxydation des filaments d'argent et à la réduction au niveau de l'électrode supérieure des ions Ag^+ ainsi formés.

Un tel mécanisme d'oxydoréduction réversible est associé à une caractéristique courant-tension (I-V) bipolaire, dont un exemple est présenté à la figure 4.2. Les différentes phases observées lors d'un cycle de balayage en tension y sont également illustrées.

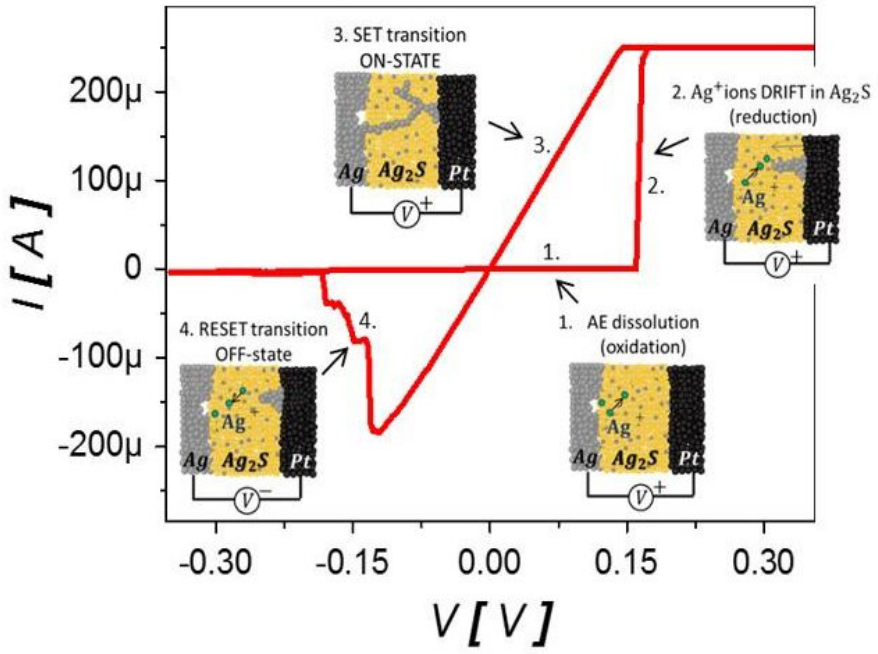


FIGURE 4.2 – Caractéristique courant-tension (I-V) observée lors d'un cycle complet de balayage en tension. Les mécanismes physiques associés aux différentes phases sont indiqués de façon schématique par les illustrations 1 à 4. *Source :* S. LA BARBERA.

4.2 Des composants aux multiples comportements plastiques

Cette partie présente les différents comportements plastiques qui coexistent au sein d'une même cellule électrochimique métallique Ag_2S, ainsi que le modèle analytique développé pour décrire l'évolution de la conductance d'une cellule mémoire soumise à des impulsions de programmation toutes identiques. À l'aide de ce modèle, un accent particulier est mis sur la compréhension des transitions entre les différents régimes de plasticité. Il s'agit en effet d'une étape importante pour nous permettre ensuite d'exploiter l'interaction entre ces multiples dynamiques pour réaliser un apprentissage à l'échelle d'un système complet.

4.2.1 Plasticité synaptique non « hebbienne »

La transition plastique détaillée dans cette partie a récemment été l'objet d'un article de la part de nos collaborateurs[133]. Nous rappelons ici une partie de ces résultats, nécessaire à la compréhension du reste du chapitre.

Remarque

Dans l'ensemble de ce chapitre, les impulsions de programmation appliquées sur les cellules électrochimiques métalliques sont toutes identiques et de forme rectangulaire, d'une amplitude de 0,4 V pour une durée de 50 µs.

4.2.1.1 D'une plasticité à court terme vers une plasticité à plus long terme

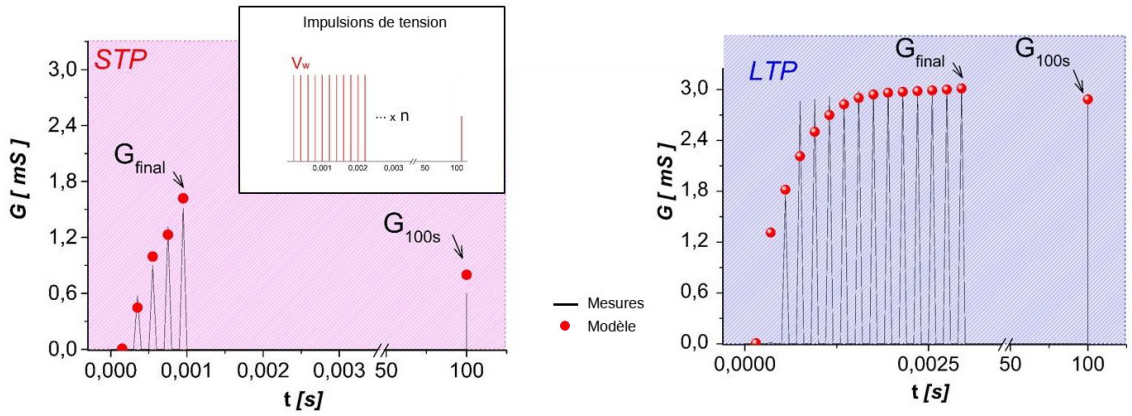


FIGURE 4.3 – Les graphes dont l'arrière-fond est coloré présentent l'évolution au cours du temps t de la conductance électrique G lorsqu'un train d'impulsions de tension tel que celui esquissé en insert est appliqué sur un échantillon. Les pointes noires sont des relevés expérimentaux, tandis que les symboles rouges (●) sont les prédictions d'un modèle ajusté, présenté dans la partie 4.2.1.2. Ces graphes mettent en évidence l'existence de différentes échelles temporelles de plasticité : à court terme (à gauche, fond rose) ou à plus long terme (à droite, fond bleu). *Source :* adapté d'une figure de S. LA BARBERA.

Appliquées de façon répétée à un échantillon, des impulsions de tension carrées simples induisent une potentialisation de ce dernier, associée à la croissance d'un filament d'argent métallique. Appliquer un nombre restreint de telles impulsions de programmation se traduit par la formation de filaments fins, qui tendent à se dissoudre rapidement en l'absence de stimulation électrique, tandis que l'application d'un nombre important de ces mêmes impulsions entraîne le développement de filaments plus larges, qui sont ainsi plus stables dans le temps lorsque la stimulation électrique est arrêtée. Ces deux approches permettent d'obtenir une plasticité respectivement à court terme (STP) et à plus long terme (LTP), comme le montre la figure 4.3. La stabilité filamentaire est interprétée comme le résultat d'une compétition entre une énergie de surface qui tend à dissoudre les filaments (plasticité à court terme) et une énergie de volume qui au contraire les stabilise (plasticité à plus long terme).

À l'issue de l'application d'un train d'impulsions de programmation, la conductance atteint une valeur G_{final}. Il est possible de suivre l'évolution de la conductance depuis cette valeur en appliquant des impulsions de lecture d'une amplitude suffisamment faible pour ne pas perturber

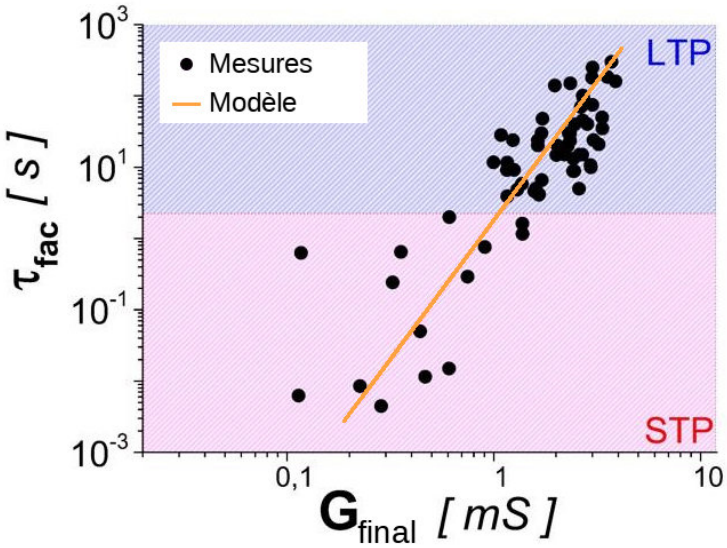


FIGURE 4.4 – Observation d'une loi de puissance entre la constante de temps τ_{fac} de la relaxation exponentielle et la conductance G_{final} atteinte à l'issue du train d'impulsions de programmation. Les symboles (●) indiquent les points expérimentaux, tandis que la ligne en trait plein orange est un ajustement du modèle (équation (4.1)). Les différents états de potentialisation (c'est-à-dire les différentes valeurs de G_{final}) ont été obtenus en modifiant le nombre d'impulsions de programmation de 3 à 150, ainsi qu'en faisant varier la fréquence de ces dernières de 1 à 5 kHz. Les aires colorées classent les valeurs de constante de temps τ_{fac} selon deux type de plasticité : à court terme (STP, en rose) et à plus long terme (LTP, en bleu). La valeur numérique de la durée employée pour distinguer les deux zones est ici arbitraire ; dans la pratique, elle dépend de la tâche à résoudre. *Source :* adapté d'une figure de S. LA BARBERA.

l'échantillon. Dans le cas présent, ce suivi a été réalisé durant 100 s après la dernière impulsion de programmation, $G_{100\,\mathrm{s}}$ désignant la conductance de l'échantillon après une telle durée. La relaxation peut être ajustée avec un modèle de décroissance exponentielle, de constante de temps τ_{fac}. La figure 4.4 montre cette constante de temps τ_{fac} présente avec la conductance G_{final} à l'issue de la programmation une relation de type loi de puissance. Autrement dit, la rétention de la cellule mémoire dépend de sa potentialisation, selon la formule

$$\tau_{\mathrm{fac}} = C_0 \times (G_{\mathrm{final}})^{C_1}. \tag{4.1}$$

Un ajustement de cette formule sur des données obtenues pour différentes fréquences et longueurs de trains d'impulsions de programmation nous permet d'obtenir $C_0 = 3{,}4 \times 10^{12}\,\mathrm{s{\cdot}S^{-C_1}}$ et $C_1 = 4$.

Remarque

Il est à noter que la frontière entre « court terme » et « long terme » est une notion qui dépend de la tâche visée. Ainsi, la durée moyenne nécessaire à l'apprentissage d'un motif doit correspondre à la dynamique à court terme des synapses électrochimiques métalliques, de façon à observer une transition vers un régime à plus long terme après la répétition de ce motif.

4.2.1.2 Vers un premier modèle d'évolution de la conductance

De façon à décrire l'évolution des cellules électrochimiques métalliques considérées lorsque ces dernières sont soumises à des impulsions de tension positive[I] rectangulaires, un modèle a été développé par nos collaborateurs de l'Institut d'Électronique de Microélectronique et de Nanotechnologie[133]. Il s'appuie sur une description analytique simple (dérivée d'un modèle phénoménologique de synapses utilisé en biologie) pour décrire le comportement observé expérimentalement.

Le modèle retenu est basé sur l'observation de deux dynamiques distinctes au sein des cellules mémoires étudiées :

1. une relaxation de leur conductance G en l'absence de stimulation électrique ;
2. à l'opposé, un incrément de conductance provoqué l'application d'une impulsion de tension positive rectangulaire de programmation.

Soit un échantillon ayant atteint après $n-1$ impulsions une conductance de valeur $G_{\mathrm{n-1}}$, comprise entre une valeur minimum G_{min} (environ 1 µS dans notre cas) et une valeur maximale A (autour de 3 mS pour les échantillons étudiés). Après une durée δt sans stimulation, sa

I. Avec les conventions de la figure 4.2 page 128.

conductance n'est plus que

$$G_{\text{relax}}(\delta t) = (G_{\text{n}-1} - G_{\text{min}}) \times \exp\left(\frac{-\delta t}{\tau_{\text{fac}}}\right) + G_{\text{min}}, \tag{4.2}$$

avec la constante de temps τ_{fac} suivant une loi de puissance calquée sur l'équation (4.2) :

$$\tau_{\text{fac}} = C_0 \times (G_{\text{n}-1})^{C_1}. \tag{4.3}$$

Par ailleurs, lors de l'application d'une nouvelle impulsion de programmation (d'indice n) à l'issue d'un délai δt, la conductance de ce même échantillon devient

$$G_{\text{n}} = \underbrace{G_{\text{relax}}(\delta t)}_{\text{Relaxation}} + \underbrace{U \times (A - G_{\text{relax}}(\delta t))}_{\text{Incrément programmation}}. \tag{4.4}$$

Le facteur U est une constante numérique comprise entre 0 et 1, qui module l'incrément de potentialisation induit par une impulsion de programmation.

Il a été constaté après ajustement de ce modèle aux relevés expérimentaux obtenus lors de l'application de trains d'impulsions de programmation à fréquence fixe, que les paramètres A et U (relatifs à l'effet potentialisant d'une impulsion) demeurent constants jusqu'à des fréquences d'au moins 5 kHz (les instants *initiaux* de deux impulsions successives sont alors séparés de 200 µs). Dans cette gamme de fréquence, ce premier modèle analytique simple donne de bonnes prédictions quantitatives du comportement plastique des cellules électrochimiques métalliques, en particulier lors de leur transition d'une plasticité à court terme vers une plasticité à plus long terme, comme le montrent les ajustements de la figure 4.3 page 129.

Le modèle constitué des équations (4.2) à (4.4) capture ainsi les principales propriétés des cellules électrochimiques métalliques Ag_2S intéressantes pour un usage synaptique. Tout d'abord, en l'absence de stimulation électrique, la relaxation naturelle des filaments entraîne la dépression de toutes les synapses faiblement potentialisées, jusqu'à ce qu'elles atteignent leur valeur minimale de conductance. À l'issue de ce processus, seules les synapses fortement potentialisées avant l'arrêt des stimulations électriques forment encore des connexions synaptiques fortes, du fait de constantes de temps τ_{fac} de valeur élevée. Un comportement similaire a été observé au sein de systèmes biologiques, où les synapses faibles tendent à disparaître au cours du temps[261]. Ceci n'est d'ailleurs pas sans rappeler également les stratégies de pénalisation des connexions de poids faible en apprentissage automatique[262]. Par ailleurs, d'après l'équation (4.4), l'incrément de potentialisation devient nul lorsque la conductance G atteint la valeur maximale A, ce qui conduit à l'obtention de poids synaptiques bornés, évitant ainsi toute divergence de ces derniers.

261. G. CHECHIK, I. MEILIJSON et E. RUPPIN, Neural Computation, 1999.
262. R. SETIONO, Neural Computation, 1997.

4.2.2 Plasticité synaptique « hebbienne »

Les comportements plastiques décrits dans la partie 4.2.1 demeurent d'un intérêt limité en termes de capacités de traitement de l'information, notamment en l'absence de comportements additionnels, tels que des effets hebbiens. Ces derniers permettent en effet des apprentissages basés sur les coïncidences entre signaux d'entrée et signaux de sortie, comme par exemple les comportements de type *Spike-Timing Dependent Plasticity* présentés dans les chapitres précédents.

Dans les systèmes biologiques, les événements présynaptiques et postynaptiques sont définis de manière univoque en raison du caractère intrinsèquement unidirectionnel des mécanismes biochimiques ayant lieu au sein des connexions synaptiques. Les cellules électrochimiques métalliques étudiées constituent quant à elles des connexions bidirectionnelles, ce qui nécessite de pouvoir différencier autrement les interactions entre événements présynaptiques et événements postsynaptiques. Une stratégie envisageable à l'échelle du système repose sur l'hypothèse d'être en mesure de contraindre la distance temporelle minimale δt_{hebb} entre deux événements synaptiques de même nature. Ceci peut par exemple être réalisé en imposant une période réfractaire de durée *ad hoc* à l'échelle des neurones. Seule une interaction entre un neurone présynaptique et un neurone postsynaptique est alors en mesure de produire une paire d'impulsions séparées par une durée δt inférieure à δt_{hebb}.

4.2.2.1 Résultats expérimentaux

Dans cette partie, nous présentons des résultats expérimentaux qui démontrent que les cellules électrochimiques métalliques Ag_2S disposent, outre les comportements plastiques décrits précédemment, d'une forme de plasticité supplémentaire qu'il est possible d'utiliser pour mettre en œuvre un apprentissage de type hebbien. En adoptant la stratégie présentée ci-dessus, ces nanocomposants permettent d'obtenir de façon *intrinsèque* un comportement plastique de type *Spike-Timing Dependent Plasticity*, basé sur la distance temporelle entre impulsions présynaptique et postsynaptique. La règle de type *Spike-Timing Dependent Plasticity* obtenue est symétrique, ne distinguant pas les paires d'impulsions causales de celles d'événements anti-causaux. Ce type de comportement plastique a également été observé en biologie, au sein du système GABAergique [I] des neurones de l'hippocampe[263].

Les échantillons sont soumis à des impulsions de tension toutes identiques (0,4 V durant 50 µs) [II]. Le protocole expérimental est analogue à celui présenté à la partie 4.2.1.2, si ce n'est que les stimulations sont désormais constituées d'un train à fréquence fixe de *paires* d'impulsions séparées par une durée δt. Différentes fréquences de trains de paires, ainsi que plusieurs distances temporelles δt entre les impulsions d'une paire sont étudiées.

I. Lié à l'acide gamma-aminobutyrique (GABA), qui est un neurotransmetteur.

263. L. F. ABBOTT et S. B. NELSON, Nature Neuroscience, 2000.

II. Pour un échantillon donné, les impulsions sont toutes appliquées sur la même électrode. Ceci est équivalent à un cas réaliste dans lequel certaines de ces impulsions sont appliquées sur l'autre électrode, avec une polarité opposée.

La conductance atteinte à l'issue d'un train de paires d'impulsions est notée G_{final} et illustre la potentialisation synaptique, tandis que $G_{100\text{ s}}$ désigne la conductance mesurée 100 s après l'arrêt des stimulations. Le rapport $G_{100\text{ s}}/G_{\text{final}}$ constitue une mesure de la rétention synaptique à long terme. La figure 4.5 montre une augmentation de ces deux quantités lorsque la distance temporelle δt entre les impulsions d'une même paire est inférieure à la centaine de microsecondes. Cet effet est d'autant plus marqué que la distance temporelle δt décroît jusqu'à 50 µs, constituant ainsi une forme de *Spike-Timing Dependent Plasticity*. Des expériences témoins, de stimulation par un train d'impulsions non appairées, sont représentées à l'aide de disques verts sur ces mêmes figures. Elles montrent quant à elles une potentialisation modérée (G_{final} de l'ordre du millisiemens seulement) et une rétention à long terme très faible ($G_{100\text{ s}}/G_{\text{final}}$ proche de zéro). Les distances temporelles δt inférieures à 50 µs correspondent à des situations de superposition des impulsions de tension : le doublement de tension qui en résulte tend à potentialiser fortement et à long terme les cellules mémoires électrochimiques, comme le montrent les symboles carrés sur les figures 4.5 (a) et (b).

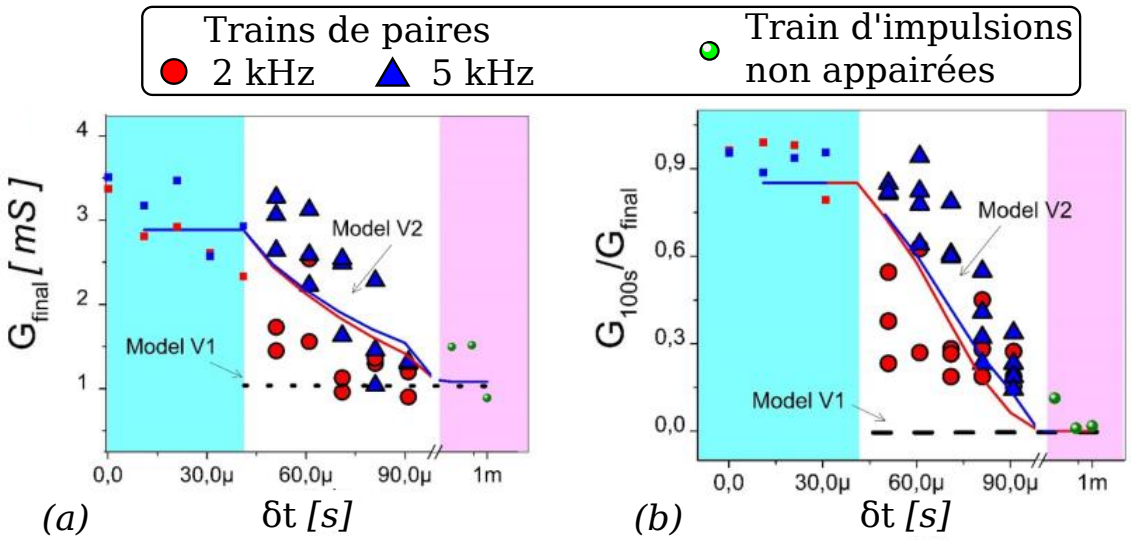


FIGURE 4.5 – Évolutions *(a)* de la potentialisation G_{final} à l'issue d'un train de paires d'impulsions et *(b)* de la rétention après 100 s en fonction de la distance temporelle δt entre les deux impulsions d'une même paire, mettant en évidence un comportement plastique additionnel aux temps courts. Les symboles sont les résultats des mesures, tandis que les lignes en trait plein correspondent aux ajustements du modèle présenté ci-après (V2). Deux fréquences de trains d'impulsions sont présentées : 2 kHz (●) et 5 kHz (▲). Des points de mesures analogues aux expériences précédentes sont également ajoutés (●). Les couleurs d'arrière-fond correspondent aux différents régimes : superposition des impulsions (cyan), similaire aux précédentes expériences aux temps longs (rose) et nouveau aux temps courts (blanc). *Source :* adapté d'une figure de S. LA BARBERA.

4.2.2.2 Extension du modèle précédent

Le modèle constitué des équations (4.2) à (4.4), qualifié de V1 sur les figures, ne permet pas de reproduire le comportement observé à la figure 4.5 pour les distances temporelles δt inférieures à 100 µs. Nous allons voir comment modifier ces équations pour obtenir une nouvelle version du modèle (V2 sur les figures), capable de mieux décrire le nouveau comportement plastique découvert pour les délais courts.

Adoptons l'idée de ne plus considérer nécessairement indépendants de la distance temporelle δt les paramètres A (la conductance d'une cellule totalement potentialisée) et U (qui module l'incrément de potentialisation apporté par une impulsion). Au contraire, pour chacune des distances temporelles δt étudiées expérimentalement, ajustons numériquement ces paramètres A et U avec les points de mesures. Les figures 4.6 (a) et (b) montrent le résultat de ces ajustements sur les données présentées à la figure 4.5 (trains de paires d'impulsions à une fréquence de 2 ou 5 kHz). Pour des distances temporelles δt comprises entre 50 et 100 µs, l'accroissement de la potentialisation et de la rétention observé lorsque δt diminue peut être modélisé par un accroissement analogue des paramètres A et U. Par ailleurs, dans le cas de valeurs de δt inférieures à 50 µs, il est considéré que les paramètres A et U sont de valeur constante en raison de la superposition des impulsions de tension, qui tend à faire saturer les cellules électrochimiques métalliques.

La conductance d'une cellule totalement potentialisée A (exprimée ci-après en siemens), est désormais décrite par une fonction continue affine par morceaux

$$A(\delta t) = \begin{cases} 3{,}4 \times 10^{-3} & \text{si } \delta t < 50\ \mu\text{s}, \\ \sigma_A \times \delta t + b_A & \text{si } 50\ \mu\text{s} \leq \delta t < 100\ \mu\text{s}, \\ 2{,}7 \times 10^{-3} & \text{sinon}, \end{cases} \tag{4.5}$$

avec $\sigma_A = -18\,\text{S}\cdot\text{s}^{-1}$ et $b_A = 4{,}32 \times 10^{-3}$ S. Par ailleurs, la modulation de l'incrément de potentialisation d'une impulsion U (sans unité) est désormais modélisée par un plateau suivi d'une décroissance exponentielle

$$U(\delta t) = \begin{cases} 85 \times 10^{-3} & \text{si } \delta t < 50\ \mu\text{s}, \\ \sigma_U \times \exp\left(\frac{-\delta t}{\tau_U}\right) + U_\infty & \text{sinon}, \end{cases} \tag{4.6}$$

avec $\sigma_U = 271{,}7 \times 10^{-3}$, $\tau_U = 34{,}1\,\mu\text{s}$ et $U_\infty = 2{,}67 \times 10^{-3}$. Ces modélisations sont tracées en trait plein et en pointillé sur les figures 4.6 (a) et (b).

Le nouveau modèle est ainsi composé des précédentes équations (4.2) à (4.4), dans lesquelles sont réinjectées les équations (4.5) et (4.6). Les traits pleins de la figure 4.5 montrent que ce modèle améliore sensiblement la description qualitative de la potentialisation et de la rétention à long terme induites par des paires d'impulsions de faible distance temporelle δt, en comparaison de la première version du modèle. Il ne permet toutefois pas de vraiment discriminer

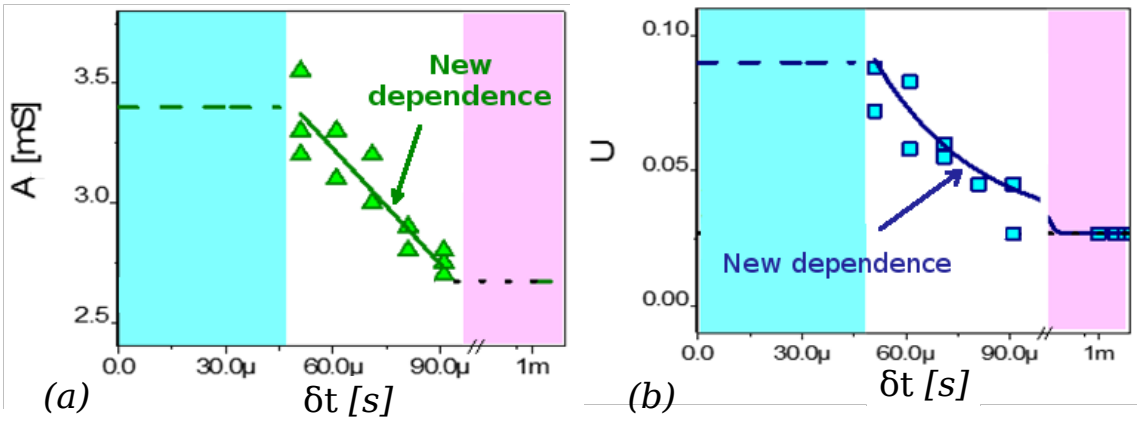


FIGURE 4.6 – Évolution des paramètres A (a) et U (b) du modèle, ajustés sur les expériences de stimulation par des trains de paires d'impulsions, en fonction de la distance temporelle δt entre deux impulsions d'une même paire. Les symboles sont les résultats des mesures, tandis que les lignes en trait plein sont le modèle moyen adopté. Les couleurs d'arrière-fond correspondent aux différents régimes : superposition des impulsions (cyan), similaires aux précédentes expériences aux temps longs (rose) et de type *Spike-Timing Dependent Plasticity* aux temps courts (blanc). *Source :* adapté d'une figure de S. LA BARBERA.

le cas de paires à 2 kHz (traits pleins rouges) de celui de paires à 5 kHz (traits pleins bleus), là où les relevés expérimentaux indiquent une différence quantitative. Ceci suggère l'existence dans les expériences précédentes d'une dépendance des paramètres synaptiques A et U à la fréquence des trains d'impulsions non appairées, qui n'aurait pas été observée en raison de la forte variabilité des caractéristiques de commutation entre les différents échantillons, tandis que les conséquences de cette dépendance se trouveraient exacerbées aux faibles distances temporelles entre impulsions. Néanmoins, en attendant de mieux comprendre les sources de la forte variabilité entre échantillons, ce qui permettrait d'affiner le modèle que nous venons de présenter, par exemple en adoptant une approche partant d'une modélisation physique de la commutation, nous utiliserons la description qualitative du modèle développé dans cette partie pour illustrer dans la suite du chapitre l'intérêt des cellules électrochimiques métalliques comme synapses artificielles.

4.2.2.3 Hypothèse sur l'origine de ce comportement additionnel

Plusieurs hypothèses sont envisageables pour expliquer l'existence de ce comportement plastique de type *Spike-Timing Dependent Plasticity* au sein des cellules électrochimiques métalliques Ag_2S.

Différents régimes de relaxation ? La première possibilité est une relaxation non linéaire, telle que récemment proposée par DU et coll.[131] pour des cellules mémoires à base d'oxydes de métaux de transition, selon laquelle la dynamique de relaxation de la conductance est modulée

par une seconde variable d'état. Les auteurs proposent ainsi un modèle mémristif du second ordre dans lequel les deux variables d'état retenues sont dotées de constantes de temps différentes pour associer les effets à court et à long terme aux multiples dynamiques ioniques internes qui peuvent exister au sein de cette famille de nanocomposants mémristifs. Cette modélisation a notamment permis d'expliquer les effets de plasticité à court terme et de type *Spike-Timing Dependent Plasticity* sans superposition des impulsions de tension observés expérimentalement au sein de cellules mémoires utilisant une couche intermédiaire faite d'oxyde de tungstène.

Dans le cas des cellules électrochimiques métalliques Ag_2S, des mesures de la relaxation de conductance au cours du temps ont été réalisées, de 500 ns à 100 s après la fin de la stimulation. À l'issue de ces mesures, la présence de plusieurs régimes de relaxation n'a pas été observée, ce qui suggère que l'origine des multiples comportements plastiques présentés auparavant est à trouver ailleurs.

Effet de la température ? Une seconde origine possible des effets plastiques observés dans le cas d'impulsions rapprochées est la dissipation thermique liée à un échauffement de la cellule mémoire programmée. Une seconde impulsion rapprochée pourrait bénéficier d'un échauffement local de la région du filament concerné, ce qui permettrait d'accroître l'effet potentialisant de cette deuxième impulsion.

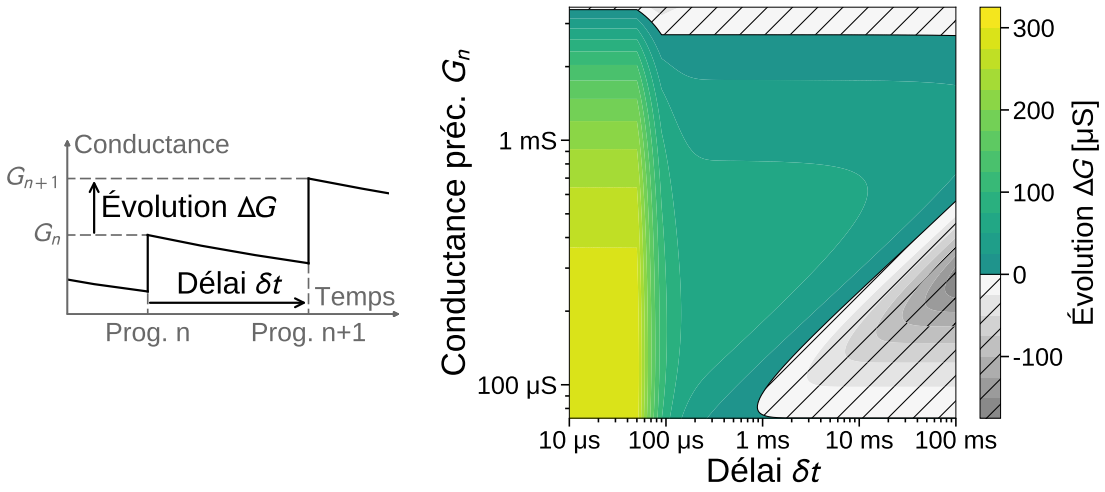
Afin d'évaluer les effets de la température sur le comportement plastique des cellules électrochimiques métalliques Ag_2S, de nouvelles mesures avec des trains de paires d'impulsions ont été réalisées sur des échantillons portés à une température de 420 K. Ces expériences ont permis d'observer une nette augmentation de la potentialisation maximale et de la rétention à long terme des cellules mémoires par rapport aux expériences précédentes effectuées à température ambiante. Dans le cas d'une situation à 300 K, l'accroissement des paramètres A et U lorsque l'intervalle temporel δt entre impulsions d'une même paire diminue pourrait ainsi être la signature d'une température locale plus élevée lorsque la seconde impulsion est appliquée. Si ces expériences ne sont pas suffisantes pour attribuer de manière exclusive une nature thermique aux effets d'interaction entre impulsions rapprochées, elles suggèrent néanmoins que la température y joue un rôle.

4.2.3 Cartographie de l'évolution de conductance

Cette nouvelle version du modèle d'évolution de la conductance des cellules électrochimiques métalliques Ag_2S capture la plasticité à court terme, la transition vers une plasticité à plus long terme, ainsi que les effets de type *Spike-Timing Dependent Plasticity* dans le cas d'impulsions rapprochées. Les effets d'échauffement local supposés donner lieu à ce troisième comportement plastique sont implicitement décrits par l'accroissement des paramètres A et U lorsque la distance temporelle δt entre deux impulsions successives diminue.

Afin de mettre en valeur la façon dont ces différents comportements plastiques et leur interaction peuvent être utiles pour un usage synaptique au sein d'un système neuromorphique

impulsionnel, nous nous intéressons tout d'abord à l'évolution de conductance induite par l'application de deux impulsions de programmation successives. Une telle situation, esquissée à la figure 4.7a, conduit à une évolution de conductance ΔG, pouvant être positive ou négative, qui dépend du délai δt entre les deux impulsions et de la conductance G_{n} atteinte à l'issue de la première impulsion. La figure 4.7b est un paysage de cette évolution de conductance ΔG en fonction des deux paramètres G_{n} et δt. L'application d'impulsions à fréquence fixe correspond à un déplacement vertical dans ce paysage, tandis qu'un déplacement horizontal y est la signature d'une modification de la distance temporelle entre impulsions successives.



(a) Situation étudiée et notations.

(b) Paysage d'évolution de la conductance.

FIGURE 4.7 – *(a)* Schéma des variables utilisées pour décrire l'effet d'impulsions consécutives (et identiques) sur la conductance d'une cellule électrochimique métallique Ag_2S. *(b)* Paysage d'évolution de la conductance en fonction du délai δt écoulé depuis la précédente impulsion et de la conductance G_{n} atteinte à l'issue de cette dernière. Les couleurs indiquent la valeur de l'évolution de conductance ΔG : les aires hachurées en niveaux de gris correspondent aux valeurs négatives de ΔG, tandis que celles colorées de jaune à vert correspondent aux valeurs positives de ΔG. La ligne noire en trait plein qui sépare la zone hachurée de la zone sans hachures est le contour nul.

Une telle représentation facilite la compréhension des phénomènes d'apprentissage qu'il est possible d'obtenir avec des synapses artificielles constituées de cellules électrochimiques métalliques Ag_2S. Nous pouvons distinguer trois régions principales :

1. Une région jaune à vert clair, correspondant aux délais $\delta t \lesssim 100\,\mu s$. Cette zone est celle des plus fortes valeurs d'évolution de conductance ΔG. Elle correspond aux délais qui excitent la dynamique plastique de type *Spike-Timing Dependent Plasticity*, ce qui donne lieu aux incréments de conductance les plus élevés. L'évolution de conductance ΔG étant positive en tout point de cette zone, toute impulsion de programmation s'y trouvant potentialise la cellule mémoire par rapport à l'impulsion précédente.

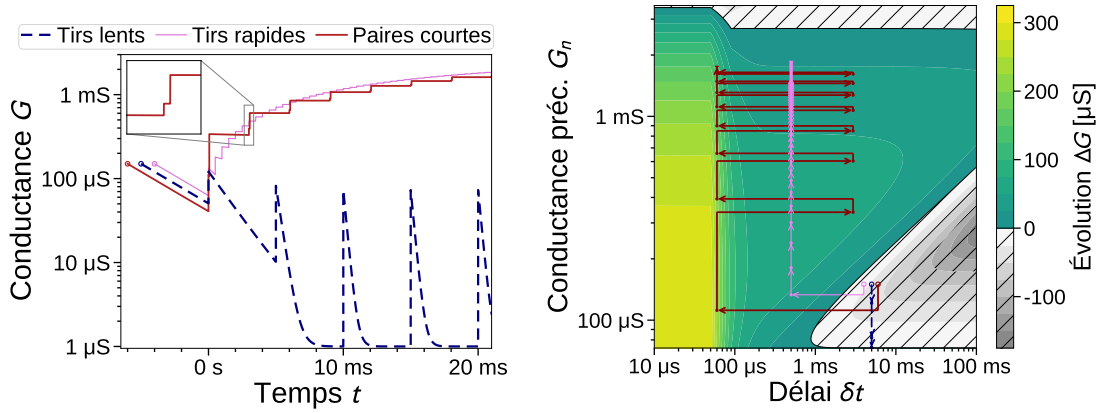
2. Une région vert foncé pour des délais $\delta t \gtrsim 100\,\mu s$. L'évolution de conductance ΔG y est également partout positive et toute impulsion y échouant participe aussi à la potentialisation de la cellule mémoire par rapport à l'impulsion précédente. Cependant, les valeurs de ΔG y sont d'amplitude plus faible que celles de la région 1 : davantage d'impulsions de programmation sont requises pour atteindre une même valeur de conductance. Ceci s'explique par le fait que le délai δt entre impulsions successives devient suffisamment long pour permettre aux paramètres A et U de décroître jusqu'à s'approcher de leur valeur asymptotique. Les impulsions se trouvant dans cette zone ne bénéficient donc plus de l'effet de type *Spike-Timing Dependent Plasticity* et cette région présente un profil relativement plat.

3. Une région en niveaux de gris hachurée, pour des délais $\delta t \gtrsim 1\,ms$. La cellule électrochimique métallique Ag_2S programmée par une impulsion se trouvant dans cette région subit une évolution ΔG de sa conductance négative, en raison d'un incrément de conductance insuffisant pour compenser la relaxation naturelle importante. Une subite diminution de la fréquence d'excitation après une faible potentialisation peut par exemple mener la cellule mémoire dans cette zone, ayant alors pour conséquence la relaxation de son état de conductance. Si une cellule électrochimique métallique Ag_2S très faiblement potentialisée est excitée par une série d'impulsions à faible fréquence (c'est-à-dire dont tous les délais $\delta t > 1\,ms$), elle demeurera dans un régime de plasticité à court terme proche de sa conductance minimale, oscillant près du contour nul ($\Delta G = 0$) inférieur. Le contour nul supérieur délimitant cette troisième région correspond à des états instables pouvant potentialiser ou au contraire déprimer la cellule mémoire selon que la distance temporelle δt entre impulsions est légèrement modifiée. Par ailleurs, la présence d'une zone hachurée au sommet de la figure illustre simplement la saturation de la conductance de ces nanocomposants.

La figure 4.7b permet de conclure que pour potentialiser à long terme une cellule mémoire électrochimique Ag_2S depuis son état de conductance minimale, des événements de programmation espacés de moins d'une milliseconde sont nécessaires. Dans le cadre d'un usage synaptique, si une telle situation est envisageable en jouant par exemple simplement sur la fréquence des impulsions présynaptiques, nous verrons dans la partie suivante comment utiliser ce comportement à notre avantage pour distinguer, à l'échelle du système, les paires d'impulsions pré- et postsynaptiques (sans notion de causalité) des paires d'événements de programmation de même nature.

La figure 4.8 illustre l'effet de quelques stratégies types de programmation : le panneau (a) montre l'évolution temporelle de la conductance dans chacune des situations, tandis que le panneau (b) indique à l'aide de flèches les positions associées dans le paysage d'évolution de la conductance. Les trois cas considérés débutent tous d'un même état de potentialisation modérée ($G_n = 150\,\mu S$).

Le cas en pointillés bleu foncé est celui d'un train d'impulsions régulières de faible fréquence



(a) Évolution temporelle de la conductance associée à différents schémas types de programmation.

(b) Trajectoires associées dans le paysage d'évolution de la conductance.

FIGURE 4.8 – Effet de situations types de programmation sur la conductance d'une cellule électrochimique métallique Ag_2S. *(a)* Évolution de la conductance au cours du temps, selon trois schémas de programmation différents : un train d'impulsions non appairées à fréquence fixe lente (pointillé bleu foncé) ou élevée (trait plein fin rose) et un train de paires d'impulsions de courte distance temporelle entre les deux impulsions d'une même paire (trait plein rouge foncé). L'insert en haut à gauche est un zoom sur l'un des doubles incréments de conductance qui ont lieu lors d'une paire de ce troisième schéma de programmation. *(b)* Paysage d'évolution ΔG de la conductance, similaire à celui de la figure 4.7b. Les flèches schématisent la transition de conductance associée à chaque nouvelle impulsion des traces illustrées en *(a)*. Les flèches rouge foncé en trait plein correspondent au train de courtes paires d'impulsions, qui potentialise avec succès la cellule mémoire à long terme. Les fines flèches roses sont associées au train à fréquence élevée d'impulsions seules, également en mesure de potentialiser la cellule mémoire à long terme. Enfin, les flèches en pointillé bleu foncé sont pour le schéma à fréquence faible du train d'impulsions seules, qui maintient la cellule dans son régime de plasticité à court terme. Les symboles circulaires évidés (∘) indiquent les conditions initiales.

(période de 5 ms). L'effet de la première impulsion (à $t = 0$ s) est une évolution négative de la conductance par rapport à la valeur à l'issue de l'impulsion précédente (symbole ○). Les impulsions suivantes maintiennent la cellule dans un état de faible conductance, sans évolution de cette dernière d'une impulsion à la suivante, le délai δt avant chaque impulsion étant suffisamment long pour assurer la relaxation asymptotique de la cellule avant chaque nouvelle programmation.

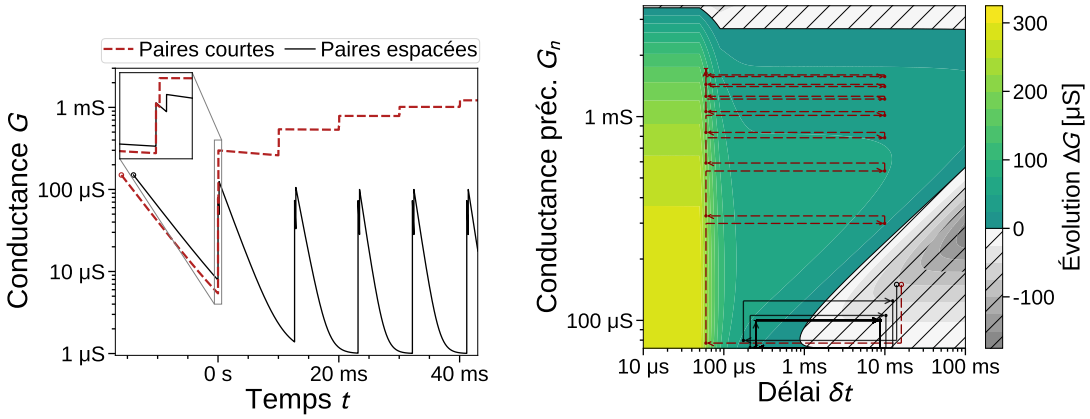
Le cas en traits pleins rouge foncé comporte des paires d'événements rapprochés capables d'exciter la dynamique de type *Spike-Timing Dependent Plasticity* et ainsi de mener à un état de forte potentialisation à long terme. Après une évolution négative de la conductance à l'issue de la première impulsion (à $t = 0$ s), la cellule électrochimique métallique Ag_2S reçoit une seconde impulsion après seulement 60 µs, qui induit alors une forte évolution positive de la conductance. Une répétition de quelques paires d'impulsions similaires est en mesure de potentialiser à long terme la cellule mémoire, même si la fréquence des paires demeure faible (période de 3 ms). Une telle trajectoire est un exemple de la transition d'une plasticité à court terme vers une plasticité à plus long terme déclenchée par la dynamique de type *Spike-Timing Dependent Plasticity*.

Enfin, la situation en traits fins roses illustre le cas d'une potentialisation à long terme induite par un train régulier d'impulsions de fréquence élevée (période de 0,5 ms) mais qui n'exploite pas la dynamique de type *Spike-Timing Dependent Plasticity*. Toujours après une évolution négative de la conductance à l'issue de la première impulsion (à $t = 0$ s), la cellule mémoire connaît un accroissement net de sa conductance après chaque impulsion ultérieure. Cependant, en raison de l'évolution de conductance ΔG plus faible induite par ces impulsions isolées, l'obtention d'une forte potentialisation requiert un nombre d'événements de programmation sensiblement plus élevé que celui d'une stratégie qui exploite la dynamique de type *Spike-Timing Dependent Plasticity*.

Remarque

Nous considérons uniquement des impulsions causant un incrément de conductance, c'est-à-dire de polarité positive par rapport à l'électrode supérieure, et exploitons la relaxation naturelle des composants pour assurer la dépression synaptique. Des impulsions de polarité opposée pourraient en principe induire une dépression analogue, autorisant même possiblement une plus grande flexibilité d'utilisation. Néanmoins, dans le cas particulier des cellules électrochimiques métalliques Ag_2S étudiées, il a été observé qu'une unique impulsion entraîne une dépression très importante, ce qui constitue un frein à tout apprentissage (déterministe) graduel. Se reposer sur la relaxation naturelle est par conséquent préférable pour assurer la dépression dans le cadre d'un usage synaptique analogique de ces nanocomposants.

Comme la figure 4.9 le montre, même dans le cas d'une stratégie de programmation utili-



(a) Évolution de la conductance associée à différents trains de paires d'impulsions.

(b) Trajectoires associées dans le paysage d'évolution de la conductance.

FIGURE 4.9 – Programmation par un train de paires d'impulsions : effet de la distance temporelle δt entre les impulsions d'une même paire. *(a)* Évolution temporelle de la conductance au cours du temps, selon deux schémas de programmation analogues, et *(b)* transitions associées dans un paysage d'évolution de la conductance similaire à celui de la figure 4.7b. Les styles de lignes sont cohérents entre les deux sous-figures. En pointillé rouge foncé est illustré le cas d'un train de courtes paires dont les impulsions sont suffisamment rapprochées pour exciter la plasticité de type *Spike-Timing Dependent Plasticity*, ce qui entraîne la transition de la cellule mémoire d'un régime de plasticité à court terme vers celui d'une potentialisation à long terme. Les traits pleins noirs correspondent au cas d'un train de paires de fréquence moyenne analogue à la situation précédente mais dont les impulsions d'une même paire sont trop espacées pour exciter la dynamique précédente : d'un état initial similaire à celui de la première situation, la cellule mémoire tombe après quelques paires dans un cycle limite sans jamais se potentialiser à long terme. *NB* : dans la seconde situation, les distances temporelles entre impulsions consécutives sont légèrement modulées autour de deux valeurs, afin de pouvoir mieux observer l'apparition du cycle limite. Les symboles circulaires évidés (○) indiquent les conditions initiales.

sant des paires d'impulsions, la potentialisation de la cellule électrochimique métallique reste dépendante de la distance temporelle entre les deux impulsions d'une même paire : certains scénarios n'aboutissent pas à une potentialisation à long terme de la cellule mémoire. Le cas en pointillé rouge foncé est similaire à celui en trait plein rouge foncé de la figure 4.8 : un train régulier de courtes paires d'impulsions, capable de potentialiser à long terme une synapse. Les traits pleins noirs illustrent quant à eux la situation d'un train de paires d'impulsions d'une fréquence analogue mais dont la distance temporelle entre les impulsions d'une même paire (autour de 0,25 ms) est désormais trop longue pour exciter la dynamique de type *Spike-Timing Dependent Plasticity*. La synapse demeure alors dans son régime de plasticité à court terme, adoptant même dans le cas présent un cycle limite d'oscillations entre deux niveaux d'évolution de conductance ΔG, sans jamais démarrer une trajectoire de saturation de sa conductance. Ce type de situation est une vision simplifiée des comportements que nous souhaitons obtenir dans le cas d'une synapse excitée par des impulsions postsynaptiques qui ne sont pas fortement corrélées à des impulsions présynaptiques.

Remarque

Durant une simulation réaliste, pouvant en outre inclure du bruit ou de la variabilité entre composants, la conductance des synapses évolue de manière sensiblement plus complexe que celle des quelques cas schématiques présentés dans cette partie. Nous verrons notamment par la suite que les synapses potentialisées durant un apprentissage sont sujettes à une combinaison de ces différents schémas primitifs.

4.3 Preuve de concept d'un apprentissage

Dans le but d'évaluer le potentiel synaptique des cellules électrochimiques métalliques Ag_2S, qui disposent à la fois d'une plasticité pouvant évoluer de court à plus long terme et d'un comportement de type *Spike-Timing Dependent Plasticity*, nous avons recours à des simulations à l'échelle d'un système utilisant ces nanocomposants comme synapses artificielles. La tâche cognitive consiste en la détection d'objets se déplaçant (vers le bas) à vitesse constante dans une voie aléatoire parmi trois voies verticales sans chevauchement. Ce problème rappelle celui de détection de véhicules sur une autoroute présenté dans le chapitre 3, dans une version simplifiée et davantage ajustable afin de faciliter dans un premier temps l'étude d'un apprentissage à l'échelle d'un petit système.

4.3.1 Méthodologie de simulation à l'échelle d'un système

4.3.1.1 Architecture générale

De façon analogue aux simulations du chapitre 3, les circuits associés aux neurones sont décrits de manière fonctionnelle, tandis que les synapses, constituées de cellules électrochimi-

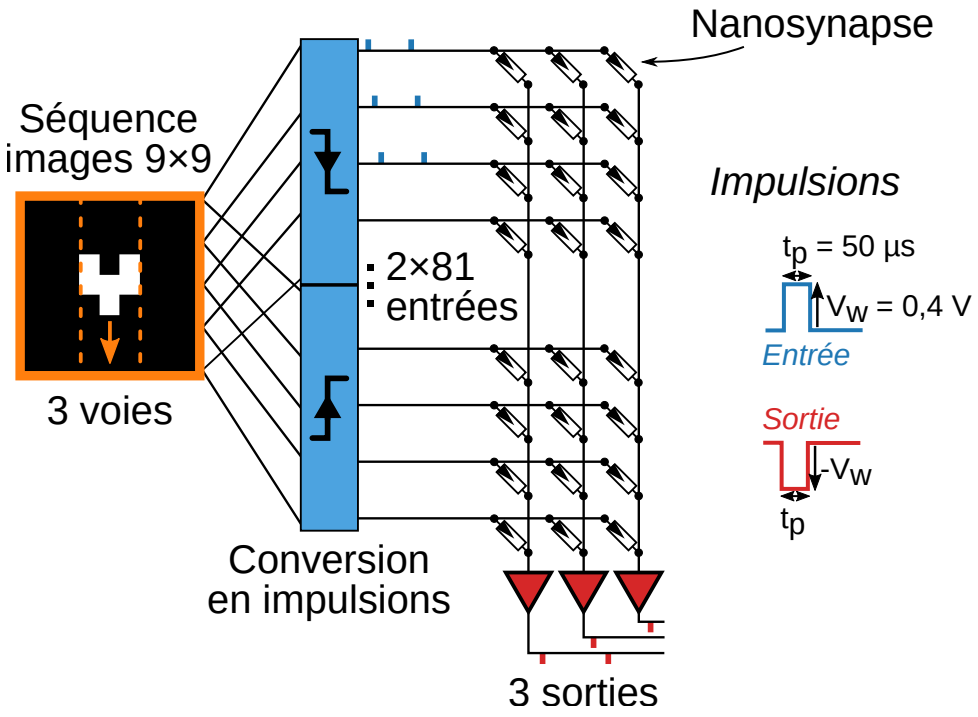


FIGURE 4.10 – Schéma de principe des simulations à l'échelle d'un système. La vignette en noir et blanc est un exemple d'image tirée de l'une des vidéos artificielles utilisées comme entrée brute. L'intensité de chaque pixel est codée dans son niveau de gris. Les symboles au cœur des deux blocs bleus indiquent le type de variation d'intensité qui est utilisé pour générer des impulsions présynaptiques à partir d'une vidéo présentée en entrée. Le symbole du haut et le symbole du bas correspondent aux entrées présynaptiques sensibles respectivement à une diminution et à une augmentation d'intensité des pixels. Les trois symboles triangulaires rouges désignent les trois neurones de sortie. À droite sont esquissées les formes d'onde des impulsions utilisées par le simulateur (similaires à celles des mesures expérimentales). La masse constitue la référence de chacune de ces formes d'onde.

ques métallique Ag_2S, sont modélisées par les équations (4.2) à (4.6).

Les neurones d'entrée du système présentent des impulsions asynchrones générées à partir d'une vidéo selon une approche inspirée par une rétine biologique, reposant sur les variations d'intensité de chaque pixel. Chaque pixel de la vidéo est associé à *deux* neurones d'entrée, l'un sensible aux variations d'intensité positives, le second aux variations négatives. Ces neurones émettent des impulsions lorsqu'ils sont sujets à une évolution d'intensité qui excède un seuil de détection arbitraire.

Les neurones d'entrée sont connectés à trois neurones de sortie, chaque couple possible étant relié par une unique cellule électrochimique métallique Ag_2S, formant ainsi une matrice de $2\times81\times3$ synapses. Les effets résistifs et capacitifs des lignes de cette matrice ne sont pas pris en compte lors des simulations. Par ailleurs, les courants parasites de fuites (*sneak path currents* en anglais) ne sont pas à priori critiques dans la situation considérée puisque les neurones au repos imposent une masse virtuelle et que les signaux sont transmis en parallèle. Une situation comparable sera étudiée au chapitre 5.

Lorsqu'un neurone de sortie est déclenché, il émet une impulsion de sortie et applique un retour identique sur le port postsynaptique des cellules électrochimiques métalliques auxquelles il est connecté. Cette fonctionnalité peut par exemple être mise en œuvre avec un convoyeur de courant[144]. Par ailleurs, l'émission d'une impulsion par un neurone de sortie remet également à zéro les deux autres neurones de sortie.

Cette architecture est schématisée à la figure 4.10 et des précisions sur les différents sous-systèmes impliqués sont données dans la partie suivante.

4.3.1.2 Précisions techniques

Entrées vidéo artificielles. Afin de mieux contrôler la dynamique temporelle des entrées appliquées au système étudié, les impulsions présynaptiques sont obtenues à partir d'une vidéo artificielle, composée d'une succession à fréquence fixe d'images de 9×9 pixels. Ces images représentent 3 voies de circulation verticales, sans chevauchement et de largeur identique (3 pixels). Des motifs de 6 pixels, de forme, vitesse et direction identiques, se déplacent de haut en bas au centre de ces voies. Il n'y a jamais plus d'un motif à la fois sur une image.

Le fichier vidéo d'entrée dure 7,2 s, présentant 90 cycles de 80 ms. Durant chaque cycle, 3 motifs sont présentés successivement : un sur chaque voie, dans un ordre aléatoire. Ceci permet de forcer un minimum d'activité sur chacune des synapses du système, notamment lors du début de l'apprentissage. En l'absence de bruit, le délai entre deux impulsions successives sur une même synapse est compris entre 1/3 et 5/3 de cycle, c'est-à-dire entre environ 27 et 133 ms.

La vidéo est cadencée à 450 images par seconde. Un motif traverse l'extension spatiale définie par les images en 12 images successives, activant pour chacune d'entre elles les pixels qui le composent. En l'absence de bruit, les pixels activés par un motif sont tous de même valeur

144. G. LECERF et coll., IEEE International Symposium on Circuits and Systems, 2014.

unitaire (dans une unité arbitraire d'intensité), tandis que tous les autres pixels sont de valeur nulle.

Du bruit peut être ajouté sur cette vidéo artificielle. Chaque pixel possède ainsi une probabilité $Pr(\text{ON} \rightarrow \text{OFF}|\text{ON})$ de s'éteindre s'il est actif et une probabilité $Pr(\text{OFF} \rightarrow \text{ON}|\text{OFF})$ de s'activer s'il est éteint. Ce bruit blanc est indépendant et identiquement distribué sur l'ensemble des pixels de chacune des images.

D'une séquence d'images à des impulsions asynchrones. La vidéo artificielle présentée est ensuite convertie en impulsions présynaptiques asynchrones selon l'évolution de l'intensité de chaque pixel entre deux images successives. Cette façon de faire est une version significativement simplifiée du comportement de la rétine neuromorphique *Dynamic Vision Sensor*[62] présentée dans les chapitres 1 et 3. Chacun des 81 pixels d'une même image est ainsi associé à *deux* neurones présynaptiques. Au sein de chacune de ces 81 paires, un neurone émet une impulsion lorsqu'il détecte un accroissement d'intensité au-delà d'un seuil T^+, tandis que le second procède de même pour une diminution d'intensité en deçà d'un seuil T^-. Dans les simulations présentées ci-après, les seuils employés sont tous de même valeur absolue ($T^+ = -T^- = 0{,}5$).

En nous basant sur les observations de la partie précédente, une période réfractaire d'une milliseconde est utilisée pour empêcher deux tirs successifs d'un même neurone présynaptique durant ce laps de temps. Il devient ainsi possible de distinguer l'interaction d'une impulsion présynaptique proche d'une impulsion postsynaptique de celle entre deux entrées présynaptiques consécutives.

D'un point de vue pratique, considérant la situation de deux images successives k et $k+1$, l'instant exact de tir d'un neurone présynaptique dû à un changement suffisant d'intensité entre ces deux images est tiré de façon aléatoire et uniforme dans une fenêtre temporelle d'une milliseconde, juste avant le début de l'image $k+1$. Ceci permet d'obtenir une situation plus asynchrone, en évitant que les impulsions présynaptiques aient toutes lieu exactement à l'instant où une nouvelle image est présentée.

Enfin, le bruit ajouté dans certaines des vidéos artificielles employées comme entrées des simulations présentées ci-après correspond en moyenne à 2 impulsions supplémentaires et 0,1 impulsion manquante[I] pour une centaine d'impulsions correctes.

Quid des sorties ? Les neurones de sortie sont de type « intègre-et-tire avec fuites ». De manière analogue aux simulations présentées dans le chapitre 3, ces neurones sont implémentés de façon purement fonctionnelle : de simples intégrateurs intégrant un terme de fuite et émettant une impulsion lorsque leur variable interne atteint un seuil donné. Après avoir tiré, un neurone de sortie est remis à zéro durant une période réfractaire de 20 ms. Par ailleurs, le neurone ayant tiré remet également à zéro durant 15 ms l'ensemble des autres neurones de sortie. Ce

62. P. LICHTSTEINER, C. POSCH et T. DELBRÜCK, IEEE Journal of Solid-State Circuits, 2008.

I. Pour rappel, les motifs activent bien moins de pixels qu'il n'y en a d'éteints.

mécanisme d'inhibition latérale permet de limiter la redondance entre les motifs appris.

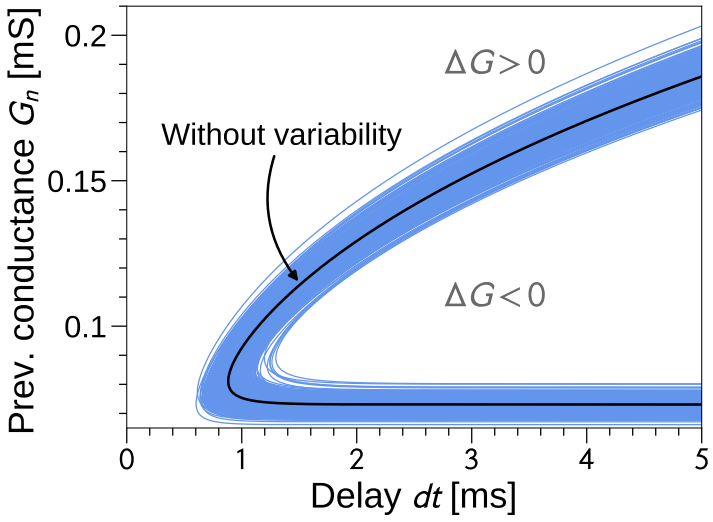


FIGURE 4.11 – Chaque courbe bleue est l'isoligne zéro du paysage d'évolution de la conductance ΔG (figure 4.7b) associée à l'une de 3×162 synapses utilisées lors de l'une des simulations présentant de la variabilité entre composants synaptiques. Des dispersions gaussiennes indépendantes de 3 % de coefficient de variation sont ajoutées sur σ_U, U_∞ et τ_U de l'équation (4.6), et de 5 % de coefficient variation sur le préfacteur exponentiel C_0 de l'équation (4.3). La courbe noire illustre la situation sans variabilité.

Ajout de variabilité entre composants synaptiques. Pour étudier en simulation la résilience du système à la variabilité entre synapses, une solution consiste à ajouter une dispersion sur tout ou partie des paramètres du modèle synaptique, constitué des équations (4.2) à (4.6). Dans les simulations présentées ci-après et incluant de la variabilité synaptique, une dispersion gaussienne indépendante a été ajoutée sur chacun des paramètres de l'équation (4.6), c'est-à-dire sur σ_U, τ_U et U_∞, avec un coefficient de variation de 3 %, ainsi que sur le préfacteur exponentiel C_0 de l'équation (4.3) décrivant la constante de temps τ_{fac} de la relaxation naturelle, avec un coefficient de variation de 5 %. La figure 4.11 illustre les conséquences de ces dispersions entre composants sur l'isoligne nulle de l'évolution de conductance ($\Delta G = 0$) dans le paysage présenté à la figure 4.7b. Chaque ligne bleue correspond à l'une des 3×162 synapses utilisées lors d'une simulation à l'échelle du système, tandis que la courbe noire correspond à la situation sans variabilité. En présence de variabilité, nous observons un comportement sensiblement différent d'une synapse à l'autre. Nous pouvons notamment noter sur cette figure que l'ajout d'une variabilité de dynamique à priori contenue provoque un déplacement significatif de la pointe de l'isoligne nulle, celle-ci fluctuant approximativement d'un tiers de part et d'autre de sa valeur en l'absence de variabilité synaptique.

4.3.2 Résultats d'apprentissage des simulations

Pour apprendre, les synapses exploitent les différents mécanismes plastiques des cellules électrochimiques métalliques Ag_2S, qu'il s'agisse de la transition d'une plasticité à court terme vers une plasticité à plus long terme, ou encore du comportement de type *Spike-Timing Dependent Plasticity*. Parce que les impulsions présynaptiques et postsynaptiques sont de formes exactement opposées, l'apprentissage est entièrement dirigé par les distances temporelles entre les impulsions reçues par les synapses. Le fait que la règle d'apprentissage soit intrinsèque aux composants synaptiques est intéressant en termes de circuits CMOS : les événements de programmation ont des formes d'onde simples et il n'est pas nécessaire d'utiliser d'autres circuits que ceux des neurones pour mettre en œuvre la règle d'apprentissage, contrairement à des systèmes utilisant d'autre nanocomposants mémristifs[264].

4.3.2.1 Évolution des cartes de conductance

Comparaison entre situation initiale et situation finale. Les deux ensembles de six images de la figure 4.12 présentent les cartes de conductance des synapses au début ($t = 0\,s$) et à l'issue ($t = 7{,}2\,s$) d'une simulation. La couleur de chacun de leurs pixels indique la valeur de la conductance de la cellule électrochimique métallique Ag_2S connectée au couple de neurones correspondant. Les pixels de ces cartes sont organisés de la même façon que ceux des vidéos d'entrée.

À $t = 0\,s$, avant l'application des premières impulsions, les synapses sont dans un état de potentialisation modérée (conductance moyenne de 0,2 mS, avec un coefficient de variation de 16 %). À l'issue de la simulation ($t = 7{,}2\,s$), nous pouvons observer sur les cartes de conductance correspondantes le résultats d'un apprentissage fructueux : chaque neurone de sortie s'est spécialisé dans une voie de circulation, comme l'indiquent les motifs de 3 synapses fortement potentialisées sur chacune des cartes. Les motifs appris par les cartes de conductance supérieures et inférieures correspondent respectivement au front arrière et au front avant des motifs se déplaçant dans la vidéo d'entrée. Ceci suggère que l'apprentissage est principalement sensible aux corrélations temporelles très rapprochées entre les impulsions d'entrée et les impulsions de sortie.

Étant donné que les neurones d'entrée sont réglés pour ne pas pouvoir tirer plus d'une fois par milliseconde[I] et que les synapses sont initialement potentialisées de façon modérée, la conductance de ces dernières fait l'objet d'une relaxation si elles sont uniquement exposées à des impulsions présynaptiques : les trajectoires associées dans le paysage de la figure 4.7b demeurent dans la zone hachurée d'évolution négative de la conductance. Contrairement à d'autres nanocomposants mémristifs, la dépression des cellules électrochimiques métalliques Ag_2S n'est pas obtenue en superposant des impulsions de forme d'onde *ad hoc*. À la place, nous

264. S. SAÏGHI et coll., Frontiers in neuroscience, 2015.

I. La présence de délais entre impulsions présynaptiques inférieurs à 1 ms ne peut être due qu'à un possible bruit.

exploitons la relaxation naturelle des composants dans leur régime de plasticité à court terme pour mettre en œuvre la dépression synaptique (équation (4.2)). Cette relaxation est régie par la constante de temps τ_{fac} du modèle analytique décrit précédemment. Ce paramètre dépend de la conductance de la cellule électrochimique métallique Ag_2S : la relaxation de la cellule est d'autant plus rapide que la conductance de cette dernière est faible (équation (4.3)). Seuls des couples formés d'une impulsion présynaptique et d'une impulsion postsynaptique faiblement espacées dans le temps sont en mesure de déclencher l'apprentissage synaptique, en excitant la dynamique de type *Spike-Timing Dependent Plasticity* qui permettra de basculer du régime de plasticité à court terme vers une potentialisation à long terme.

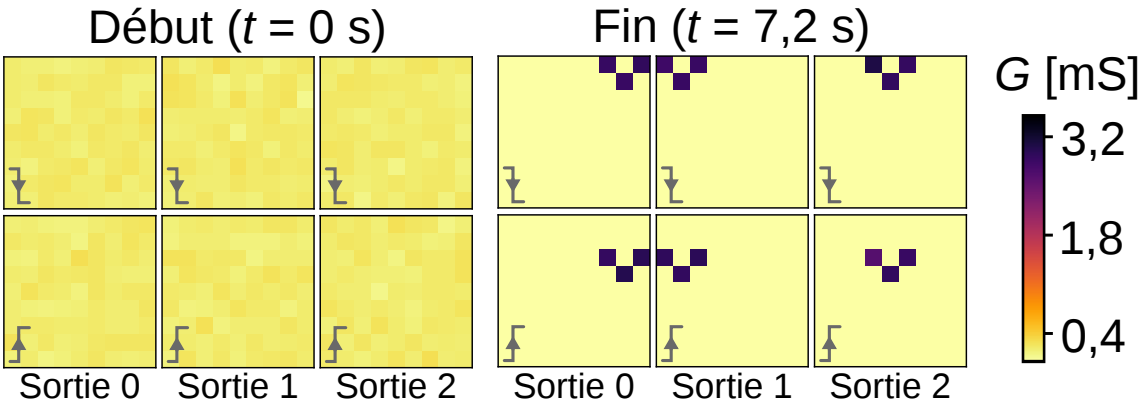


FIGURE 4.12 – Les 3×2 cartes de conductance du réseau synaptique, à gauche juste avant de commencer à présenter les premières entrées ($t = 0\,\text{s}$) et à droite juste après les dernières impulsions présynaptiques ($t = 7{,}2\,\text{s}$). Les annotations « Sortie 0/1/2 » font référence au neurone de sortie, tandis que le petit symbole gris en bas à gauche de chacune des vignettes indique le type de neurones présynaptiques auquel les synapses sont connectées (voir la figure 4.10). La couleur de chaque pixel indique la valeur de la conductance G pour la synapse associée.

Évolution synaptique au cours de l'apprentissage. La figure 4.13 présente l'évolution au cours du temps de la conductance des 81 synapses appartenant à l'une des deux cartes de conductance associées au neurone de sortie n° 2 de la figure 4.12, qui s'est spécialisé dans la détection des motifs circulant sur la voie centrale. Le motif appris par les synapses de cette carte de conductance est constitué de trois éléments fortement potentialisés à l'issue de l'apprentissage (en violet foncé dans la carte de conductance en haut à droite). Les courbes en rouge foncé sur le graphe principal correspondent à ces trois synapses, tandis que les courbes associées aux autres éléments synaptiques, qualifiées de « fond », sont tracées en bleu clair. À $t = 0\,\text{s}$, les valeurs de conductance sont distribuées selon une distribution proche d'une gaussienne, avec une moyenne de 0,2 mS et un coefficient de variation de 16 %. Les impulsions qui seraient employées dans la pratique pour atteindre un tel état sont supposées s'être terminées 80 ms auparavant. Entre 0 et 7,2 s, les impulsions présynaptiques issues d'une vidéo artificielle sont appliquées sur la matrice de cellules électrochimiques métalliques Ag_2S.

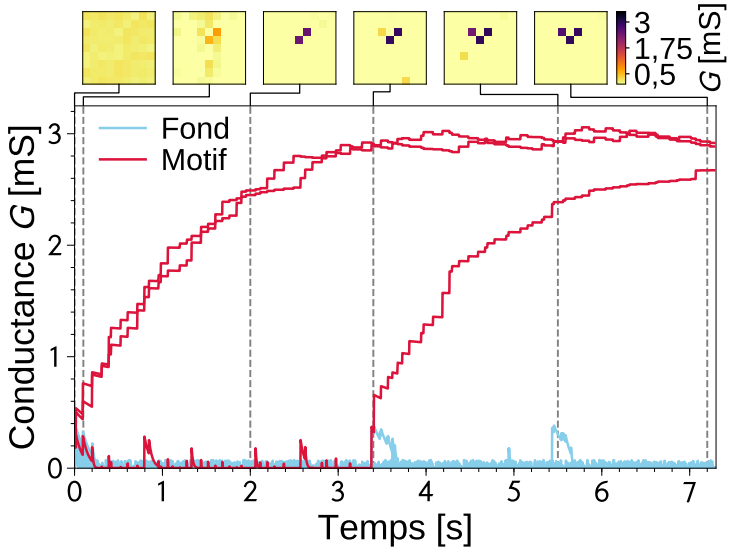


FIGURE 4.13 – Évolution de la conductance des 81 cellules électrochimiques métalliques Ag_2S connectées au même neurone de sortie ainsi qu'aux neurones d'entrée sensibles à un incrément d'intensité des pixels. Les courbes rouges correspondent aux synapses qui forment le motif appris, tandis que les lignes bleues correspondent quant à elles à toutes les autres synapses, qualifiées de « fond ». Les vignettes du haut sont des cartes de conductance des 81 synapses étudiées. Elles sont tracées à six instants différents, indiqués par des lignes verticales en pointillé gris.

Durant les premières centaines de millisecondes a lieu une relaxation de la conductance de la totalité des synapses mis à part deux d'entre elles. Un motif partiel, composé de deux pixels sur trois, commence à être appris par le biais de couples d'impulsions présynaptiques et d'impulsions postsynaptiques, qui potentialisent graduellement les deux synapses en question. Comme nous le verrons par la suite avec notamment la figure 4.15, ces paires d'impulsions n'excitent pas toutes la dynamique de type *Spike-Timing Dependent Plasticity*. Néanmoins ce comportement plastique à court terme de type *Spike-Timing Dependent Plasticity* est essentiel au processus d'apprentissage : des simulations où cette plasticité est désactivée à dessein montrent que dans ce cas l'ensemble des synapses connait une relaxation de leur conductance et aucun apprentissage n'est alors observé.

Au cours de trois premières secondes, la synapse manquante du motif en cours d'apprentissage ne connaît que de faibles potentialisations, ce qui la maintient dans son régime de plasticité à court terme : elle ne parvient pas à sortir de la zone d'évolution ΔG négative dans le paysage de la figure 4.7b page 138. Autour de $t = 3{,}4\,\text{s}$, deux paires d'impulsions excitant la dynamique de type *Spike-Timing Dependent Plasticity* sont appliquées coup sur coup sur cette troisième synapse, ce qui déclenche l'évolution de cette dernière vers une potentialisation à long terme. Des paires d'impulsions supplémentaires, excitant ou non la dynamique de type *Spike-Timing Dependent Plasticity*, finissent ensuite de la potentialiser, permettant d'obtenir un motif complet.

Robustesse au bruit. À environ $t = 3{,}4\,\text{s}$ et $t = 5{,}4\,\text{s}$, nous pouvons observer la potentialisation partielle erronée de deux synapses appartenant au fond. Sur les vignettes des cartes conductance correspondantes de la figure 4.13, ces deux synapses apparaissent en jaune plus foncé que le reste du fond. Sachant que le délai entre des impulsions présynaptiques correctes appliquées sur une même synapse est toujours supérieur à 1 ms, ces potentialisations accidentelles sont dues au bruit qui peut être à l'origine de paires d'impulsions non significatives. Dans le cas présent, le bruit est ajouté selon la méthodologie présentée dans la partie 4.3.1.2. Au contraire des paires d'impulsions répétées liées aux motifs sujets à un apprentissage, les occurrences de tels événements restent rares et ne déclenchent donc pas de potentialisation à long terme : les synapses concernées sont déprimées naturellement grâce à la relaxation intrinsèque de leur conductance. À l'issue de la relaxation de la conductance des synapses du fond, le système présente ainsi une tolérance à un bruit de fréquence modérée[I].

4.3.2.2 Résultats synthétiques

À l'issue d'une simulation, une stratégie simple pour juger de la qualité de l'apprentissage se base sur une évaluation visuelle des cartes de conductance finales : un motif est considéré comme « appris nettement » s'il présente une paire de cartes de conductance sur lesquelles exactement 2×3 synapses sont fortement potentialisées à long terme et forment un arrangement similaire à ceux des cartes de conductance finales de la figure 4.12 page 149. Il convient de noter que ce critère est particulièrement drastique. Une étude de visu des impulsions postsynaptiques montre qu'un neurone de sortie associé à une paire de cartes de conductance totalisant 5 ou 7 synapses fortement potentialisées peut se révéler capable de détecter correctement les motifs se déplaçant le long d'une voie.

La table 4.1 récapitule la proportion de simulations à l'issue desquelles au moins deux motifs (sur les trois possibles) étaient appris nettement, dans différentes situations. Sans variabilité synaptique, le système apprend nettement au moins deux motifs dans plus de 93 % des simulations dans le cas d'entrées idéales sans bruit et demeure capable de maintenir une performance de 70 % dans le cas où les impulsions présynaptiques sont sujettes au bruit décrit dans la partie 4.3.1.2 (et que nous avons pu observer indirectement sur la figure 4.13).

Les simulations effectuées en ajoutant une dispersion artificielle sur certains paramètres clefs du modèle synaptique (détails dans la partie 4.3.1.2) montrent une tendance analogue, avec un apprentissage net d'au moins deux motifs à l'issue de 85 % des simulations effectuées en l'absence de bruit et de 60 % d'entre elles dans la situation d'entrées présynaptiques bruitées. Contrairement à ce que l'observation de la figure 4.11 aurait pu laisser craindre, ces premiers résultats suggèrent une assez bonne tolérance du système à la variabilité entre synapses. Et ce malgré les dimensions réduites de ce dernier, qui limitent les effets de redondance. Même

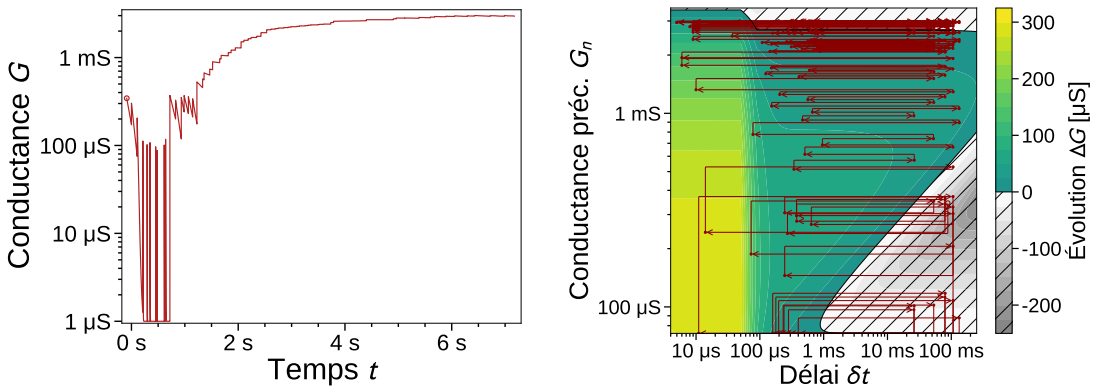
I. Dans les simulations présentées, en raison de la façon dont le bruit est ajouté sur les données vidéo d'entrée, la fréquence de ce dernier ne peut pas excéder 450 Hz. De plus, une telle situation est peu probable : elle constitue le *pire* des cas possibles pour un pixel, avec un événement de bruit par image.

	Entrées non bruitées	Entrées bruitées
Sans variabilité synaptique	93,3 %	70 %
Avec variabilité synaptique	85 %	60 %

TABLE 4.1 – Proportion des simulations à l'issue desquelles au moins deux des trois paires de motifs attendues étaient nettes sur les cartes de conductances. 120 et 60 simulations ont été effectuées, respectivement avec et sans variabilité entre composants synaptiques (que les entrées aient été bruitées ou non).

en durcissant encore la métrique adoptée, pour ne retenir par exemple que les situations d'apprentissage de trois motifs nets après présentation d'entrées non bruitées, la performance du système en présence de variabilité synaptique se maintient à 38,3 %, contre 43,3 % pour un système aux synapses de caractéristiques toutes identiques.

4.3.2.3 Interaction entre les différents comportements plastiques



(a) Évolution temporelle de la conductance dans le cas d'une succession réaliste d'impulsions.

(b) Trajectoire associée dans le paysage d'évolution de la conductance.

FIGURE 4.14 – figure similaire aux figures 4.8 et 4.9, présentant en rouge foncé *(a)* l'évolution temporelle de la conductance d'une synapse lors de l'apprentissage d'un motif et *(b)* les déplacements associés dans le paysage d'évolution de conductance ΔG de la figure 4.7b. Ces résultats sont tirés d'une simulation à l'échelle d'un système et illustrent l'interaction entre les différents comportements plastiques des synapses électrochimiques métalliques Ag_2S sur laquelle repose l'apprentissage.

La figure 4.14 illustre la manière dont interagissent les différents mécanismes de plasticité d'une même cellule électrochimique métallique Ag_2S pour mener à la potentialisation à long terme de cette synapse lors d'une simulation à l'échelle d'un système. Sur le panneau (b), la trajectoire de l'évolution de conductance ΔG dans le paysage de la figure 4.7b est plus complexe que celles des quelques situations schématiques des figures 4.8 et 4.9. Néanmoins il est possible de reconnaître certains des comportements précédemment décrits.

Partant d'une situation initiale dans la zone d'évolution de conductance ΔG négative (symbole ○ dans la région hachurée), la synapse étudiée demeure tout d'abord dans son régime de plasticité à court terme, effectuant des cycles entre des niveaux de faible potentialisation. Après environ 1 s, plusieurs paires d'impulsions finissent par exciter son comportement de type *Spike-Timing Dependent Plasticity*, causant des incréments de conductance suffisamment importants pour déclencher la transition vers le régime de plasticité à plus long terme. Les impulsions reçues par la suite achèvent la potentialisation à long terme de cette synapse, quelle que soit leur nature.

La figure 4.14b permet de mieux comprendre le rôle primordial du comportement plastique de type *Spike-Timing Dependent Plasticity* observé à la partie 4.3.2.1 : sans lui, même la potentialisation causée par une paire d'impulsions demeure insuffisante pour sortir du régime de plasticité à court terme une synapse de très faible conductance.

La figure 4.15 présente la fonction de répartition empirique des délais entre impulsions inférieurs à une milliseconde en agrégeant à l'issue d'une simulation toutes les distances temporelles entre les impulsions appliquées sur chacune des 3×162 synapses du système. Ces délais sont nécessairement le résultat d'une paire constituée d'une impulsion présynaptique et d'une impulsion postsynaptique puisque les neurones d'entrée comme ceux de sortie sont réglés pour ne pas pouvoir émettre deux impulsions consécutives espacées de moins d'une milliseconde. Les valeurs sont distribuées de façon relativement continue, chaque comportement plastique de dynamique rapide des cellules électrochimiques métalliques Ag_2S étant stimulé. En particulier, environ un tiers des paires excite le comportement de type *Spike-Timing Dependent Plasticity*, absolument nécessaire au bon apprentissage des motifs : dans les simulations où ce comportement est désactivé[I], aucun apprentissage n'est observé. Par ailleurs, cette figure permet d'estimer à approximativement un tiers la proportion de ces paires stimulant la dynamique de type *Spike-Timing Dependent Plasticity* sans être constituées d'impulsions qui se superposent.

I. En remplaçant les expressions des paramètres A et U du modèle par leur valeur asymptotique.

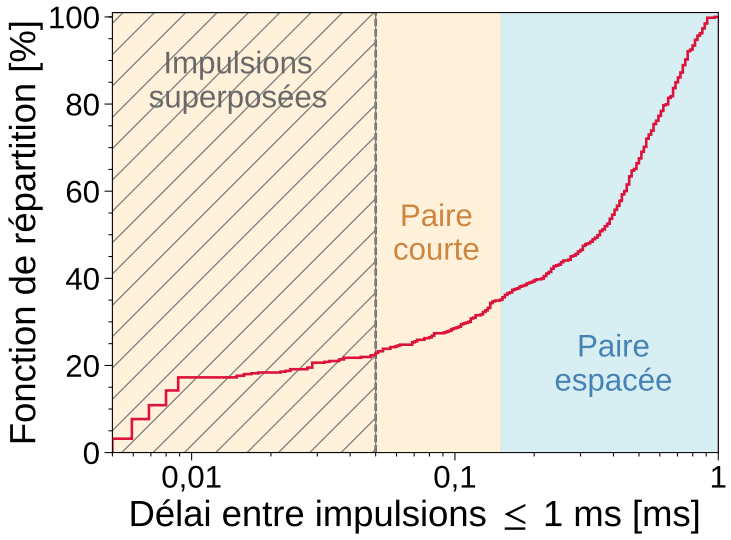


FIGURE 4.15 – La courbe rouge représente la fonction de répartition empirique des distances temporelles entre impulsions qui sont inférieures à 1 ms, dans un cas où les trois motifs ont été appris. Les délais employés sont tirés de l'agrégation des 3×162 ensembles de délais entre impulsions appliquées sur chacune des 3×162 synapses du système. La région hachurée désigne les délais inférieurs à la durée d'une impulsion. Les zones orange et bleue indiquent respectivement si un délai excite ou non la dynamique de type *Spike-Timing Dependent Plasticity* des cellules électrochimiques métalliques Ag_2S modélisées. La valeur de 150 µs pour la frontière entre ces deux régions s'appuie sur l'observation du paysage d'évolution de la conductance ΔG de la figure 4.7b. *NB* : les transitions très marquées pour les délais inférieurs à 10 µs sont le fruit du pas de temps employé lors des simulations.

4.4 Conclusions

4.4.1 Comparaison avec les jonctions tunnel magnétiques à transfert de spin

L'usage synaptique des cellules électrochimiques métalliques Ag_2S se révèle assez différent de celui des jonctions tunnel magnétiques à transfert de spin étudié au chapitre 3, chacune de ces familles de nanocomposants présentant ses avantages et ses inconvénients.

Des états de résistance extrêmes différents. En termes de rapport R_{OFF}/R_{ON} entre la valeur de résistance à l'état OFF et celle à l'état ON, l'avantage est nettement aux cellules électrochimiques métalliques Ag_2S, puisqu'il est de trois ordres de grandeur même au sein des composants académiques étudiés dans ce chapitre, alors qu'il peine à dépasser quelques fois l'unité dans les meilleures jonctions tunnel magnétiques à transfert de spin. Par ailleurs, l'état OFF des premières est de meilleure qualité puisque leur résistance y est de plusieurs centaines de kiloohms voire du mégaohm, contre quelques dizaines de kiloohms tout au plus pour les composants spintroniques. Ces deux caractéristiques font à priori des cellules électrochimiques métalliques Ag_2S des bons candidats pour une matrice synaptique sans composants de sélection (structure de type « 1R »), là où une étude plus approfondie des problématiques au niveau circuit pourrait révéler la nécessité d'ajouter des composants de sélection dans une matrice de synapses exploitant des jonctions tunnel magnétiques (structure de type « 1T-1R »), notamment lorsque les dimensions de cette dernière deviennent importantes. Toutefois, l'usage binaire des jonctions tunnel magnétiques proposé au chapitre 3 permet d'utiliser des circuits de lecture spécialisés[190] capables de distinguer aisément l'état synaptique malgré le faible rapport R_{OFF}/R_{ON}, là où l'usage analogique des cellules électrochimiques métalliques est davantage exposé au bruit et à la variabilité entre composants. Il est cependant à noter que l'utilisation d'une architecture neuromorphique permet dans les deux cas d'être tolérant à des niveaux significatifs de variabilité entre composants.

Énergie consommée. En termes de consommation énergétique des composants synaptiques[I], l'avantage semble pour le moment aux composants spintroniques : avec les paramètres utilisés pour le chapitre 3, un événement de programmation d'une jonction magnétique requiert une énergie de l'ordre du picojoule, tandis que dans le cas des cellules électrochimiques métalliques Ag_2S étudiées, cette énergie est typiquement de l'ordre de la centaine de picojoules voire du nanojoule. En effet, bien que les amplitudes des impulsions de tension et les états de plus faible résistance soient de valeurs similaires, les composants filamentaires souffrent d'une durée d'impulsion supérieure de trois ordres de grandeur à celle des composants spintroniques. En outre, en séparant les phases d'écriture de celles de lecture à l'aide de circuits de lecture spécialisés

190. W. S. ZHAO et coll., IEEE Transactions on Magnetics, 2009.

I. Les valeurs sont calculées en considérant uniquement les synapses, c'est-à-dire en faisant abstraction des possibles circuits de contrôle externes.

qui utilisent des impulsions d'amplitude réduite, les jonctions tunnel magnétiques permettent possiblement de réduire encore davantage la consommation associée aux phases d'inférence[I], là où les cellules électrochimiques métalliques Ag_2S utilisent des impulsions identiques durant les deux types de phase.

Une différence d'approche à propos de la *Spike-Timing Dependent Plasticity*. Dans les deux cas, le schéma d'apprentissage de type *Spike-Timing Dependent Plasticity* adopté est original en comparaison de la règle d'évolution généralement recherchée avec d'autres nanocomposants de la littérature[120]. Les caractéristiques de la règle retenue et sa mise en œuvre diffèrent néanmoins sensiblement entre les cellules électrochimiques métalliques du présent chapitre et les jonctions tunnel magnétiques à transfert de spin.

Dans le cas des composants d'électronique de spin, l'apprentissage reste contrôlé par un ensemble de circuits externes aux synapses, à priori dans une technologie CMOS. Un exemple de tel circuit est celui chargé de définir le caractère causal ou non de l'association d'un événement postsynaptique et d'une impulsion présynaptique, en fonction d'une fenêtre temporelle qui dépend de l'application visée. Seul l'aspect probabiliste des phases de programmation, nécessaire pour pallier le caractère binaire des synapses en question, bénéficie des propriétés intrinsèques des nanocomposants. Ceci n'est cependant pas négligeable puisque le recours à des circuits CMOS pour générer des nombres aléatoires demeure une solution coûteuse en surface de silicium, comme nous avons par exemple pu le constater dans le cas de la puce TrueNorth[13].

En ce qui concerne les cellules électrochimiques métalliques, une différence majeure est le caractère *symétrique* de la règle d'apprentissage. Une impulsion présynaptique peut avoir le même effet sur le comportement plastique qu'une impulsion postsynaptique, ce qui autorise une stratégie d'apprentissage reposant davantage sur la notion de proximité temporelle des impulsions appliquées sur une même synapse plutôt que l'existence d'une causalité entre ces événements. Ceci permet d'utiliser uniquement des impulsions électriques identiques de forme d'onde rectangulaire, ce qui simplifie les circuits électroniques associés. L'obtention d'un apprentissage hebbien, c'est-à-dire reposant sur une activation jointe de l'entrée présynaptique *et* de la sortie postsynaptique, requiert toutefois de pouvoir distinguer les deux types d'événements. Dans ce chapitre, nous avons ainsi proposé une solution à l'échelle du système, au travers de la limitation en fréquence des entrées présynaptiques, qui permet de séparer artificiellement les paires d'impulsions pertinentes de la simple succession d'événements de même nature. Dans la pratique, cette approche suppose la bonne adéquation de la dynamique des entrées présynaptiques aux échelles de temps des cellules électrochimiques métalliques employées comme synapses, et à défaut le recours à une étape de prétraitement des signaux d'entrée.

I. À condition de montrer que la consommation du circuit de lecture *ad hoc* ne vient pas réduire à néant les gains obtenus.

13. P. A. MEROLLA et coll., Science, 2014.

4.4.2 Bilan et perspectives

Un apprentissage résultant de l'interaction de plusieurs mécanismes de plasticité synaptique. Ce chapitre a permis de présenter les différents mécanismes plastiques coexistant dans des cellules électrochimiques métalliques Ag_2S fabriquées et caractérisées par nos collaborateurs de l'Institut d'Électronique, de Microélectronique et de Nanotechnologie de l'Université de Lille I. En sus d'une forme de plasticité non hebbienne et d'une transition d'un régime de rétention à court terme vers un régime à plus long terme, nous avons observé avec nos collaborateurs l'existence d'un comportement additionnel de type *Spike-Timing Dependent Plasticity* au sein de ces composants mémristifs.

L'extension du modèle d'évolution de la conductance précédemment développé par nos collaborateurs nous a permis de mieux comprendre comment ces différentes plasticités et leur interaction pouvaient être exploitées dans le cadre d'un usage synaptique pour des systèmes neuromorphiques. Des simulations fonctionnelles d'un système de taille modeste utilisant ces composants comme synapses analogiques artificielles ont permis d'obtenir une preuve de concept d'un apprentissage reposant sur l'interaction de ces différents mécanismes plastiques. Le fait d'utiliser des comportements internes aux éléments synaptiques permet notamment de réduire significativement la complexité des circuits externes de programmation ou de contrôle associés.

Par ailleurs, les premières études réalisées montrent qu'un tel système ne voit pas ses performances s'écrouler en présence de bruit sur les entrées présynaptiques ou de variabilité entre les éléments synaptiques.

Pas de panique, il reste des choses à faire. Les résultats précédents encouragent à poursuivre ces travaux sur un usage synaptique des cellules électrochimiques métalliques qui exploitent les multiples comportements plastiques offerts ces composants.

Dans le domaine de la fabrication, se pose notamment la question de l'ingénierie de leurs caractéristiques pertinentes pour concevoir un système neuromorphique. En particulier la possibilité de moduler les échelles de temps auxquelles ils réagissent. Un certain contrôle de la transition d'une rétention à court terme vers une rétention à plus long terme est possible en jouant sur le courant utilisé lors de l'étape de formation[133], mais pouvoir ajuster les caractéristiques du comportement de type *Spike-Timing Dependent Plasticity*, qui semble être au moins en partie d'origine thermique, pourrait par exemple nécessiter d'optimiser la structure même de la cellule, à la manière de ce qui est fait pour les cellules mémoires à changement de phase (dont la transition repose sur un phénomène thermique). La mise au point d'un nouveau modèle analytique du comportement de ces composants, partant de la physique de ces derniers, pourrait être un point de départ pour mener à bien cette étude.

À l'échelle d'un système, il serait bien entendu intéressant d'entreprendre l'étude de systèmes de plus grandes dimensions, ce qui permettrait d'évaluer les effets de la variabilité entre composants ou encore du bruit dans un cadre plus réaliste, notamment en termes d'effets de

redondance entre composants synaptiques. Si d'un point de vue du simulateur développé pour les travaux du présent chapitre, ce passage à l'échelle ne pose pas d'autres problèmes que celui d'un allongement des durées de simulation, il risque néanmoins de nécessiter la mise au point d'un protocole d'évaluation des performances du système plus automatisé que celui utilisé pour le moment, les motifs appris devenant potentiellement plus complexes. Par ailleurs, il serait envisageable d'exploiter la flexibilité offerte par les simulations pour guider les travaux portant sur la fabrication des composants. Une étude visant à évaluer l'influence des différents paramètres synaptiques sur les capacités d'apprentissage d'un système complet pourrait par exemple permettre d'identifier les principaux paramètres qu'il serait utile de pouvoir moduler.

Chapitre 5

Usage synaptique de jonctions tunnel magnétiques à transfert de spin : réflexions à l'échelle du circuit

« Qui ne sait les détails ignore l'ensemble. »

Proverbe turc, rapporté par Charles CAHIER

“CE CHAPITRE *explore des pistes de réflexion sur des problématiques spécifiques à l'échelle du circuit dans le cadre d'un usage synaptique des jonctions tunnel magnétiques à transfert de spin. Ceci constitue une occasion d'évaluer la pertinence des stratégies retenues dans les chapitres précédents à l'échelle d'un système complet. Les réflexions présentées ci-après se focalisent sur le choix d'une stratégie de fonctionnement analogique ou numérique et du potentiel de passage à l'échelle associé. Sont abordées la phase de lecture liée à l'inférence, ainsi que la phase d'écriture mettant en œuvre la règle d'apprentissage. La motivation de ce chapitre est de permettre d'orienter une future étude à l'échelle du circuit plus approfondie, reposant sur l'utilisation d'outils de conception assistée par ordinateur, notamment de type SPICE. ”*

5.1 Introduction

Quelques questions qui se posent à l'échelle du circuit. Les simulations fonctionnelles à l'échelle du système présentées dans les chapitres 3 et 4 nous ont permis de mettre en lumière l'intérêt des jonctions tunnel magnétiques à transfert de spin et des cellules électrochimiques métalliques Ag_2S comme synapses artificielles au sein d'une architecture neuro-inspirée. Elles ne tiennent néanmoins pas compte d'un certain nombre de problématiques, propres aux *circuits électroniques* nécessaires à la mise en œuvre des étapes de fonctionnement du système complet.

Dans ce chapitre, nous réfléchissons sur la plausibilité du fonctionnement des systèmes précédemment étudiés, à l'échelle des circuits électroniques que ces derniers requièrent. Les conclusions de ce type d'étude peuvent sensiblement dépendre de la technologie des composants employés. Sachant que les cellules électrochimiques métalliques Ag_2S du chapitre 4 sont encore dans une phase de développement académique, nous nous focaliserons donc dans ce chapitre sur un système analogue à celui du chapitre 3, exploitant des jonctions tunnel magnétiques à transfert de spin, des composants dont la maturité technologique et industrielle est désormais établie.

Le système étudié au chapitre 3 présente deux phases distinctes de fonctionnement, avec d'un côté une phase de lecture liée à l'inférence et de l'autre une phase d'écriture, où les éléments synaptiques sont programmés selon la règle d'apprentissage adoptée. Dans la suite de ce chapitre, nous étudierons ces deux phases indépendamment, en nous intéressant notamment à la façon dont le passage à l'échelle de la matrice synaptique affecte chacune d'entre elles. Nous observerons en particulier le rôle limitant de l'entrance et de la sortance des circuits neuronaux associés à chacune de ces phases.

Comme cela a été évoqué dans le chapitre 1, en comparaison des autres technologies de mémoires résistives innovantes, les jonctions tunnel magnétiques présentent un faible rapport d'états de résistance R_{OFF}/R_{ON} (il est à peine supérieur à deux avec les valeurs données dans la partie suivante). Il s'agit donc d'évaluer, du point de vue des circuits, les défis que pose ce faible écart entre les éléments synaptiques potentialisés et ceux déprimés.

Une problématique pour le moment ouverte et en partie liée au rapport R_{OFF}/R_{ON} est le recours ou non à un composant de sélection pour chaque cellule mémoire de la matrice synaptique. L'un des problèmes limitant les dimensions d'une matrice synaptique sans composant de sélection (structure 1R) sont les courants de fuite parasites[234,265] (*sneak path currents* en anglais) circulant à travers les points mémoires seulement sélectionnés à moitié. Il s'agit donc également de comparer les avantages et les inconvénients qu'il y a à ajouter un transistor de sélection aux cellules mémoires de la matrice synaptique (structure 1T-1R).

234. W. S. ZHAO et coll., IEEE Transactions on Nanotechnology, 2012.
265. A. SERB et coll., IEEE International Symposium on Circuits and Systems, 2015.

Conditions de l'étude. Les réflexions de ce chapitre se placent majoritairement dans le cadre de la tâche présentée au chapitre 3. Il s'agit d'un système constitué d'une matrice entièrement connectée de jonctions tunnel magnétiques à transfert de spin, dont les impulsions émises par $2 \times 128 \times 128$ entrées présynaptiques sont issues d'un enregistrement réalisé par une rétine neuromorphique de type *Dynamic Vision Sensor*[62] observant le passage de véhicules sur une autoroute.

Remarque

Une particularité notable des signaux d'entrée utilisés, liée au caractère asynchrone du capteur les ayant produites ainsi qu'à la précision temporelle de ce dernier (de l'ordre de la microseconde), est le fait que chaque impulsion présynaptique est disjointe des autres. Ceci signifie qu'une impulsion présynaptique est toujours appliquée seule, sur une unique ligne de la matrice d'éléments synaptiques.

Le modèle du comportement stochastique des jonctions tunnel magnétiques à transfert de spin présenté au chapitre 2 a été adapté en Verilog-A par Nicolas LOCATELLI[I]. Ce langage de description matérielle à temps continu permet la simulation de notre composant dans des circuits électriques à l'aide d'outils logiciels industriels de type SPICE (dans notre cas le simulateur Spectre de Cadence Design Systems®). L'un des intérêts de telles simulations est de décrire de façon réaliste le comportement des transistors présents dans les circuits, en utilisant un kit de conception de procédés (PDK[II]) fourni par une entreprise de fonderie. Dans notre cas, le kit utilisé est celui d'un nœud 28 nm basse consommation d'un acteur majeur franco-italien de l'industrie des semi-conducteurs. En nous basant sur des simulations Spectre avec ce kit de conception de procédés, sous une tension d'alimentation $V_{dd} = 1{,}2\,\text{V}$, des transistors dopés négativement (NMOS) et positivement (PMOS) de longueur de grille $L = 30\,\text{nm}$ sont capables de conduire respectivement 1000 et 615 microampères par micromètre de largeur de grille (notée W).

Les jonctions tunnel magnétiques à transfert de spin simulées sont à aimantation perpendiculaire, d'un diamètre de 32 nm. Leur résistance R_{ON} dans l'état magnétique parallèle (également notée R_P) est de $6\,\text{k}\Omega$. La magnétorésistance tunnel retenue étant quant à elle de 120 %, la résistance R_{OFF} dans l'état antiparallèle (également notée R_{AP}) est proche de $13\,\text{k}\Omega$. Enfin, l'amplitude I_{c_0} du courant critique est de l'ordre de $30\,\mu\text{A}$.

Ce chapitre demeure une première étude visant à orienter une éventuelle étude plus approfondie sur les problèmes spécifiques à l'échelle du circuit. Il s'agit notamment d'évaluer rapidement, par exemple par des calculs d'ordre de grandeur, la plausibilité de certaines stratégies envisagées dans les chapitres précédents, sans exclure la possibilité de remettre en cause ou d'adapter ces dernières.

62. P. LICHTSTEINER, C. POSCH et T. DELBRÜCK, IEEE Journal of Solid-State Circuits, 2008.

I. À l'époque sous contrat post-doctoral dans l'équipe NANOARCHI du laboratoire.

II. De l'anglais *Process Design Kit*.

5.2 Phase de lecture

Différentes stratégies sont envisageables pour assurer la lecture des états synaptiques durant la phase d'inférence des architectures neuromorphiques présentées dans les chapitres précédents. Les circuits associés peuvent suivrent une approche analogique[266] ou au contraire numérique[267], chacune ayant ses avantages et ses inconvénients[268].

Dans cette partie, nous étudions la plausibilité d'une approche analogique relativement conventionnelle dans le cadre d'un usage synaptique de composants mémristifs innovants, et celle d'une approche exploitant quant à elle un circuit numérique initialement développé pour les puces mémoires à base de jonctions tunnel magnétiques.

5.2.1 Stratégie analogique exploitant un convertisseur à transimpédance

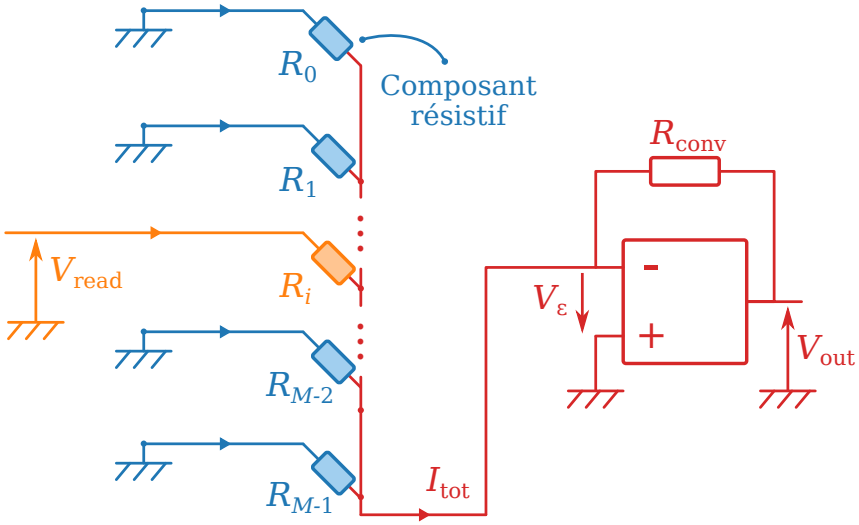


FIGURE 5.1 – Lecture d'un composant i (en orange) parmi une colonne de M (en bleu), tous reliés au même neurone de sortie. L'état de résistance R_{ON} ou R_{OFF} de l'élément synaptique est déterminé à partir du courant I_{tot} sortant de la colonne, qui est converti en une tension V_{out} à l'aide d'un convertisseur à transimpédance (en rouge). Dans un circuit de conversion idéal, la tension de décalage V_ε est nulle.

Lecture par un convertisseur à transimpédance. Cette approche consiste à utiliser un convertisseur à transimpédance pour convertir en une tension V_{out} le courant I_{tot} constitué de la somme des courants traversant les M éléments synaptiques connectés à un même neurone de sortie[266]. Un exemple de cette stratégie est illustré sur la figure 5.1 avec un convertisseur à transimpédance utilisant un amplificateur opérationnel. Avec les notations de cette figure et sous l'hypothèse d'idéalité de l'amplificateur opérationnel ($V_\varepsilon = 0$), la relation entre la tension

266. M. S. QURESHI et coll., IEEE International Symposium on Circuits and Systems, 2011.
267. L. ZHANG et coll., IEEE Transactions on Magnetics, 2014.
268. B. RAJENDRAN et coll., IEEE Transactions on Electron Devices, 2013.

de sortie V_{out} et le courant postsynaptique incident I_{tot} est

$$V_{\text{out}} = -R_{\text{conv}} \times I_{\text{tot}}. \tag{5.1}$$

Dans le cas d'une impulsion présynaptique d'amplitude V_{read}, présentée seule sur la ligne i comme à la figure 5.1, l'équation (5.1) devient alors simplement celle d'un amplificateur inverseur, à savoir

$$V_{\text{out}} = -\frac{R_{\text{conv}}}{R_i} \times V_{\text{read}}, \tag{5.2}$$

où R_i désigne la résistance de la jonction tunnel magnétique lue. L'ajout d'un comparateur à une tension de seuil bien choisie (non représenté) permet ensuite de rendre binaire le signal de sortie d'une telle lecture, autorisant de la sorte le recours à une stratégie numérique pour le reste du circuit neuronal.

Remarque

La situation présentée est celle d'une lecture analogique suivie d'une numérisation par seuillage, qui se prête à l'utilisation de circuits numériques pour le reste du circuit neuronal. D'autres approches existent, comme par exemple la solution davantage analogique proposée par G. LECERF et coll. dans la référence [235], dont l'étage d'entrée est un convoyeur de courant de seconde génération. Le point qui importe ici est la présence d'un étage différentiel similaire à celui de l'amplificateur opérationnel de la figure 5.1, qui est généralement rencontré dans un circuit postsynaptique afin notamment de fixer la tension d'entrée de ce dernier. Les conclusions ultérieures découlant de l'existence d'une tension de décalage V_ε non nulle entre les deux entrées d'un tel étage s'appliquent donc également à ces solutions analogiques.

Dimensions limites de la matrice d'éléments synaptiques. En tenant compte de l'existence d'une tension de décalage V_ε à l'entrée de l'étage de lecture, l'expression du courant postsynaptique incident devient

$$I_{\text{tot}} = \overbrace{\frac{V_{\text{read}} + V_\varepsilon}{R_i}}^{\text{Utile}} + \overbrace{\sum_{j \neq i} \frac{V_\varepsilon}{R_j}}^{\text{Parasites}}, \tag{5.3}$$

faisant apparaître le courant circulant à travers la jonction tunnel magnétique de la ligne i, utile à la lecture, et la somme des contributions des courants parasites traversant l'ensemble des autres composants en raison d'une tension de décalage V_ε non nulle. En appelant M le nombre de synapses dans leur état de résistance R_{OFF} (c'est-à-dire déprimées), l'équation (5.3) devient

$$I_{\text{tot}} = \frac{V_{\text{read}} + V_\varepsilon}{R_i} + \left(m_{\text{OFF}} \times \frac{V_\varepsilon}{R_{\text{OFF}}} + (M - 1 - m_{\text{OFF}}) \times \frac{V_\varepsilon}{R_{\text{ON}}} \right). \tag{5.4}$$

Remarque

La dispersion entre les composants n'est pas prise en compte dans les calculs présentés au sein de ce chapitre. Ce choix est guidé par l'idée que ces derniers servent davantage à obtenir des ordres de grandeurs et dégager les tendances d'évolution qu'à obtenir des valeurs précises. Dans le cadre de cette approche, le gain marginal à utiliser des variables aléatoires dans les équations de ce chapitre est donc limité, essentiellement à la définition de marges par rapport aux résultats fournis par une approche déterministe sans variabilité. Cette étape est donc laissée pour la future étude de type SPICE, plus approfondie, qui permettra par ailleurs de décrire de façon plus réaliste la dispersion touchant les transistors des circuits.

De l'équation (5.4), nous pouvons déduire une estimation du nombre maximal d'entrées présynaptiques $M_{\max}$ qu'il est possible d'utiliser dans le cadre d'une telle structure 1R sans sélecteur. Plaçons nous dans le pire cas de lecture d'une synapse, à savoir celui où la résistance de l'élément lue est maximale (R_{OFF}), tandis que celle de tous les autres éléments est minimale (R_{ON}). Le terme m_{OFF} de l'équation (5.4) est alors nul. Ceci revient à minimiser dans l'équation (5.3) le terme de courant utile et maximiser celui des contributions parasites. Étudions la situation limite où le terme de courant parasite devient juste suffisamment important pour que le courant postsynaptique incident I_{tot} soit interprété comme le résultat de l'activation d'une entrée reliée à une synapse potentialisée, et non déprimée. Mathématiquement, cela se traduit par le franchissement d'un seuil de détection en courant I_{det}. En l'absence d'informations supplémentaires sur la dispersion des composants, nous prendrons ce seuil égal à la moyenne arithmétique du courant traversant une synapse potentialisée et celui circulant à travers une synapse déprimée :

$$I_{\mathrm{det}} = \frac{V_{\mathrm{read}}}{2} \times \left(\frac{1}{R_{\mathrm{ON}}} + \frac{1}{R_{\mathrm{OFF}}} \right). \tag{5.5}$$

En égalisant les équations (5.4) et (5.5) dans le cas où le nombre m_{OFF} de synapses déprimées et non sélectionnées est nul, nous obtenons l'estimation

$$M_{\max} = \left\lfloor \left(\frac{1}{2} \times \frac{V_{\mathrm{read}}}{V_{\varepsilon}} + 1 \right) \times \left(1 - \frac{R_{\mathrm{ON}}}{R_{\mathrm{OFF}}} \right) \right\rfloor \tag{5.6}$$

du nombre maximal d'entrées présynaptiques [I] qu'il est possible d'utiliser sans qu'il ne devienne possible de masquer le courant utile à la lecture par les contributions parasites.

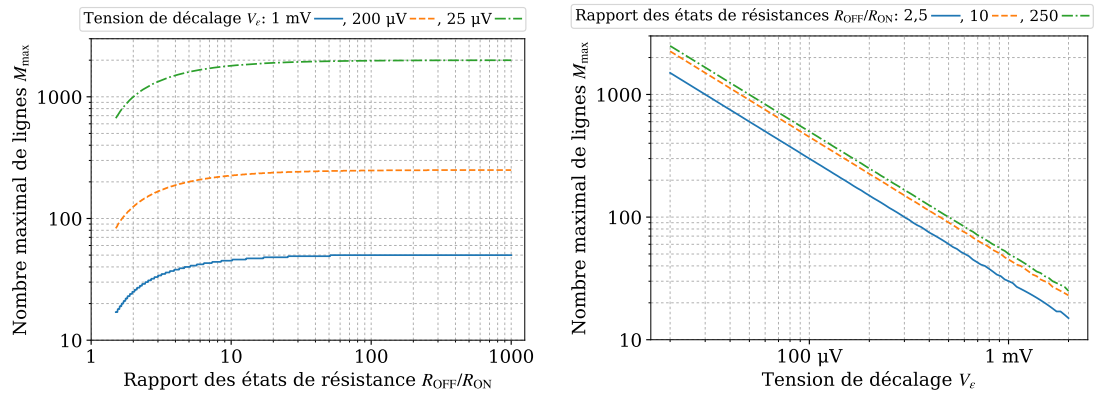
D'après l'équation (5.6), à amplitude V_{read} des impulsions de lecture fixée, le nombre maximal d'entrées présynaptiques $M_{\max}$ dépend de deux paramètres. Tout d'abord de la tension de décalage V_{ε} de l'étage d'entrée du neurone postsynaptique. Étant généralement dans une situation où $V_{\mathrm{read}} \gg V_{\varepsilon}$, la tension de décalage constitue le principal facteur qui limite le nombre

I. Les symboles $\lfloor\rfloor$ désignent l'opérateur partie entière.

d'entrées de la matrice synaptique. L'autre facteur est le rapport $R_{\mathrm{OFF}}/R_{\mathrm{ON}}$ des états de résistance des synapses. Cette dépendance est toutefois moins marquée que la première puisque dans un cas réaliste où $V_\varepsilon/V_{\mathrm{read}} \simeq 0{,}01$ et $R_{\mathrm{OFF}}/R_{\mathrm{ON}} \simeq 2$, diviser par un facteur deux le premier rapport double presque le nombre maximal d'entrées présynaptiques utilisables, tandis que ce dernier n'est même pas doublé lorsque le second rapport est décuplé. Ainsi, le fait pour les jonctions tunnel magnétiques de présenter un rapport des états de résistance $R_{\mathrm{OFF}}/R_{\mathrm{ON}}$ inférieur de plusieurs ordres de grandeur à celui des autres technologies mémristives apparaît moins dramatique qu'au premier abord quant à la limitation du nombre d'entrées d'une matrice synaptique sans sélecteur. Quel que soit le rapport $R_{\mathrm{OFF}}/R_{\mathrm{ON}}$ des technologies concurrentes, ces dernières permettront tout au plus de doubler le nombre d'entrées présynaptiques par rapport à celui d'une matrice de jonctions tunnel magnétiques. Le facteur limitant le plus la dimension d'entrée de la matrice synaptique est en réalité la tension de décalage V_ε de l'étage d'entrée du circuit postsynaptique.

La figure 5.2 présente l'évolution du nombre maximal d'entrées présynaptiques M_{max} donné par l'équation (5.6), selon les deux paramètres susmentionnés que sont le rapport des états de résistance $R_{\mathrm{OFF}}/R_{\mathrm{ON}}$ et la tension de décalage V_ε. L'amplitude V_{read} de l'impulsion de lecture est prise égale à 0,1 V. Les courbes tracées montrent que dans le cadre d'hypothèses raisonnables ($R_{\mathrm{OFF}}/R_{\mathrm{ON}} \simeq 2$) et d'un circuit postsynaptique de faible complexité ($V_\varepsilon \simeq 1\,\mathrm{mV}$), le nombre d'entrées présynaptiques M_{max} envisageable ne dépasse pas quelques dizaines de lignes. Comme analysé précédemment, bénéficier d'un écart relatif entre les états de résistance plus important ne permet pas de plus que doubler ce nombre M_{max}. À l'inverse, une tension de décalage V_ε proche de celle d'un bon amplificateur opérationnel discret (de l'ordre de la centaine de microvolts) permet d'atteindre un nombre M_{max} de l'ordre de la centaine de lignes. Enfin, réduire la tension de décalage V_ε jusqu'à une valeur proche de celle d'amplificateurs d'instrumentation (de l'ordre de la dizaine de microvolts) permet de gagner encore un ordre de grandeur sur la valeur de M_{max}. Ces deux dernières solutions se font néanmoins au détriment de la complexité et de la surface de silicium des circuits postsynaptiques, puisque la réduction de la tension de décalage V_ε nécessite d'avoir recours à des circuits (analogiques) de compensation supplémentaires. Elles ne sont donc intéressantes qu'à condition que l'augmentation du nombre d'entrées présynaptiques permette de compenser la surface et la consommation supplémentaires requises par ces stratégies.

Quel que soit le jeu de valeurs ($R_{\mathrm{OFF}}/R_{\mathrm{ON}}$, V_ε) retenu parmi ceux présentés précédemment, nous constatons qu'aucun ne permet de réellement espérer pouvoir lire (dans le pire des cas) une matrice synaptique comportant $2\times128\times128 = 32\,768$ entrées. Il faut néanmoins remarquer que ces conclusions sont basées sur l'étude du cas le plus défavorable possible (lecture d'une synapse déprimée parmi un fond entier de synapses potentialisées). Les dimensions envisageables en pratiques sont à priori légèrement supérieures, du fait du nombre m_{OFF} non nul d'éléments synaptiques dans leur état de faible résistance, l'équation (5.6) devenant dans ce



(a) (b)

FIGURE 5.2 – Évolution du nombre maximal $M_{\max}$ d'entrées présynaptiques envisageable lors d'une phase de lecture d'une structure 1R sans sélecteur, avec une amplitude V_{read} des impulsions de lecture de 0,1 V. Le panneau *(a)* illustre la dépendance au rapport $R_{\text{OFF}}/R_{\text{ON}}$ des états de résistance synaptiques, tandis que le panneau *(b)* présente celle liée à la tension de décalage V_ε. Les courbes sont calculées à partir de l'équation (5.6).

cas

$$M_{\max} = \left\lfloor \left(\frac{1}{2} \times \frac{V_{\text{read}}}{V_\varepsilon} + (1 + m_{\text{OFF}}) \right) \times \left(1 - \frac{R_{\text{ON}}}{R_{\text{OFF}}} \right) \right\rfloor . \tag{5.7}$$

Dans une situation usuelle, où plus de la moitié des éléments synaptiques est déprimée[I], il est donc envisageable d'utiliser une matrice synaptique disposant d'un nombre d'entrées plus important que celui du pire cas étudié. Par ailleurs, le rôle de cette matrice d'éléments synaptiques n'est pas de constituer une mémoire à accès aléatoire au fonctionnement déterministe mais d'être utilisée au sein d'un système neuromorphique capable d'*apprendre* sa tâche. Nous pouvons espérer que le système complet soit capable de fonctionner correctement malgré le caractère possiblement erroné de certains événements de lecture synaptique, tant que le nombre de ces derniers reste limité.

Coût énergétique de la lecture d'une synapse. Un autre aspect important de la structure matricielle 1R à prendre en compte est sa consommation énergétique durant la phase d'inférence du système. Nous n'examinons ici que la consommation intrinsèque liée aux synapses elles-mêmes. La consommation du circuit de conversion, qui s'y ajoute, n'est pas en général négligeable mais dépend considérablement de l'implémentation choisie.

L'énergie absorbée par un élément synaptique de résistance R_i lors d'une impulsion rectan-

I. Ceci est particulièrement le cas du système étudié : les cartes de conductance obtenues à l'issue de l'apprentissage présentées au chapitre 3 montrent un nombre de synapses potentialisées bien plus faible que celui des synapses déprimées.

gulaire de lecture d'amplitude V_{read} et de durée t_{read} est

$$E_{\text{read}} = \frac{V_{\text{read}}^2}{R_i} \times t_{\text{read}}. \tag{5.8}$$

En considérant le cas d'un composant d'une résistance de l'ordre de 10 kΩ soumis à une impulsion de lecture de durée $t_{\text{read}} = 1\,\text{ns}$ et d'amplitude $V_{\text{read}} = 0{,}1\,\text{V}$, l'énergie absorbée est alors de l'ordre du femtojoule (10^{-15} J).

Malheureusement, il faut également tenir compte de la puissance consommée en continu par les éléments synaptiques non lus mais soumis à la tension de décalage V_ε. Intéressons nous au cas d'un ensemble de $M = 32\,768$ éléments synaptiques, chacun d'une résistance $R_j \simeq 10\,\text{k}\Omega$ et soumis à une tension de décalage V_ε relativement faible de 200 µV. La puissance dissipée de manière statique par ces composants est

$$P_{\text{offset}} = M \times \frac{V_\varepsilon^2}{R_j}, \tag{5.9}$$

dont l'application numérique se situe autour de 0,1 µW.

L'enregistrement présenté en entrée du système comporte environ 5 millions d'événements pour une durée de 82 s. Une estimation haute de la fréquence moyenne des entrées présynaptiques est alors de l'ordre de la centaine de kilohertz, ce qui correspond à une durée moyenne d'environ 10 µs entre deux événements successifs de lecture. Durant ce délai, une colonne de la matrice synaptique dissipe une énergie de l'ordre du *picojoule*, c'est-à-dire mille fois plus que celle E_{read} dissipée lors de la lecture elle-même !

La consommation énergétique associée au fait qu'un système neuromorphique puisse passer la majeure partie de son temps dans un état de repos apparaît donc comme une limitation majeure de la structure matricielle sans sélecteur. Ce découplage entre la dynamique temporelle de la tâche cognitive à réaliser et celle du circuit électronique associé est d'ailleurs l'une des raisons pour lesquelles la puce neuromorphique TrueNorth[13] exploite une technologie de transistors à haute tension de seuil.

5.2.2 Stratégie exploitant un circuit de lecture numérique

Ajout d'éléments de sélection à la matrice synaptique. La partie précédente nous a permis de constater que la circulation de courants au travers des éléments synaptiques non sélectionnés est un problème fondamental de la structure matricielle 1R dès lors que ces derniers sont soumis à une tension parasite non nulle. Une solution naturelle à cela est d'adjoindre un composant de sélection à chacun des éléments synaptiques. La figure 5.3 présente l'exemple d'une matrice de jonctions tunnel magnétiques équipées de transistors de sélection. Cette approche permet réduire de façon drastique les courants traversant les synapses non sélectionnées, ces derniers étant alors les courants de fuite des transistors.

13. P. A. MEROLLA et coll., Science, 2014.

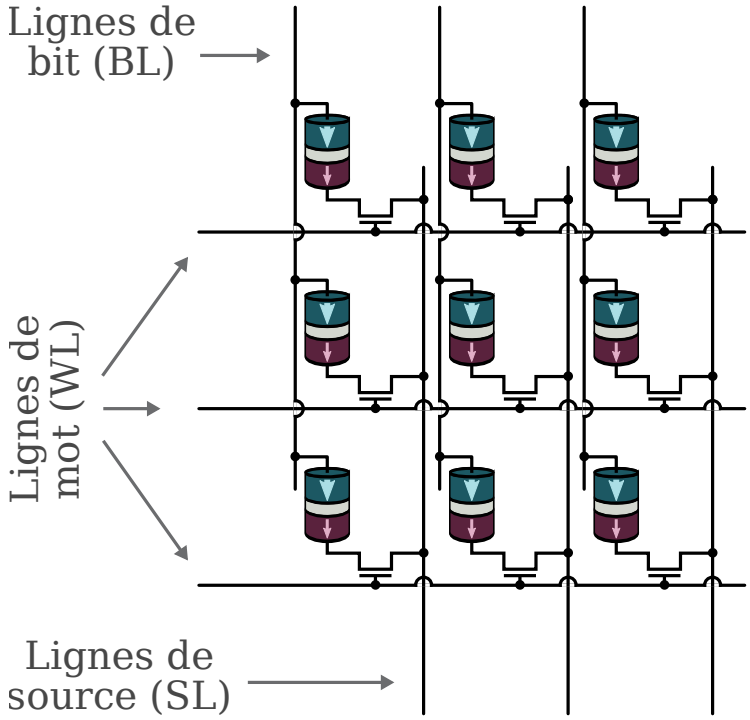


FIGURE 5.3 – Matrice synaptique de jonctions tunnel magnétiques à transfert de spin munies de transistors de sélection (structure 1T-1R). Une entrée présynaptique active sélectionne une ligne de mot.

Conséquences sur la densité d'intégration. Un inconvénient à priori important de l'ajout d'un transistor de sélection est la réduction associée du potentiel d'intégration de la matrice synaptique. La surface occupée par le transistor de sélection, placé à la verticale de la jonction tunnel magnétique qu'il sélectionne, est en effet supérieure à la surface de cette dernière. Dans le cas d'un usage mémoire conventionnel, les dimensions d'une cellule élémentaire actuelle composée d'une jonction et de son transistor d'accès sont typiquement d'environ 30 F^2 (F désignant la dimension caractéristique du nœud technologique), pour tenir compte des problématiques de dispersion technologique et de fiabilité mémoire[269].

Intéressons-nous au cas de jonctions tunnel magnétiques à transfert de spin de 32 nm dont l'amplitude du courant critique est de l'ordre de 30 µA. D'après le kit de conception de procédés à notre disposition, des transistors NMOS et PMOS de respectivement 30 nm et 50 nm de largeur suffisent à assurer le passage d'un tel courant. Il est à noter que ceci est cohérent avec les prévisions de la littérature[270,271] suggérant la possibilité d'utiliser des cellules mémoires élémentaires de 6 F^2 pour les nœuds technologiques inférieurs à 45 nm. Sachant qu'une cellule mémoire élémentaire *sans* sélecteur occupe une surface de 4 F^2, la densité d'intégration n'est jamais réduite que de 50 % dans le cas d'une cellule avec sélecteur de 6 F^2. Il est à noter toutefois que l'uti-

269. R. DORRANCE et coll., IEEE Transactions on Electron Devices, 2012.
270. S. A. WOLF et coll., Proceedings of the IEEE, 2010.
271. U. K. KLOSTERMANN et coll., IEEE International Electron Devices Meeting, 2007.

lisation de transistors de sélection empêche de superposer une telle matrice synaptique sur une couche de circuits neuronaux CMOS : seule une approche de connexion planaire est alors possible.

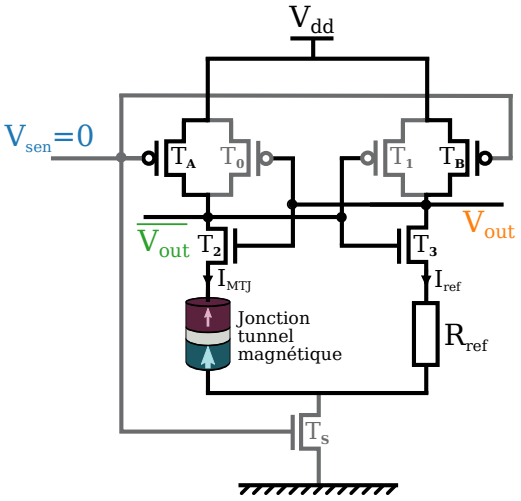
Néanmoins, quand bien même la surface d'une cellule élémentaire à considérer en pratique serait de 30 F^2 ou davantage, la bonne question à résoudre est en réalité de savoir si cet inconvénient est compensé en termes de fonctionnalités à l'échelle du système : phase de lecture ou d'écriture plus simples ou plus économes en énergie, matrice synaptique de grandes dimensions envisageable, etc.

De nouvelles possibilités de circuits de lecture. Le caractère binaire des jonctions tunnel magnétiques utilisées comme synapses rend la lecture différente de celle d'autres composants mémristifs à plus de deux niveaux de résistance. Au sein d'une colonne de synapses toutes connectées au même circuit postsynaptique, il s'agit moins de lire précisément la valeur du courant pondérée par la conductance de l'*unique* synapse sélectionnée que d'être capable d'en classer l'état de résistance entre « haute » et « faible ». Cette opération est à rapprocher de la lecture d'un bit mémoire d'une cellule conventionnelle à deux états.

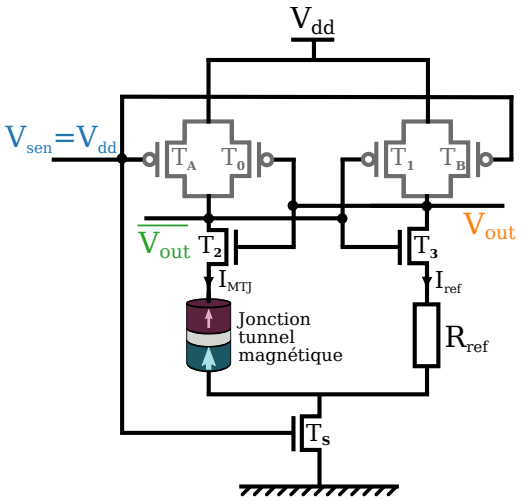
Dans cette partie, nous présentons l'utilisation d'un circuit de lecture conçu à l'origine pour des puces de mémoire non volatile conventionnelle. Il s'agit du circuit de *Pre-Charge Sense Amplifier*[190,267] (PCSA). La constitution classique de ce circuit est donnée à la figure 5.4, qui en détaille par ailleurs les grandes phases de fonctionnement. Schématiquement, ce circuit est similaire à une (partie de) cellule de mémoire vive statique (SRAM), faite des transistors T_0 à T_4, que l'on place de manière forcée dans un état initial métastable au moyen des transistors T_A et T_B, avant de la laisser basculer vers un état final stable qui dépend de la valeur de la résistance de la jonction tunnel magnétique relativement à celle de la résistance de référence placée dans l'autre branche. Ces deux phases de fonctionnement sont détaillées ci-après.

1. Dans un premier temps, le signal V_{sen} est au repos (figure 5.4a), ce qui bloque le transistor T_S et rend passant les transistors T_A et T_B. Cette configuration a pour effet de bloquer les transistors T_0 et T_1, tout en rendant passants T_2 et T_3. Durant cette phase, les signaux de sortie V_{out} et $\overline{V_{out}}$ sont tous les deux dans leur état haut.
2. Dans un second temps, le signal V_{sen} est activé, ayant pour effet de bloquer les transistors T_A et T_B (figure 5.4b). Le transistor T_S étant quant à lui rendu passant, du courant peut circuler à travers la jonction tunnel magnétique ainsi qu'au travers de la résistance de référence R_{ref}. Ces courants sont le fruit des charges électriques accumulées sur les grilles des transistors lors de la phase précédente. Si par exemple la résistance de la jonction tunnel magnétique est supérieure à la valeur de référence R_{ref}, la tension V_{out} diminue plus rapidement que $\overline{V_{out}}$. Le transistor T_0 est alors le premier des transistors PMOS à devenir passant, faisant alors remonter $\overline{V_{out}}$ vers son niveau actif, tandis que V_{out} achève de décroître jusqu'à son niveau bas, bloquant le transistor T_2 (figure 5.4c). Une situation

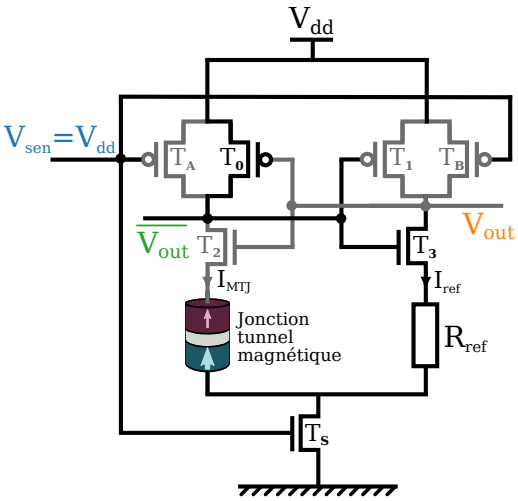
190. W. S. ZHAO et coll., IEEE Transactions on Magnetics, 2009.



(a) Phase de charge anticipée.



(b) Début de la phase de mesure.



(c) Fin de la phase de mesure.

FIGURE 5.4 – Circuit et fonctionnement schématique d'un circuit de type *Pre-Charge Sense Amplifier*, utilisé pour déterminer la configuration d'une jonction tunnel magnétique. Lorsque le signal V_{sen} est inactif, le circuit est dans un état métastable présenté en *(a)*. Les illustrations *(b)* et *(c)* décrivent les situations respectivement initiale et finale de la phase de mesure, durant laquelle le signal V_{sen} est actif, dans le cas d'une jonction tunnel magnétique de résistance supérieure à la référence R_{ref} placée dans l'autre branche. *NB* : les transistors dessinés en gris sont dans leur état bloqué. *Source* : adapté d'une figure de N. LOCATELLI.

opposée ($\overline{V_{\text{out}}}$ dans son état bas et V_{out} dans son état haut) est obtenue de manière analogue dans le cas où la résistance de la jonction tunnel magnétique est inférieure à celle la valeur de référence R_{ref}.

Si la résistance de référence R_{ref} est d'une valeur comprise entre les deux valeurs de résistance possibles des jonctions tunnel magnétiques employées, ce circuit permet ainsi de lire l'*état* de résistance de ces composants.

Dans le cas de la structure 1T-1R de la figure 5.3, un circuit de type *Pre-Charge Sense Amplifier* est utilisé par colonne de synapses reliées au même circuit postsynaptique, réparti de part et d'autre de ladite colonne. Les entrées présynaptiques sont présentées sur les grilles des transistors de sélection, et déclenchent en outre le signal de lecture V_{sen}. La sortie V_{out} constitue quant à elle le signal transmis au circuit neuronal, à priori numérique. Comme nous l'avons vu précédemment, ce signal est un « un » logique lorsque la jonction tunnel magnétique est dans son état de faible résistance (c'est-à-dire potentialisée) et un « zéro » logique lorsque la synapse est au contraire dans son état de haute résistance (c'est-à-dire déprimée).

Avantages et inconvénients de cette stratégie de lecture. L'un des premiers avantages de ce système est d'effectuer directement la conversion de la lecture en un signal numérique, ce qui permet au signal V_{out} d'être moins affecté par la variabilité synaptique ou le bruit électronique. Par ailleurs, le fait que le signal V_{out} soit numérique permet de masquer aux étages ultérieurs la faible valeur du rapport des états de résistance $R_{\text{OFF}}/R_{\text{ON}}$, sans nécessiter le recours à un étage d'amplification supplémentaire.

Dans le but d'évaluer la consommation énergétique liée à la lecture d'une synapse par un circuit de type *Pre-Charge Sense Amplifier*, simulons un tel circuit sous le logiciel Spectre Cadence Design Systems® (figure 5.5). L'évolution des signaux V_{sen}, V_{out} et $\overline{V_{\text{out}}}$, est tracée à la figure 5.6, ainsi que celle de la puissance P_{tot} fournie par l'alimentation. Nous pouvons observer la décroissance des signaux V_{out} et $\overline{V_{\text{out}}}$ lorsque la lecture est déclenchée par le passage du signal V_{sen} à son état haut à $t = 1\,\text{ns}$. Après environ 300 ps, le signal $\overline{V_{\text{out}}}$ remonte vers son état haut, tandis que V_{out} poursuit sa chute, indiquant que la synapse lue est déprimée. Les traces de la figure 5.6 montrent qu'il est envisageable de lire l'état de la jonction tunnel magnétique en moins d'une nanoseconde, même lorsque le circuit débite sur des charges capacitives de valeurs non négligeables.

Par ailleurs, l'évolution du signal P_{tot} montre que lors de la phase de lecture, l'amplitude de la puissance absorbée par l'ensemble du circuit de lecture est du même ordre de grandeur (la dizaine de microwatts) que celle dissipée par la synapse dans le cas d'une lecture analogique. Néanmoins, un avantage du circuit de lecture numérique est qu'il consomme uniquement lors des phases dynamiques. En régime établi, chacune des branches du circuit comporte un transistor dans son état bloqué, limitant alors la puissance consommée aux fuites des transistors, bien inférieure à celle liée à la tension de décalage non nulle dans le scénario analogique de la partie 5.2.1. Ainsi, l'*ensemble* du circuit de lecture de type *Pre-Charge Sens Amplifier* de la figu-

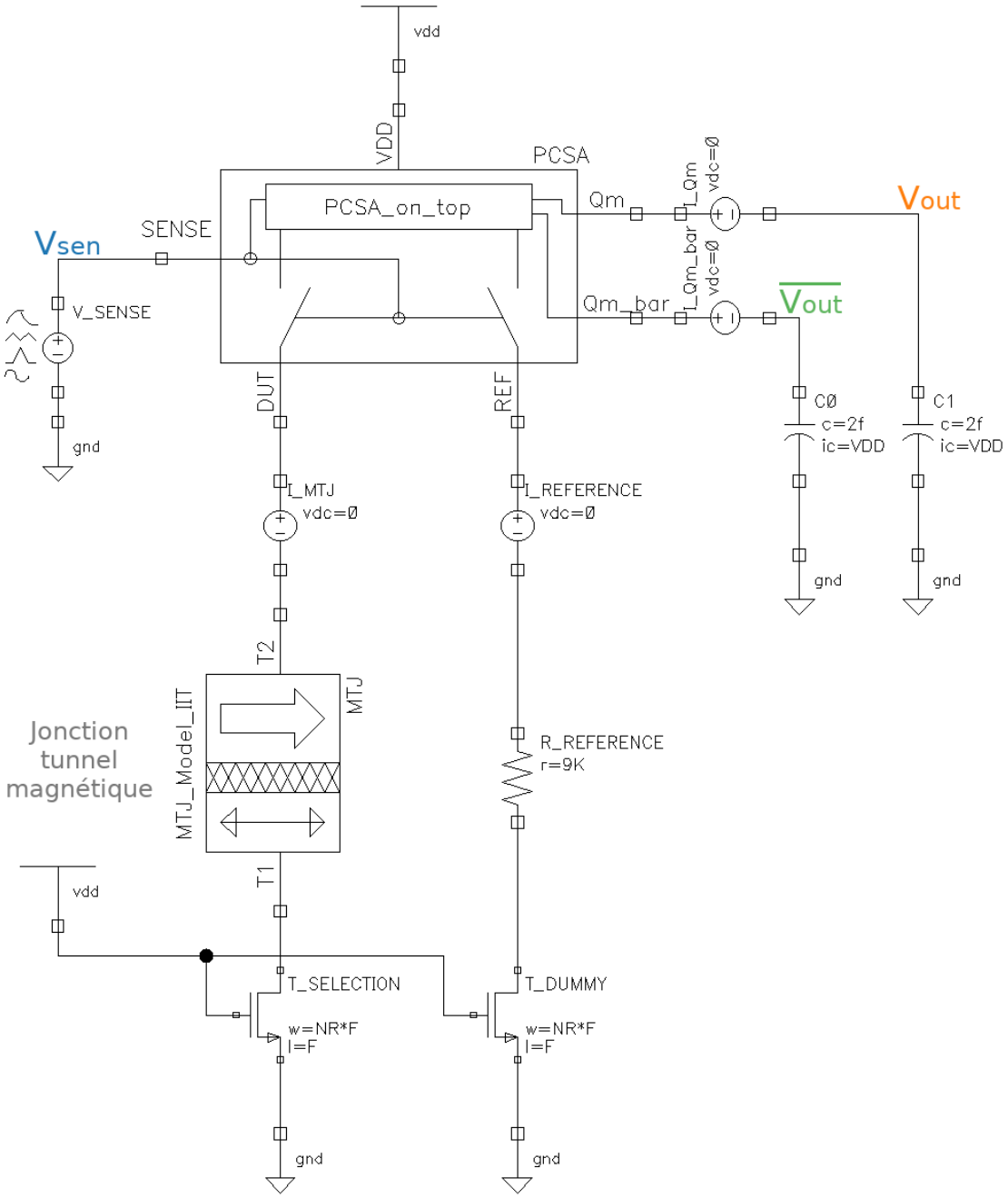


FIGURE 5.5 – Circuit de type *Pre-Charge Sense Amplifier* simulé sous Spectre®. Les signaux V_{sen}, V_{out} et $\overline{V_{\text{out}}}$ sont similaires à ceux de la figure 5.4. La jonction tunnel magnétique à transfert de spin est décrite par le modèle Verilog-A adapté par Nicolas LOCATELLI à partir des équations du modèle analytique du chapitre 2. La largeur de grille L des transistors employés est $F = 30\,\text{nm}$. La largeur W des transistors NMOS est de 30 nm, tandis que celle des transistors PMOS vaut 60 nm. La tension d'alimentation V_{dd} est de 1,2 V. Les sorties V_{out} et $\overline{V_{\text{out}}}$ débitent sur des condensateurs de 2 fF et la résistance de référence présente une valeur de 9 kΩ, située entre les deux valeurs de résistance de la jonction tunnel magnétique. *NB* : le bloc « PCSA_on_top » comporte une structure proche de la partie supérieure des illustrations de la figure 5.4.

re 5.5 consomme uns énergie $E_{\text{read}} = 5{,}2\,\text{fJ}$ lors de la lecture d'une jonction tunnel magnétique dans son état antiparallèle (figure 5.6), et 5,4 fJ dans le cas d'un état parallèle (non représenté).

De ces nombres, il est possible d'obtenir une estimation de la puissance moyenne $P_{\text{col}}^{(\text{inf})}$ dissipée lors de la phase d'inférence par une colonne de synapses connectées au même neurone de sortie postsynaptique et le circuit de lecture numérique associé. En considérant que chaque événement de lecture consomme 10 fJ et sachant qu'il y a environ 5 millions d'impulsions présynaptiques durant les 82 s de l'enregistrement utilisé, nous obtenons une puissance moyenne $P_{\text{col}}^{(\text{inf})}$ inférieure au *nanowatt*. Pour rappel, l'estimation similaire de $P_{\text{col}}^{(\text{inf})}$ dans le cas analogique étudié précédemment était de l'ordre de la *centaine* de nanowatts, sachant que ce nombre n'incluait pas la consommation du circuit de conversion à transimpédance. En termes de consommation énergétique lors de la phase d'inférence, un circuit numérique de type *Pre-Charge Sense Amplifier* associé à une structure matricielle pourvue de transistors de sélection semble donc constituer un meilleur choix qu'une approche analogique.

En termes surfaciques, l'utilisation d'un circuit numérique de type *Pre-Charge Sense Amplifier* semble extrêmement avantageux, puisqu'il requiert uniquement des transistors de dimensions minimales (figure 5.5). En comparaison, du fait de son fonctionnement analogique, un circuit de conversion à transimpédance nécessite des transistors sensiblement plus larges (pour la plupart dix à vingt fois) et dont la quantité peut devenir significative selon la complexité du circuit de réduction de la tension de décalage V_ε employé. Néanmoins, la topologie du *Pre-Charge Sense Amplifier* se rapproche de celle d'une cellule de mémoire vive statique, dont la surface minimale est principalement limitée par les interconnexions métalliques nécessaires. La surface de silicium exigée par ce circuit de lecture devrait donc être supérieure à la centaine de F^2. Par ailleurs, rappelons que la densité d'intégration de la matrice synaptique est également dégradée en raison de la nécessité d'ajouter des transistors de sélection. Il n'est pas évident que cette stratégie de lecture soit réellement avantageuse en termes de coût surfacique par rapport à la stratégie analogique, notamment selon la tension de décalage V_ε tolérée pour cette dernière, qui détermine la complexité du circuit de lecture analogique à employer.

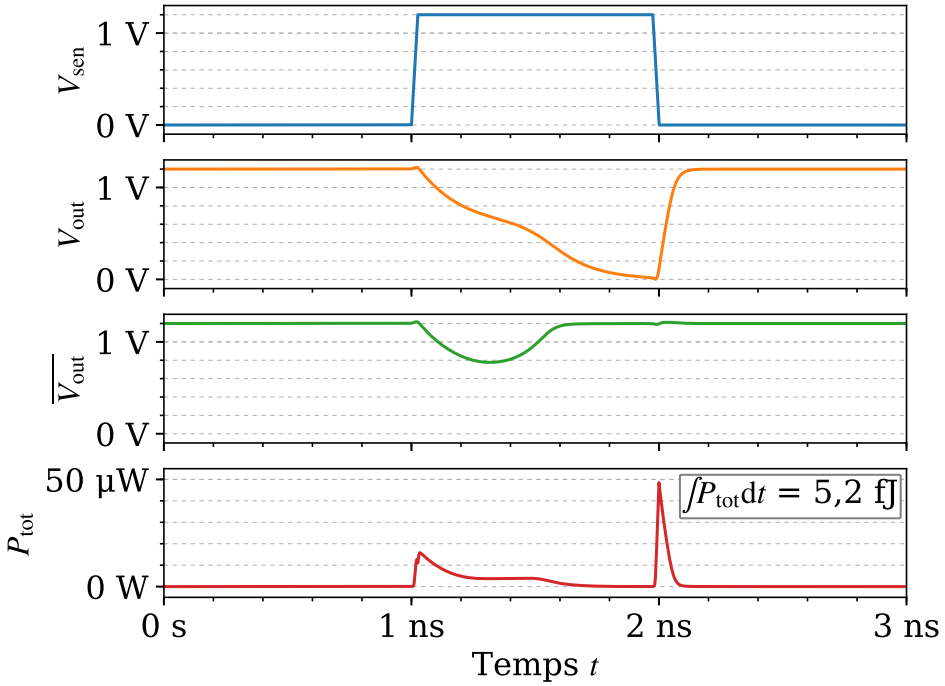


FIGURE 5.6 – Évolution des signaux issus de la simulation du circuit de la figure 5.5. La lecture est déclenchée lorsque le signal V_{sen} passe à l'état haut. Les signaux V_{out} et $\overline{V_{out}}$ sont les deux sorties du circuit de lecture, de valeurs complémentaires à l'issue de la phase de lecture. Enfin, l'évolution de la puissance P_{tot} délivrée par l'alimentation de 1,2 V est également tracée. L'annotation placée sur ce dernier graphe indique l'énergie consommée à l'issue des 3 ns de simulation, obtenue par intégration numérique de P_{tot}.

5.3 Phase d'écriture

Dans cette partie, nous nous intéressons à la mise en œuvre à l'échelle du circuit, des phases d'écriture permettant de réaliser l'apprentissage du système. Nous y comparons la programmation parallèle d'une matrice synaptique sans composants de sélection, telle que nous l'avions imaginée dans le chapitre 3, à une approche par programmation séquentielle des lignes d'une matrice de synapses munies de transistors de sélection.

Afin de simplifier l'étude, nous considérons une unique valeur de l'ordre de la dizaine de microampères pour le courant I_{prog} nécessaire à la programmation d'une jonction tunnel magnétique, sans distinction entre les deux scénarios de commutation possibles. Cette valeur numérique constitue une estimation optimiste, liée aux dernières réalisations technologiques publiées[187].

5.3.1 Cellules d'écriture

La figure 5.7 illustre les deux situations que nous envisageons, selon la structure matricielle considérée.

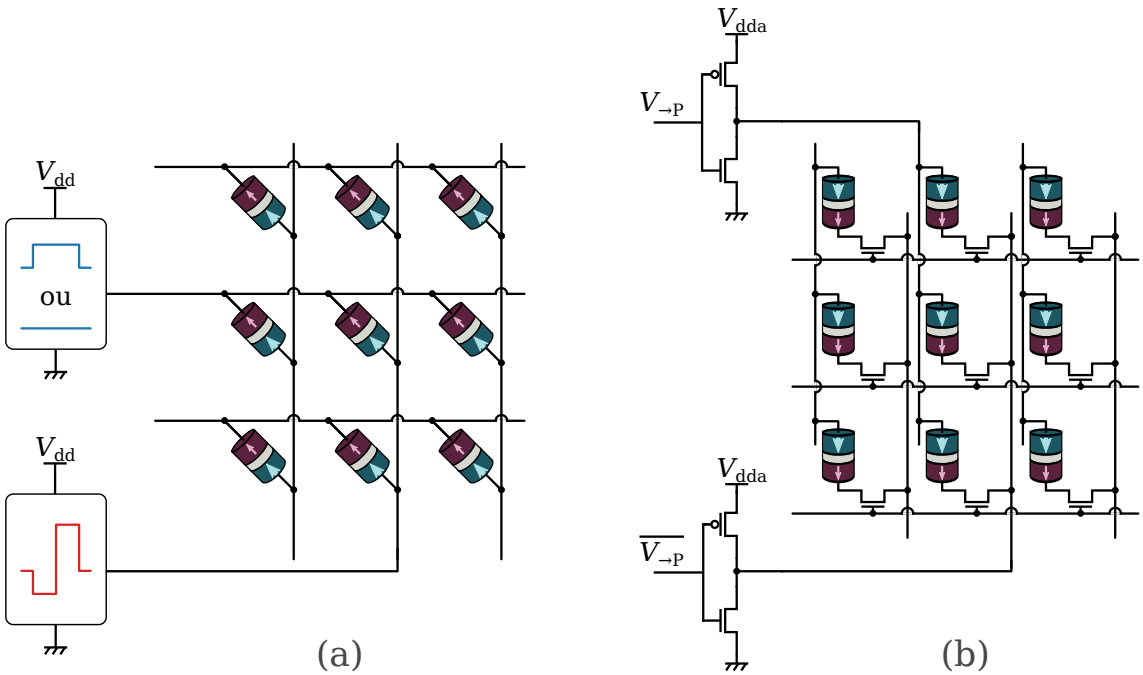


FIGURE 5.7 – Cellules d'écriture. *(a)* Cas d'une matrice de synapses sans sélecteurs (structure 1R). Les blocs sur la gauche appliquent les tensions présentées au chapitre 3. *(b)* Pont en H pour une matrice de synapses avec transistor de sélection (structure 1T-1R). Le signal $V_{\rightarrow P}$ est à l'état haut lorsque la jonction tunnel magnétique à transfert de spin considérée doit être programmée vers son état parallèle (P).

Dans le cas d'une configuration sans sélecteur (structure 1R), chaque cellule d'écriture est

187. J. J. NOWAK et coll., IEEE Magnetics Letters, 2016.

constituée de deux parties distinctes :

- un circuit postsynaptique, déclenché par l'activation du neurone postsynaptique associé ;
- un circuit présynaptique qui applique un potentiel électrique dont la forme d'onde dépend de l'activité récente du neurone présynaptique associé.

Les potentiels appliqués par ces circuits sont ceux présentés dans le chapitre 3.

Dans le cas où chaque synapse de la matrice est associée à un transistor de sélection (structure 1T-1R), la cellule d'écriture est un pont en H, dont chacun des deux bras est constitué d'un transistor PMOS et d'un transistor NMOS, afin de faire circuler un courant de la polarité désirée à travers la jonction tunnel magnétique à transfert de spin à programmer. Les bras d'écriture sont placés de part et d'autre de chaque colonne. Dans ce second schéma, la ligne de la jonction tunnel magnétique à programmer doit alors également être activée lors de la programmation.

5.3.2 Programmation parallèle d'une structure sans transistors de sélection

Cette approche est la mise en œuvre naturelle des idées des simulations fonctionnelles du chapitre 3 : lors du déclenchement d'une phase d'écriture par l'un des neurones de sortie, toutes les synapses de la colonne associée sont programmées en parallèle.

Les deux parties successives d'un cycle d'écriture tel qu'imaginé au chapitre 3 sont présentées à la figure 5.8. Les courants circulant dans la matrice synaptique peuvent être décomposés en trois composantes :

- *une partie utile*, composée de la somme des courants circulant à travers les synapses entièrement sélectionnées de la colonne active, c'est-à-dire soumise à une tension V_{prog} à leurs bornes ;
- *une partie inutile*, somme des courants traversant les synapses qui sont seulement à moitié sélectionnées (tension $V_{\text{prog}}/2$ à leurs bornes) de la colonne active ; ces courants ne visent pas à faire commuter les synapses mais sont présents du fait des formes d'onde de tension employées ;
- *des courants parasites de fuites*, qui circulent à travers les synapses reliant les colonnes inactives aux lignes sur lesquelles une tension $V_{\text{prog}}/2$ est appliquée.

Dimensionnement des transistors du côté présynaptique. La situation de la ligne du milieu de la figure 5.8a correspond au cas dans lequel le courant délivré par le circuit présynaptique associé est maximal. En effet, le courant I_{prog} circulant à travers la synapse sélectionnée (première colonne, en trait gras violet) s'additionne aux $N-1$ courants parasites de fuites $I_{\text{prog}}/2$ circulant dans les autres synapses à moitié sélectionnées de la ligne. Le courant présynaptique à délivrer est ainsi $(N+1) \times I_{\text{prog}}/2$ et croît de manière quasiment *proportionnelle au nombre N de neurones de sortie* en raison des courant parasites de fuites. Avec 20 colonnes de sortie (comme au chapitre 3) et les valeurs du kit de conception de procédés données dans la partie 5.1, ce courant est de l'ordre de la centaine de microampères, ce qui requiert une largeur d'environ

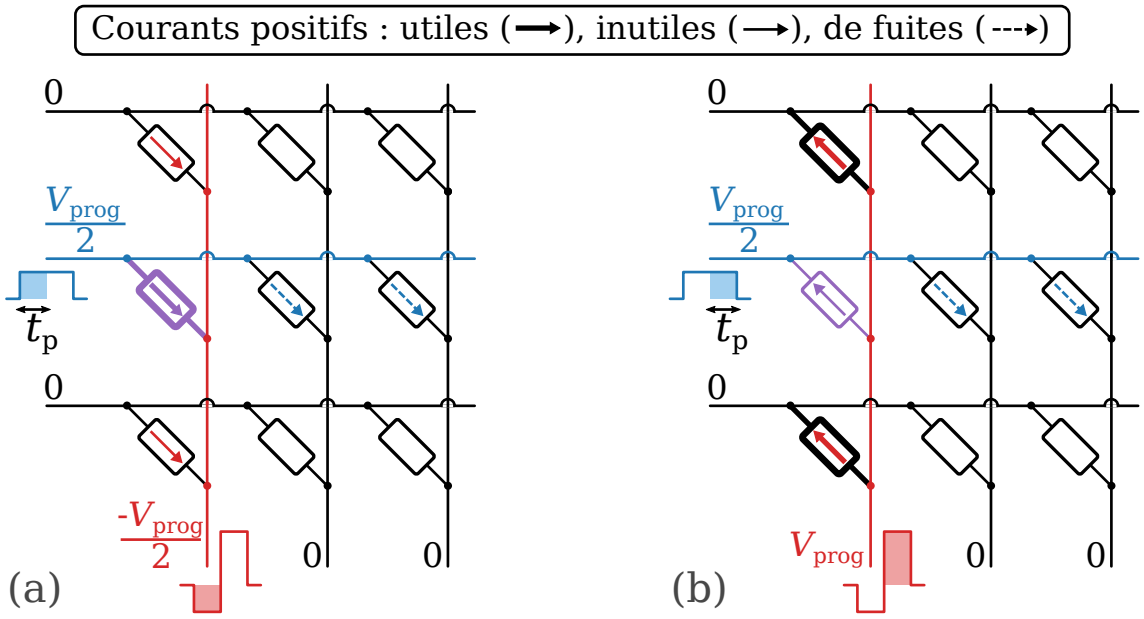


FIGURE 5.8 – Les deux phase d'un cycle d'écriture durant lequel l'ensemble des synapses reliées au même circuit postsynaptique est programmé de manière parallèle. Le neurone postsynaptique venant de se déclencher est associé à la première colonne de gauche. Les symboles le long des fils désignent leur potentiel électrique et les aires colorées des formes d'onde indiquent la portion utilisée des impulsions appliquées.
(a) La première moitié t_p du cycle d'écriture permet de sélectionner entièrement les synapses telles que celle représentée en trait gras violet, les tensions pré- et postsynaptiques se combinant pour appliquer une tension V_{prog} à leurs bornes. Il s'agit de la composante utile des courants circulant durant cette première phase (flèche en trait gras). Les autres synapses de la colonne sont quant à elles sélectionnées seulement à moitié (tension $V_{prog}/2$ à leurs bornes), faisant circuler une composante de courant inutile (flèches en trait fin). Enfin, les synapses de lignes activées mais appartenant aux autres colonnes de la matrice sont également à moitié sélectionnées, ce qui entraîne la circulant de courants parasites de fuites (flèches en pointillé) à travers les jonctions tunnel magnétiques en question.
(b) Durant la seconde moitié t_p du cycle d'écriture, la superposition des tensions présynaptiques à la tension postsynaptique appliquée par le neurone de sortie activé inverse la sélection des synapses de ladite colonne, ainsi que le caractère utile ou inutile des courants impliqués. Les courants parasites de fuites restent quant à eux identiques.

100 nm pour les transistors NMOS et 170 nm pour les transistors PMOS.

Un circuit présynaptique d'écriture étant utilisé par entrée, il est par ailleurs à noter que la surface occupée par les transistors de l'ensemble de ces circuits devient conséquente dans l'hypothèse d'un système possédant un grand nombre d'entrées présynaptiques.

Dimensionnement des transistors du côté postsynaptique. Le circuit postsynaptique d'écriture d'une colonne doit pouvoir délivrer un courant constitué de la somme des courants $\pm I_{\text{prog}}$ circulant à travers les synapses entièrement sélectionnées de la colonne, et des courants $\pm I_{\text{prog}}/2$ traversant les autres synapses sélectionnées seulement à moitié[I]. Durant un cycle complet d'écriture, chaque synapse est entièrement sélectionnée la moitié du temps et à moitié sélectionnée l'autre moitié. Ainsi, le scénario qui requiert le plus faible courant postsynaptique est celui de l'activation en écriture de la moitié des M entrées présynaptiques, qui a pour expression $M \times (5/4) \times I_{\text{prog}}$. Avec 32 768 entrées présynaptiques (comme au chapitre 3), l'application numérique donne une valeur d'environ 400 mA ! Les transistors postsynaptiques NMOS et PMOS nécessaires à ce scénario *optimal* devraient être respectivement d'une largeur de 400 µm et 650 µm.

En conclusion, la stratégie d'une programmation entièrement parallèle semble peu réaliste dans le cas d'une matrice synaptique dont l'une des dimensions devient importante.

5.3.3 Programmation séquentielle

Dans le cas d'une matrice synaptique incluant des transistors de sélection (structure 1T-1R), telle que celle utilisée pour la lecture par des circuits numériques de type *Pre-Charge Sense Amplifier*, des schémas de programmation différents de l'approche parallèle sont possibles. La lecture en question ayant recours à un décodeur de lignes pour délivrer les impulsions de lecture, il est envisageable que l'activation d'un neurone postsynaptique déclenche un balayage séquentiel des entrées présynaptiques, ligne à ligne, durant chaque étape duquel les tensions adéquates sont appliquées sur les deux ports de l'*unique* jonction tunnel magnétique à transfert de spin sélectionnée.

Nécessité d'un découplage entre la vitesse d'écriture d'une synapse et les temps caractéristiques de la tâche à résoudre. La stratégie d'une écriture séquentielle n'est réalisable qu'à la condition que l'ensemble des entrées présynaptiques puissent être balayées pendant la durée minimale entre la période réfractaire du neurone de sortie venant de se déclencher et les périodes d'inhibition des autres neurones de sortie, qui dépendent toutes de la tâche cognitive à réaliser. Dans les simulations présentées au chapitre 3, la durée minimale en question est de l'ordre de 10 ms. Avec 32 768 entrées présynaptiques à balayer, chaque programmation synaptique doit alors être effectuée en moins de 300 ns. Cette approche ne peut donc pas être utilisée

I. Le cas de la circulation de la somme des courants parasites de fuites est écarté puisque cette somme est nécessairement plus faible que la valeur du courant postsynaptique lorsque la colonne est activée.

dans le régime de programmation à faible courant des jonctions tunnel magnétiques à transfert de spin, puisque la durée des impulsions de programmation employées dans cette situation est de l'ordre de la microseconde ou plus. En revanche, cette stratégie est envisageable lorsque les jonctions tunnel magnétiques sont programmées dans leur régime de courant intermédiaire ou fort, puisque la durée des impulsions requises est alors de l'ordre de la dizaine de nanosecondes ou moins (voir partie 3.1.3.2).

Quid du dimensionnement des transistors d'écriture ? Du fait de la présence des transistors de sélection, qui permettent de sélectionner *une à une* les synapses à programmer, les transistors des bras d'écriture (figure 5.7b) doivent seulement être capables de faire circuler le courant de programmation I_{prog} d'une seule jonction tunnel magnétique à transfert de spin, ce qui simplifie le dimensionnement des transistors en question. D'après le kit de conception de procédés à notre disposition, avec une tension d'alimentation $V_{\text{dd}} = 1{,}2\,\text{V}$, des transistors de dimensions minimales (de 30 nm de largeur pour les NMOS et 60 nm pour les PMOS) seraient suffisants pour transmettre un tel courant. Les transistors de ces bras d'écriture sont donc de dimensions biens inférieures à celles des transistors des circuits d'écriture d'une stratégie de programmation parallèle.

Comparaison du nombre de circuits d'écriture. La stratégie d'écriture parallèle de la partie 5.3.2 nécessite un circuit (présynaptique) d'écriture par entrée et un circuit (postsynaptique) d'écriture par sortie (figure 5.7a). La stratégie d'écriture séquentielle, qui utilise quant à elle une matrice synaptique avec des transistors de sélection, requiert un nombre de bras d'écriture double du nombre de sorties postsynaptiques (figure 5.7b). Dans le cas du scénario envisagé au chapitre 3, qui comporte 32 768 entrées présynaptiques pour seulement 20 sorties postsynaptiques, la seconde approche est assurément plus intéressante en terme de quantité de surface de silicium occupée par les circuits d'écriture.

Influence sur la consommation électrique. L'utilisation des transistors de sélection permet de supprimer les éventuels courants parasites de fuites présents dans le cas de la programmation parallèle (flèches en pointillé sur la figure 5.8). Ceci constitue une économie d'énergie importante dans le cas d'une matrice comportant un grand nombre de sorties postsynaptiques, puisque la contribution totale de ces courants parasites de fuites devenait alors proportionnelle au nombre de sortie.

Par ailleurs, le fait de programmer les synapses une par une et non plus parallèlement permet d'éviter le recours à une combinaison d'impulsions de demi-sélection. Selon le type de commutation désirée, il ne s'agit alors plus que d'appliquer durant une durée t_{p} la tension V_{prog} ou $-V_{\text{prog}}$ sur la jonction tunnel magnétique programmée. Dans ce cas, l'énergie dissipée par la

programmation d'une synapse est réduite de 20 % puisqu'elle devient

$$t_\mathrm{p} \times \frac{V_\mathrm{prog}^2}{R}$$

alors qu'elle était de

$$\underbrace{t_\mathrm{p} \times \frac{V_\mathrm{prog}^2}{R}}_{\text{Utile}} + \underbrace{t_\mathrm{p} \times \frac{\left(\frac{V_\mathrm{prog}}{2}\right)^2}{R}}_{\text{Inutile}}$$

dans le cas de la programmation parallèle[I].

5.4 Chutes résistives et effets capacitifs

Une autre famille d'effets pouvant limiter les dimensions de la matrice synaptique est celle des chutes résistives et des effets capacitifs, liés aux pistes métalliques d'accès aux cellules mémoires[272–274]. Pour une piste donnée, la synapse la plus exposée à ces problèmes est celle qui est physiquement la plus éloignée du circuit de commande (*driver*) de ligne ou de colonne.

Dans le but de procéder à quelques applications numériques, considérons une résistance linéique $\hat{r} = 0{,}4\,\Omega{\cdot}\mu\mathrm{m}^{-1}$ et une capacité linéique $\hat{c} = 0{,}2\,\mathrm{fF}{\cdot}\mu\mathrm{m}^{-1}$. Il s'agit de valeurs numériques qu'il est possible de rencontrer pour les niveaux d'interconnexions métalliques intermédiaires d'un nœud technologique récent[79].

5.4.1 Chute résistive le long d'une ligne de synapses

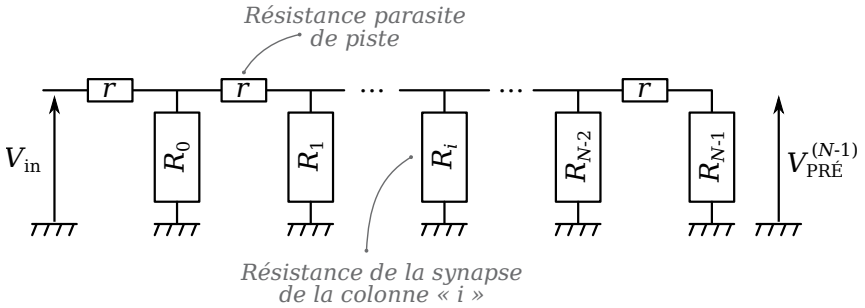


FIGURE 5.9 – Schéma d'une ligne d'éléments synaptiques incluant les résistances parasites des interconnexions métalliques : la tension $V_\mathrm{PRÉ}^{(N-1)}$ est plus faible que celle V_in appliquée à l'entrée de la ligne.

Situation considérée : le cas d'une matrice synaptique sans sélecteurs. Soit une ligne de N éléments synaptiques d'un matrice sans composants de sélection (structure 1R), telle que

I. R désigne la résistance électrique de la jonction tunnel magnétique programmée.

272. S. YU et coll., IEEE International Electron Devices Meeting, 2015.
273. S. ZULOAGA et coll., IEEE International Symposium on Circuits and Systems, 2015.
274. S. B. ERYILMAZ et coll., IEEE International Electron Devices Meeting, 2015.
79. T. GOKMEN et Y. VLASOV, Frontiers in Neuroscience, 2016.

représentée à la figure 5.9. Entre chaque colonne, la piste présente une résistance parasite r. Un calcul itératif à partir de la dernière colonne $(N-1)$ permet d'exprimer la tension appliquée sur le dernier élément synaptique comme

$$V_{\text{PRÉ}}^{(N-1)} = V_{\text{in}} \times \prod_{i=0}^{N-1} \frac{R_{\text{eq}}^{(i)}}{R_{\text{eq}}^{(i)} + r}, \tag{5.10}$$

où V_{in}est la tension appliquée à l'entrée de la ligne et $R_{\text{eq}}^{(i)}$ désigne la résistance équivalente à l'ensemble des résistances situées à droite de la synapse de la colonne $i-1$ sur la figure 5.9, à savoir

$$R_{\text{eq}}^{(i)} = \frac{R_i \times (r + R_{\text{eq}}^{(i+1)})}{R_i + r + R_{\text{eq}}^{(i+1)}} \tag{5.11}$$

La figure 5.10 présente la chute subie par la tension $V_{\text{PRÉ}}^{(N-1)}$ par rapport à la valeur d'entrée V_{in}, dans le pire cas : une ligne de synapses dans leur état de faible résistance R_{ON} (d'une valeur de 6 kΩ).

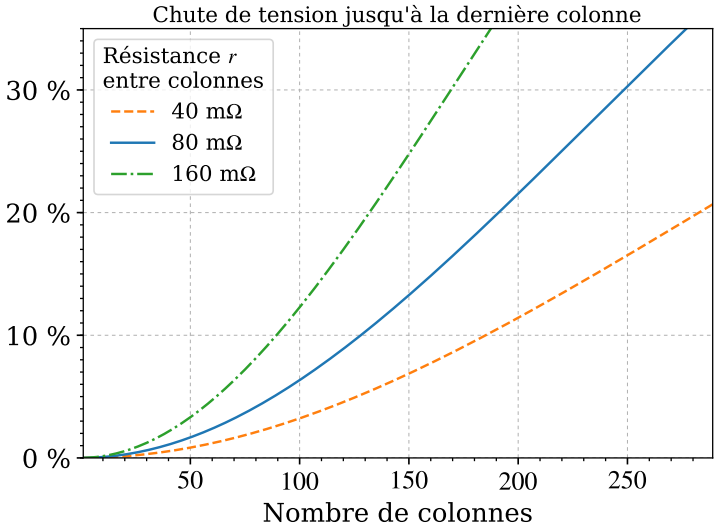


FIGURE 5.10 – Chute subie de la part de la tension $V_{\text{PRÉ}}^{(N-1)}$ par rapport à la valeur d'entrée V_{in} du schéma de la figure 5.9 dans une matrice synaptique sans sélecteurs, calculée selon l'équation (5.10). Les valeurs de la résistance parasite r correspondent, dans l'ordre croissant, à des longueurs de piste de 100 nm (pointillé orange), 200 nm (trait plein bleu) et 400 nm (ligne brisée verte). Les N synapses de la ligne sont toutes dans leur état de faible résistance $R_{\text{ON}} = 6\,\text{k}\Omega$.

Conséquences des chutes résistives. La chute de tension sur les synapses éloignées des circuits de commande des pistes se traduit par une diminution similaire du courant traversant la synapse. Ceci n'est pas réellement critique lors des phases de lecture : les jonctions tunnel magnétiques étant utilisées de façon binaire, leurs états peuvent toujours être distingués l'un de

l'autre si la chute de tension reste mesurée (par exemple 10 %). En revanche, les conséquences d'une chute de tension sont bien plus significatives dans le cadre des phases d'écriture. Comme nous l'avons vu au chapitre 2, selon le régime de programmation utilisé et la probabilité de commutation visée, une variation du courant de programmation, même limitée à 10 % par exemple, peut modifier de façon bien plus importante la probabilité de commutation de la synapse. Même si les résultats du chapitre 3 montrent que le système complet est assez tolérant à des variations de ces probabilités de commutation, il semble néanmoins difficile d'espérer pouvoir utiliser une matrice synaptique sans sélecteurs qui présente plus de quelques centaines de sorties.

Gain apporté par l'utilisation de transistors de sélection. Dans le cas d'une matrice synaptique dont chaque élément est associé à un transistor de sélection, la tension d'entrée est appliquée à l'entrée d'une colonne et tout au plus une de ses M lignes est sélectionnée. Le nombre de neurones de sortie n'est donc plus limité par les chutes résistives.

Les résultats analytiques utilisées précédemment peuvent être adaptés à cette nouvelle situation simplement en remplaçant $V_{\text{PRÉ}}^{(N-1)}$ par $V_{\text{PRÉ}}^{(M-1)}$ et en considérant une valeur de résistance R_i très importante[I] pour les $M-1$ lignes qui ne sont pas sélectionnées. Ceci permet d'évaluer le nombre d'entrées présynaptiques qu'il est envisageable d'utiliser. La figure 5.11 présente ainsi, de façon analogue à la figure 5.10, la chute de tension aux bornes de la synapse de la dernière ligne d'une colonne en fonction du nombre de lignes. À chute de tension égale, en comparaison de la structure synaptique sans sélecteur, l'ajout des composants de sélection permet de repousser d'un à deux ordres de grandeur le nombre de synapses qu'il est possible d'utiliser. Toutefois, même dans ces conditions, l'utilisation des 32 768 lignes tel qu'imaginée dans les simulations du chapitre 3 semble difficile.

5.4.2 Délai dû à l'effet capacitif d'une piste

Lorsqu'une piste devient longue, il convient de vérifier dans quelle mesure la composante capacitive qu'elle apporte peut altérer le signal V_{in} appliqué à son entrée. Prenons le cas d'une colonne de N synapses pourvues chacune d'un transistor de sélection, en supposant que l'unique ligne sélectionnée, d'indice $M-1$, est la plus éloigné du circuit de commande de la piste en question. En faisant l'hypothèse simplificatrice que les transistors de sélection présentent une résistance infinie dans leur état bloqué, cette situation peut être schématisée selon la figure 5.12, où r et c désignent respectivement la résistance parasite et la capacité parasite apportées par chaque portion de piste entre une ligne et la précédente.

En notant p la variable de LAPLACE, la fonction de transfert de ce filtre passe-bas du premier

I. Par exemple une dizaine de gigaohms.

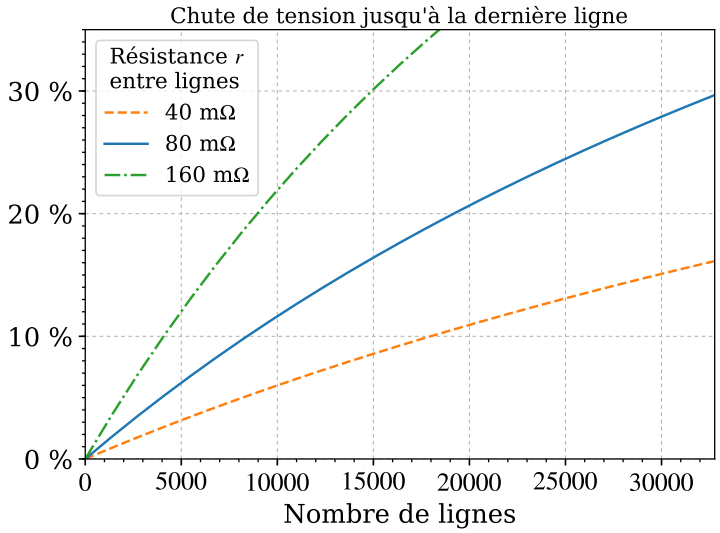


FIGURE 5.11 – Chute subie de la part de la tension $V_{\text{PRÉ}}^{(M-1)}$ aux bornes du dernier élément synaptique par rapport à la valeur V_{in} appliquée à l'entrée de cette colonne, dans une structure matricielle avec des transistors de sélection. Les valeurs de la résistance parasite r correspondent, dans l'ordre croissant, à des longueurs de piste de 100 nm (pointillé orange), 200 nm (trait plein bleu) et 400 nm (ligne brisée verte). Les lignes d'indices 0 à $M-2$ sont non sélectionnées (une valeur de résistance de 10 GΩ est utilisée pour chacune d'elles), tandis que la ligne dernière ligne (d'indice $M-1$) est sélectionnée, sa synapse étant supposée dans son état de faible résistance $R_{\text{ON}} = 6\,\text{k}\Omega$.

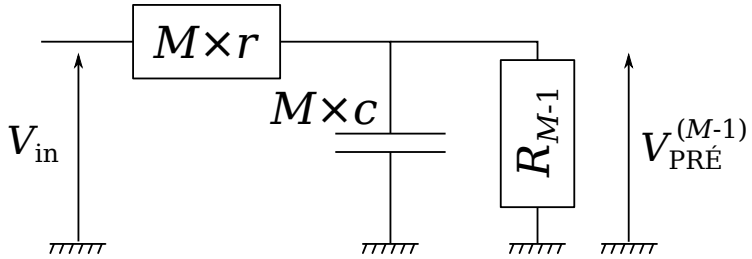


FIGURE 5.12 – Schéma du filtre RC équivalent à une colonne de M synapses munies de transistors de sélection dont les $M-1$ premiers sont bloqués et le dernier passant. V_{in} désigne la tension appliquée à l'entrée de la colonne et $V_{\text{PRÉ}}^{(M-1)}$ celle appliquée sur la synapse sélectionnée. Entre deux lignes successives, l'interconnexion métallique ajoute une résistance parasite r et une capacité parasite c.

ordre a pour expression

$$\mathscr{H}(p) \hat{=} \frac{V_{\text{PRÉ}}^{(M-1)}(p)}{V_{\text{in}}(p)} = \frac{R_{M-1}}{R_{M-1}+Mr} \times \frac{1}{1+\tau_{\text{ligne}} \times p}, \tag{5.12}$$

dont la constante de temps est

$$\tau_{\text{ligne}} = M^2 \times \frac{R_{M-1}}{R_{M-1}+Mr} \times rc. \tag{5.13}$$

La figure 5.13 trace l'évolution du temps caractéristique τ_{ligne} en fonction du nombre de lignes M d'une colonne de synapses munies de transistors de sélection et reliées au même circuit postsynaptique. Même dans le cas d'interconnexions métalliques courtes (100 nm), dès lors qu'une colonne comporte plus de 10 000 entrées présynaptiques, cette constante de temps τ_{ligne} devient proche de 10 % de la durée des impulsions de lecture (1 ns).

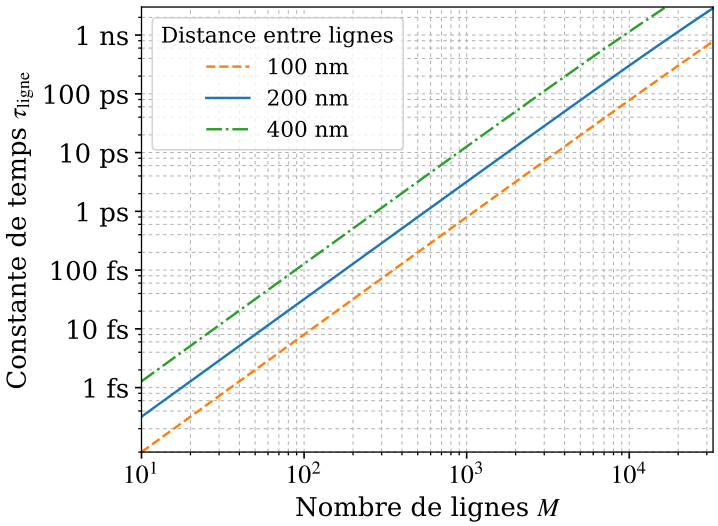


FIGURE 5.13 – Évolution de la constante de temps τ_{ligne} apportée par une piste métallique reliant M lignes d'une structure 1T-1R, calculée selon la formule (5.13). La résistance R_{M-1} de la synapse sélectionnée est prise égale à celle de son état R_{OFF}= 13,2 kΩ, nous plaçant ainsi dans le pire cas.

Dans les cas considérés, la limitation du nombre d'entrées présynaptiques d'une matrice synaptique avec transistor de sélection (structure 1T-1R) en raison de la déformation du signal appliqué sur une colonne, est sensiblement du même ordre de grandeur que les limitations précédentes liées aux effets purement résistifs : de plusieurs milliers à quelques dizaines de milliers d'entrées. Ces quantités soulèvent une interrogation sur la faisabilité d'une matrice présentant d'un seul bloc un aussi grand nombre d'entrées présynaptiques que celui des simulations à l'échelle du système du chapitre 3.

Dans les cas nécessitant à priori le recours à une matrice synaptique de dimensions importantes, il est potentiellement intéressant d'étudier en premier lieu la possibilité d'adapter la

topologie des connexions entre la couche des neurones d'entrée et celle des neurones de sortie, afin de limiter le nombre de synapses impliquées. L'adoption d'une stratégie de connexions aléatoires, ou encore d'un pavage plus parcimonieux des connexions peuvent en effet constituer des pistes pour réduire les dimensions réellement requises de la matrice de synapses. Dans l'hypothèse où de telles stratégies ne sont pas envisageables, il est également possible de découper la matrice synaptique en plusieurs blocs physiques de dimensions inférieures et de dupliquer les circuits d'entrée et de sortie associés. Cette solution donne cependant lieu à une augmentation de la surface occupée ainsi que de la consommation énergétique.

5.5 Conclusions

5.5.1 Vers la conception d'un circuit complet

La poursuite naturelle de ces réflexions est une étude plus approfondie de ces problématiques spécifiques à l'échelle du circuit, à l'aide notamment d'outils de simulation électrique de type SPICE. De telles simulations permettront notamment d'inclure les effets de la variabilité des composants des circuits, dont en particulier celle des transistors, décrite de façon réaliste par le kit de conception de procédé.

Par ailleurs, les limites proposées dans ce chapitre ont été calculées sous des hypothèses fortes, généralement dans des cas extrêmes, à la manière des études du pire cas pour les circuits dont le comportement doit être déterministe. Sachant que les systèmes neuromorphiques tolèrent assez bien la présence d'erreurs et de bruit dans les signaux qui leur sont appliqués, et que l'utilisation qui est faite des jonctions tunnel magnétiques est probabiliste, des simulations électriques plus réalistes que de simples calculs d'ordre de grandeur rendront possible la définition de limites plus proches d'une utilisation réaliste de ces systèmes. Une partie des imperfections étudiées dans ce chapitre pourrait également être directement intégrée dans les outils de simulation à l'échelle du système, comme par exemple les chutes résistives.

Les travaux de simulation électrique ont débuté, avec l'accueil d'un stagiaire en première année de master, Qifan WU, que j'ai eu l'occasion d'encadrer avec Nicolas LOCATELLI. Qifan a travaillé à la mise en place de simulations sous Spectre® d'un système réaliste inspiré de l'architecture du chapitre 3 tenant également compte des réflexions de ce chapitre. L'arrivée inopinée de 2000 m^3 d'oxyde de dihydrogène dans le sous-sol du laboratoire au début du stage a quelque peu perturbé le déroulement de ce dernier. Malgré des conditions de travail dégradées, Qifan a néanmoins réussi à réaliser la majeure partie de ce nous envisagions initialement. Malheureusement le stage s'est terminé avant d'avoir eu l'occasion de tester l'ensemble du système. Nous avons toutefois bon espoir de pouvoir achever cette étude SPICE dans un futur proche.

5.5.2 Nécessité d'une structure matricielle avec des composants de sélection

Les premières estimations réalisées dans ce chapitre suggèrent fortement qu'une matrice synaptique sans éléments de sélection (structure 1R) n'est pas envisageable dès lors qu'elle comporte un nombre important d'entrées ou de sorties. Ce constat existe à la fois dans le cadre des opérations de lecture, où la moindre tension de décalage en entrée des circuits postsynaptiques provoque une consommation non négligeable tout en noyant les signaux de lecture utiles, et dans celui de la programmation des synapses, où la contribution des courants parasites de fuites ou les dimensions requises pour les transistors d'écriture peuvent devenir importantes voire excessives.

Adopter une structure matricielle avec des transistors de sélection permet de réduire significativement ces problèmes, rendant par ailleurs possible l'utilisation de matrices synaptiques de dimensions sensiblement plus élevées : le nombre de neurones postsynaptiques n'est plus limité par les courants parasites de fuites ou les chutes résistives, et le recours à plusieurs milliers d'entrées présynaptiques semble envisageable. La programmation séquentielle associée à cette structure ne peut cependant être mise en œuvre qu'à la condition que la vitesse d'écriture des synapses soit suffisamment plus rapide que les constantes de temps de la tâche cognitive à réaliser.

Bilan et perspectives

> "The attack began at 6.18 p.m., just as he said it would. Judgment Day. The day the human race was nearly destroyed by weapons they'd built to protect themselves."
>
> John CONNOR

"CES QUELQUES *pages dressent le bilan synthétique des travaux de thèse présentés dans ce manuscrit. À la lumière de ces conclusions, il est également judicieux d'évaluer les pistes qui s'ouvrent pour les recherches à venir.* "

Bilan

De nouvelles puces électroniques, spécialisées pour de nouveaux usages émergents

L'explosion actuelle de la quantité de données numériques s'est accompagnée de l'apparition de nouveaux usages informatiques, qualifiés de « cognitifs » parce qu'ils se rapprochent des tâches que nous réalisons dans notre vie quotidienne : reconnaissance de visage, communication en langage naturel, etc. Les réseaux de neurones artificiels, qui s'inspirent du fonctionnement des cerveaux biologiques, sont de ce fait une solution prometteuse pour y répondre. Ces réseaux sont constitués d'unités de calcul (généralement simples), les neurones, massivement connectés par des synapses, qui jouent les rôle d'une mémoire distribuée. Durant une phase dite « d'apprentissage », les poids de ces synapses sont modulés de façon supervisée ou non supervisée afin d'apprendre à traiter la tâche attendue sans le recours à une programmation explicite de la solution. La capacité des synapses à voir leur poids ainsi modifié est qualifiée de « plasticité ».

L'architecture conventionnelle des ordinateurs, qui sépare les unités de calcul (et de contrôle) de la mémoire, est peu adaptée à la mise en œuvre efficiente des réseaux neuronaux artificiels. En raison de la croissance actuelle et attendue des usages cognitifs, il est pertinent de chercher des moyens d'y répondre de façon plus efficace, notamment du point de vue énergétique. Un axe désormais majeur de recherche sur les réseaux de neurones artificiels est la conception d'architectures matérielles spécialisées pour ce type d'usages, avec la co-intégration de circuits neuronaux en technologie silicium et de matrices mémoires. Deux grandes approches sont actuellement explorées, avec d'un côté les stratégies numériques prêtes à sacrifier une partie des gains énergétiques contre un aspect versatile, et de l'autre côté des stratégies à signaux mixtes, dont une partie des opérations est effectuée de manière analogique afin de gagner en sobriété énergétique.

Les opérations analogiques sont généralement basées sur l'utilisation de transistors à effet de champ, qui possèdent en régime de faible inversion une relation courant-tension exponentielle, capable de reproduire *naturellement* des comportements fréquemment observés en neurobiologie et dont les équations sont autrement coûteuses à calculer. Ce constat, formulé par C. MEAD à la fin des années 1980, a depuis permis la conception de neurones artificiels énergétiquement efficaces. La puce TrueNorth des laboratoires IBM® a par ailleurs récemment prouvé qu'il était également possible d'atteindre une efficacité énergétique importante avec une approche entièrement numérique.

Les technologies mémoires innovantes au secours des synapses artificielles

Un défi important qui demeure avec les multiples réalisations matérielles présentées dans la littérature est la mise en œuvre de l'apprentissage directement sur puce. Dans la grande majorité des cas cette étape est en effet réalisée à l'extérieur, au moyen d'outils informatiques conven-

tionnels, et seulement ensuite les poids synaptiques obtenus sont chargés dans l'architecture. Certaines réalisations expérimentales disposent néanmoins de synapses plastiques mais les caractéristiques de ces dernières restent limitées. Par ailleurs, dans l'ensemble des réalisations matérielles actuelles, la surface des circuits dédiés aux fonctions synaptiques consomment une part très significative des puces.

À ce titre, les nanocomposants mémoires innovants sont de sérieux candidats aux usages synaptiques, en offrant des avantages tels que l'augmentation de la densité d'intégration, un caractère non volatil, et la possibilité de mettre en œuvre tout ou partie de la règle d'apprentissage. À l'image de l'utilisation des transistors à effet de champ en régime de faible inversion, ce dernier point repose sur l'idée d'exploiter la physique riche des nanocomposants pour réaliser directement des opérations coûteuses autrement. Les technologies potentielles demeurent actuellement nombreuses.

Dans le cadre des mises en œuvre logicielles de réseaux neuronaux, l'approche originelle et encore largement en usage est le recours à des synapses quasiment analogiques, présentant un nombre très important de niveaux intermédiaires entre leur états extrêmes ON et OFF. Une tendance actuelle en apprentissage automatique est cependant à la réduction de la résolution des poids synaptiques utilisés *après* l'apprentissage, ce qui permet d'accélérer le traitement de la tâche cognitive apprise, tout en diminuant la consommation énergétique. De récents travaux en sciences informatiques suggèrent qu'il est également possible de réduire le nombre d'états d'une synapse *durant* l'apprentissage, généralement au moyen d'approches probabilistes.

À la lumière de ces observations du chapitre 1, notre attention durant ces travaux de thèse s'est portée sur deux familles de nanocomposants mémoires innovants. D'un côté les jonctions tunnel magnétiques à transfert de spin, composants binaires et technologiquement mûres, et de l'autre côté les cellules électrochimiques métalliques Ag_2S, en développement académique et offrant un riche comportement analogique dynamique. Après une étape de modélisation des composants, utilisée ensuite au sein d'outils de simulation développés en interne, nous avons dans chaque cas démontré les possibilités d'apprentissage qu'offre un usage synaptique de ces deux technologies.

Les jonctions tunnel magnétiques comme synapses binaires et stochastiques

Une première partie du chapitre 2 a été consacrée à la présentation succincte des principaux phénomènes physiques à l'origine du comportement des jonctions tunnel magnétiques à transfert de spin. Ces composants mémoires binaires présentent en particulier un caractère stochastique intrinsèque, généralement préjudiciable dans un usage mémoire conventionnel. Dans le cadre d'une approche macrospin, des simulations Monte-Carlo résolvant l'équation différentielle de la dynamique magnétique de tels composants nous ont permis d'obtenir des données statistiques sur le délai avant commutation lors d'une programmation. Un travail de modélisation a ensuite été réalisé en seconde partie du chapitre 2, menant à l'obtention d'un ensemble d'équations analytiques permettant de décrire la distribution statistique dudit délai

dans différents régimes de programmation. Une avancée significative de ce modèle est d'établir une continuité entre les modèles disponibles dans la littérature, valables dans les régimes de programmation extrêmes mais incompatible entre eux. Notre modèle est ainsi capable de décrire le comportement stochastique du délai avant commutation dans le régime de programmation intermédiaire entre les régimes de courant fort et de courant faible.

Dans le chapitre 3, nous avons présenté un simulateur informatique développé durant la thèse permettant d'étudier un système utilisant un grand nombre de jonctions tunnel magnétiques à transfert de spin comme synapses artificielles binaires. La description de ces dernières repose sur le modèle analytique du chapitre 2, bien plus rapide à calculer que la résolution de l'équation différentielle magnétique de chaque composant synaptique.

En adaptant de façon probabiliste une règle d'apprentissage de type *Spike-Timing Dependent Plasticity*, nous avons démontré la possibilité d'un apprentissage non supervisé du système précédent sur une tâche dynamique de détection de véhicules. Ceci met en lumière la possibilité de réaliser l'apprentissage directement sur une telle matrice de synapses artificielles binaires. L'aspect probabiliste de ce dernier est par ailleurs mis en œuvre en exploitant le comportement naturellement stochastique offert par la physique des nanocomposants synaptiques plutôt qu'un circuit électronique extérieur de génération de nombres aléatoires.

La fin du chapitre 3 a été consacrée à l'étude de l'influence des imperfections synaptiques (variabilité entre composants, fluctuations aléatoires des caractéristiques synaptiques, etc.) et du régime de programmation employé sur les performances du système. Dans le cadre de la tâche cognitive étudiée, ce travail nous a permis d'observer que le régime *intermédiaire* offre une excellente tolérance aux imperfections synaptiques, au-delà des valeurs de dispersion rapportées dans les démonstrateurs technologiques de la littérature, tout en nécessitant moins d'énergie pour la programmation de la matrice synaptique que les autres régimes. Les estimations de cette consommation sont également inférieures aux valeurs rapportées sur la même tâche pour un système similaire utilisant une autre technologie de synapses artificielles.

Les cellules électrochimiques métalliques, une richesse de comportements plastiques à exploiter

Le chapitre 4 a porté sur une autre famille de nanocomposants mémoires innovants, les cellules électrochimiques métalliques. À la différence des jonctions tunnel magnétiques, ces composants présentent un continuum d'états de conductance entre deux valeurs extrêmes. La déclinaison à base d'un électrolyte de sulfure d'argent (Ag_2S) étudiée durant la thèse a été fabriquée et caractérisée par nos collaborateurs de l'Université de Lille I. Une particularité majeure de ces composants est de présenter *plusieurs* comportements plastiques, à différentes échelles de temps. Nous avons notamment découvert une plasticité de type *Spike-Timing Dependent Plasticity*. L'originalité de ce nouveau comportement plastique est de pouvoir être excité par de simples impulsions de tension rectangulaires (identiques) et temporellement rapprochées.

Après un résumé des propriétés déjà étudiées de ces composants, nous avons présenté

dans le chapitre 4 le modèle analytiques simple développé avec nos collaborateurs pour reproduire le comportement de type *Spike-Timing Dependent Plasticity* observée expérimentalement. Ce modèle nous a ensuite permis de mieux comprendre l'interaction des échelles de temps d'excitation menant à la potentialisation de ces synapses artificielles. Cette avancée a été utilisée pour réaliser en simulation une preuve de concept d'un apprentissage à l'échelle d'un système. Les outils informatiques utilisés sont adaptés de ceux du chapitre 3 et la tâche cognitive traitée est une version simplifiée de celle présentée dans ce même chapitre, adaptée aux particularités dynamiques des cellules électrochimiques métalliques Ag_2S. L'architecture employée est conçue pour réaliser un apprentissage hebbien non supervisé original, qui exploite l'interaction des dynamiques temporelles des comportements plastiques pour potientialiser les synapses et tire profit de la relaxation naturelle de ces dernières pour les déprimer. Les premières études de cas incluant une variabilité synaptique non nulle ou la présence de bruits dans les signaux d'entrée suggèrent une bonne tolérance du système à de telles perturbations.

Attention à ne pas oublier l'échelle du circuit !

L'exploitation de nanocomposants mémoires innovants au sein d'un système neuromorphique soulève à l'échelle des circuits employés pour leur lecture et écriture des problématiques importantes, qu'il convient de ne pas négliger puisqu'elles influencent de façon significative la conception de puces réelles. Ces défis et les solutions à y apporter dépendent par ailleurs des cellules mémoires utilisées comme synapses artificielles.

Le chapitre 5 a ainsi présenté certaines interrogations qui se posent à l'échelle des circuits de lecture et d'écriture dans le cas de l'utilisation des jonctions tunnel magnétiques à transfert de spin présentée au chapitre 3, en s'aidant d'un kit de conception de procédés industriel récent. À l'issue des réflexions menées dans ce chapitre, l'adoption d'une matrice synaptique sans composant de sélection (structure 1R) ne semble pas réaliste. L'existence d'une tension de décalage, même de valeur raisonnable, en entrée d'un circuit postsynaptique de lecture suffit à limiter fortement le nombre d'entrées présynaptiques utilisables et entraîne dans le même temps une consommation statique importante en comparaison de la fraction réellement utile. Par ailleurs, la programmation entièrement parallèle permise par ce type de structure synaptique requiert des circuits d'écriture pré- et postsynaptiques capables de délivrer des courants importants, même dans le cas d'une matrice synaptique de dimensions modestes (quelques centaines d'entrées ou sorties).

Une matrice de synapses chacune associées à un transistor de sélection (structure 1T-1R) permet quant à elle d'envisager la mise en œuvre d'un système comparable à celui du chapitre 3. Par ailleurs, le kit de conception de procédés à notre disposition suggère que la diminution de la densité d'intégration due aux transistors de sélection puisse être relativement modérée. En effet, dans l'hypothèse d'une phase de programmation séquentielle, rendue plausible dans la pratique grâce à la vitesse d'écriture des jonctions tunnel magnétiques à transfert de spin, les courants d'écriture requis peuvent être délivrés par des transistors proches de leurs dimensions

minimales. Or ces dimensions se rapprochent toujours davantage de celles des jonctions tunnel magnétiques développées.

Perspectives

À court terme

Poursuivre l'étude à l'échelle du circuit débutée sur l'utilisation synaptique des jonctions tunnel magnétiques à transfert de spin. La plupart des réflexions analytiques formulées dans le chapitre 5 se placent dans l'hypothèse du pire cas, tel que cela est généralement fait lors de la conception de circuits conventionnels déterministes. Puisque nous avons pu observer dans le chapitre 3 une bonne tolérance de l'architecture étudiée à des quantités significatives d'imperfections des jonctions tunnel magnétiques, il serait pertinent de procéder à des simulations électriques de type SPICE dans des situations plus proches des usages pratiques. En couplant la description réaliste des transistors dans la technologie du kit de conception de procédés à notre disposition et le modèle Verilog-A développé par N. LOCATELLI à partir des équations analytiques du chapitre 2, nous espérons ainsi pouvoir évaluer dans quelle mesure les ordres de grandeurs théoriques calculés dans le chapitre 5 surestiment les limitations au regard de la fonctionnalité réellement obtenue en pratique. Ce travail de simulation électrique a été amorcé avec le stage de Q. WU.

Accroître les dimensions de la matrice synaptique de cellules électrochimiques métalliques Ag_2S simulée. Maintenant qu'une preuve de concept d'apprentissage par un ensemble des cellules électrochimiques métalliques Ag_2S a été faite au moyen de simulations à l'échelle d'un système de dimensions modestes (quelques centaines de synapses) et que les travaux du chapitre 4 nous ont permis de mieux comprendre comment exploiter à notre avantage les différents comportements plastiques disponibles, il semble intéressant de vouloir procéder à la simulation d'un système de plus grandes dimensions. Ceci donnerait l'occasion d'approfondir l'étude de la robustesse à la variabilité synaptique qui a déjà été débutée : l'augmentation du nombre de synapses, en apportant une forme de redondance, pourrait notamment donner lieu à une amélioration de la tolérance aux imperfections des nanocomposants.

À plus long terme

Aller plus loin avec les composants d'électronique de spin

Vers une réalisation matérielle. Il semble judicieux de chercher à vérifier les bonnes performances obtenues en simulation en réalisant un prototype matériel de l'architecture du chapitre 3, qui tienne bien entendu compte de l'éclairage apporté par les réflexions à l'échelle

du circuit du chapitre 5, ainsi que des futures conclusions des simulations électriques SPICE. Cette preuve de concept ayant de fortes chances d'être d'échelle modeste, il doit être possible de s'inspirer de la tâche cognitive simplifiée utilisée au chapitre 4.

Si les réflexions à l'échelle du circuit introduites au chapitre 5 ont mis en lumière la nécessité d'adapter en partie l'architecture envisagée dans le chapitre 3, elles n'ont toutefois pas soulevé de problème irrémédiable s'opposant au principe même d'une telle architecture. En raison de la maturité technologique des jonctions tunnel magnétiques à transfert de spin, il semble ainsi raisonnablement permis d'espérer connaître, à assez brève échéance après la conception de la preuve de concept mentionnée précédemment, le développement d'une véritable puce co-intégrant ces dernières avec des circuits neuronaux CMOS, selon un concept inspiré de celui présenté dans les chapitres 3 et 5.

Explorer les possibilités de synapses à plus de deux niveaux. Si nous avons démontré la possibilité d'un apprentissage non supervisé d'une tâche dynamique à l'aide de synapses binaires et probabilistes, cette approche ne se prête pas nécessairement à tous les usages. L'apprentissage de certaines tâches ne fonctionne pas si la variation des poids synaptiques est trop abrupte. Il serait donc utile d'évaluer dans quelle mesure l'électronique de spin permettrait de réaliser des synapses artificielles à plus de deux niveaux.

L'association en série ou en parallèle de plusieurs jonctions tunnel magnétiques à transfert de spin est à priori une solution naturelle pour concevoir des synapses dotées de quelques niveaux intermédiaires. Il faudrait alors notamment se pencher sur la densité de probabilité des modifications de conductance d'un tel ensemble, ainsi que la possibilité de partager un transistor de sélection et les conséquences en termes de densité d'intégration. De premiers résultats ont été présentés dans la référence [248], sur la base d'une étude conceptuelle.

Si un grand nombre d'états de conductances intermédiaires est requis, d'autres nanocomposants mémoires innovants d'électronique de spin faisant encore l'objet de recherches exploratoires sont envisageables. Nous pouvons par exemple citer l'approche associant le déplacement d'une paroi de domaine magnétique dans un ruban ferromagnétique, à la manière de certaines pistes recherches sur les mémoires spintroniques conventionnelles[275], avec une jonction tunnel magnétique[276,277]. Ceci permet d'obtenir un composant synaptique dont la résistance électrique n'est plus binaire mais offre au contraire un continuum d'états, selon la position de la paroi de domaine[276,277]. La séparation entre chemin d'écriture et chemin de lecture est une autre caractéristique notable de ces composants synaptiques potentiels.

Poursuivre l'étude des cellules électrochimiques métalliques Ag_2S

Mettre en place une modélisation s'appuyant davantage sur la physique. Si le modèle comportemental simple des cellules électrochimiques métalliques Ag_2S que nous avons présenté

275. S. PARKIN, M. HAYASHI et L. THOMAS, Science, 2008.
276. X. WANG et coll., IEEE Electron Device Letters, 2009.
277. S. LEQUEUX et coll., arXiv preprint arXiv :1605.07460, 2016.

au chapitre 4 nous a déjà permis de mettre en place une preuve de concept d'apprentissage, il serait intéressant de tenter d'élaborer un modèle qui se base davantage sur la physique même de ces nanocomposants. Une telle modélisation permettrait possiblement de mieux comprendre l'origine de la dynamique de type *Spike-Timing Dependent Plasticity* ou encore de mieux saisir la dépendance des paramètres du modèle actuel aux caractéristiques des impulsions employées. Ces connaissances pourraient alors être utilisées pour adapter l'utilisation synaptique qui est faite de ces nanocomposants à leurs caractéristiques propres.

Développer l'ingénierie de ces nanocomposants. Une question importante et pour le moment ouverte est la possibilité de fabriquer des cellules électrochimiques métalliques Ag_2S dont les caractéristiques intrinsèques, telles que les échelles de temps du comportement de type *Spike-Timing Dependent Plasticity* ou de la rétention à long terme, soient plus proches de tâches cognitives pratiques. À ce titre, disposer d'une modélisation plus physique telle qu'évoquée juste avant pourrait permettre de mieux comprendre la dépendance aux paramètres matériaux des comportements plastiques intéressants du point de vue applicatif.

Et l'échelle du circuit dans tout ça ? Il serait judicieux qu'à terme, l'obtention d'un modèle physique plus complet serve également à l'écriture d'un modèle compact pour ces nanocomposants. Ceci permettrait l'étude des problématiques propres à ces synapses artificielles à l'échelle du circuit et pouvant avoir une influence à d'autres échelles.

Le mot de la fin

Les travaux présentés dans ce manuscrit montrent qu'il est pertinent et réaliste d'utiliser des nanocomposants mémoires innovants pour jouer le rôle de synapses artificielles au sein d'architectures neuro-inspirées *capables d'un apprentissage non supervisé sur puce*. Nous avons démontrée que ceci est possible même avec des technologies aux caractéristiques forts différentes, *à la condition de disposer d'une bonne connaissance du comportement plastique* de ces dernières. Connaître finement les comportements qu'offrent naturellement les synapses artificielles employées permet en effet d'adapter l'utilisation qui en est faite, par exemple à travers la règle d'apprentissage, afin de bénéficier au maximum de la richesse des opérations que ces nanocomposants sont capables de réaliser *de manière intrinsèque*. Ceci implique de devoir *travailler aux multiples échelles du problème*, de la physique du composant à l'architecture du système. Il est donc fort probable que le succès futur des systèmes de calcul neuromorphiques repose sur la capacité de leur communauté à mettre en place une recherche interdisciplinaire de qualité.

Bibliographie

[1] M. HILBERT et P. LÓPEZ. The world's technological capacity to store, communicate, and compute information. *Science*, **332** : 60–65, 2011. (voir p. 2)

[2] Visual Networking Index CISCO. The zettabyte era–trends and analysis. *Cisco white paper*, 2013. (voir p. 2)

[3] J. GUBBI, R. BUYYA, S. MARUSIC et M. PALANISWAMI. Internet of Things (IoT) : A vision, architectural elements, and future directions. *Future Generation Computer Systems*, **29** : 1645–1660, 2013. (voir p. 2)

[4] B. LANNOO, S. LAMBERT, W. VAN HEDDEGHEM, M. PICKAVET, F. KUIPERS, G. KOUTITAS, H. NIAVIS, A. SATSIOU, M. TILL, A. F. BECK et coll. D8. 1. Overview of ICT energy consumption. *Network of Excellence in Internet Science, FP7*, **8** : 2013. (voir p. 2)

[5] ACADÉMIE DES TECHNOLOGIES *Impact des TIC sur la consommation d'énergie à travers le monde* rapp. tech. EDP Sciences, 2015 (voir p. 2)

[6] P. DUBEY. Recognition, mining and synthesis moves computers to the era of tera. *Technology Intel Magazine*, 1–10, 2005. (voir p. 2)

[7] A. KRIZHEVSKY, I. SUTSKEVER et G. E. HINTON. "Imagenet classification with deep convolutional neural networks" *Advances in neural information processing systems*. 2012. 1097–1105 (voir p. 2, 18)

[8] G. MOORE. Cramming More Components Onto Integrated Circuits. *Electronics*, **38** : 1965. (voir p. 3)

[9] W. ARDEN, M. BRILLOUËT, P. COGEZ, M. GRAEF, B. HUIZING et R. MAHNKOPF. More-than-Moore white paper. *ITRS*, 2010. (voir p. 3)

[10] K. ROY, B. JUNG, D. PEROULIS et A. RAGHUNATHAN. Integrated Systems in the More-Than-Moore Era : Designing Low-Cost Energy-Efficient Systems Using Heterogeneous Components. *IEEE Design Test*, **33** : 56–65, 2016. DOI : 10.1109/MDT.2011.49 (voir p. 3)

[11] K. PALEM et A. LINGAMNENI. "What to do about the end of Moore's law, probably!" *Proceedings of the Annual Design Automation Conference*. ACM 2012. 924–929 (voir p. 3)

[12] D. ATTWELL et S. B. LAUGHLIN. An energy budget for signaling in the grey matter of the brain. *Journal of Cerebral Blood Flow & Metabolism*, **21** : 1133–1145, 2001. (voir p. 3, 21)

[13] P. A. MEROLLA, J. V. ARTHUR, R. ALVAREZ-ICAZA, A. S. CASSIDY, J. SAWADA, F. AKOPYAN, B. L JACKSON, N. IMAM, C. GUO, Y. NAKAMURA et coll. A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science*, **345** : 668–673, 2014. DOI : 10.1126/science.1254642 (voir p. 4, 43, 45, 48, 103, 156, 167)

[14] C. MEAD. Neuromorphic electronic systems. *Proceedings of the IEEE*, **78** : 1629–1636, 1990. DOI : 10.1109/5.58356 (voir p. 4, 38)

[15] C. MEAD. Boston, MA, USA : Addison-Wesley Longman Publishing Co., Inc., 1989. (voir p. 4)

[16] S. SAÏGHI, Y. BORNAT, J. TOMAS, G. LE MASSON et S. RENAUD. A Library of Analog Operators Based on the Hodgkin-Huxley Formalism for the Design of Tunable, Real-Time, Silicon Neurons. *IEEE Transactions on Biomedical Circuits and Systems*, **5** : 3–19, 2011. DOI : 10.1109/TBCAS.2010.2078816 (voir p. 4, 103)

[17] G. INDIVERI, B. LINARES-BARRANCO, T. J. HAMILTON, R. ETIENNE-CUMMINGS, T. DELBRÜCK, S.-C. LIU, P. HÄFLIGER, S. RENAUD, J. SCHEMMEL, G. CAUWENBERGHS, J. ARTHUR, S. SAÏGHI, J. WIJEKOON et K. BOAHEN. Neuromorphic silicon neuron circuits. *Frontiers in Neuroscience*, **5** : 73, 2011. DOI : 10.3389/fnins.2011.00073 (voir p. 4, 103)

[18] G. INDIVERI et S.-C. LIU. Memory and information processing in neuromorphic systems. *Proceedings of the IEEE*, **103** : 1379–1397, 2015. (voir p. 5)

[19] Y. LECUN, Y. BENGIO et G. HINTON. Deep learning. *Nature*, **521** : 436–444, 2015. (voir p. 5, 19)

[20] F. A. C. AZEVEDO, L. R. B. CARVALHO, L. T. GRINBERG, J. M. FARFEL, R. E. L. FERRETTI, R. E. P. LEITE, R. LENT, S. HERCULANO-HOUZEL et coll. Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *Journal of Comparative Neurology*, **513** : 532–541, 2009. DOI : 10.1002/cne.21974 (voir p. 11)

[21] P. LENNIE. The Cost of Cortical Computation. *Current Biology*, **13** : 493 –497, 2003. DOI : 10.1016/S0960-9822(03)00135-0 (voir p. 11)

[22] W. S. MCCULLOCH et W. PITTS. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, **5** : 115–133, 1943. (voir p. 11)

[23] D. QUERLIOZ, O. BICHLER, P. DOLLFUS et C. GAMRAT. Immunity to device variations in a spiking neural network with memristive nanodevices. *IEEE Transactions on Nanotechnology*, **12** : 288–295, 2013. (voir p. 12, 104)

[24] P. MEROLLA, R. APPUSWAMY, J. ARTHUR, S. K. ESSER et D. MODHA. Deep neural networks are robust to weight binarization and other non-linear distortions. *arXiv preprint arXiv :1606.01981*, 2016. (voir p. 12, 32)

[25] D. O. HEBB. John Wiley & Sons, 1949. (voir p. 13, 125)

[26] S. LÖWEL et W. SINGER. Selection of intrinsic horizontal connections in the visual cortex by correlated neuronal activity. *Science*, **255** : 209, 1992. (voir p. 13)

[27] F. ROSENBLATT. The perceptron : a probabilistic model for information storage and organization in the brain. *Psychological review*, **65** : 386, 1958. (voir p. 13)

[28] P. WERBOS. Beyond regression : New tools for prediction and analysis in the behavioral sciences. *Ph.D. thesis*, 1974. (voir p. 15)

[29] D. B. PARKER. MIT Press, 1985. (voir p. 15)

[30] Y. LECUN. "Une procédure d'apprentissage pour réseau à seuil asymmétrique (a Learning Scheme for Asymmetric Threshold Networks)" *Proceedings of Cognitiva*. 1985. (voir p. 15)

[31] D. E. RUMELHART, G. E. HINTON et R. J. WILLIAMS *Learning internal representations by error propagation* rapp. tech. DTIC Document, 1985 (voir p. 15)

[32] S. HOCHREITER, Y. BENGIO, P. FRASCONI et J. SCHMIDHUBER. *Gradient flow in recurrent nets : the difficulty of learning long-term dependencies*. 2001. (voir p. 16)

[33] G. E. HINTON, S. OSINDERO et Y.-W. TEH. A fast learning algorithm for deep belief nets. *Neural computation*, **18** : 1527–1554, 2006. (voir p. 17)

[34] Y. BENGIO, P. LAMBLIN, D. POPOVICI, H. LAROCHELLE et coll. Greedy layer-wise training of deep networks. *Advances in neural information processing systems*, **19** : 153, 2007. (voir p. 17)

[35] M. C. CRAIR, D. C. GILLESPIE et M. P. STRYKER. The role of visual experience in the development of columns in cat visual cortex. *Science*, **279** : 566–570, 1998. (voir p. 17)

[36] G. G. BLASDEL. Orientation selectivity, preference, and continuity in monkey striate cortex. *The Journal of Neuroscience*, **12** : 3139–3161, 1992. (voir p. 17)

[37] J. BACKUS. Can Programming Be Liberated from the Von Neumann Style ? : A Functional Style and Its Algebra of Programs. *Commun. ACM*, **21** : 613–641, 1978. DOI : 10.1145/359576.359579 (voir p. 18)

[38] R. RAINA, A. MADHAVAN et A. Y. NG. "Large-scale deep unsupervised learning using graphics processors" *Proceedings of the 26th annual international conference on machine learning*. ACM 2009. 873–880 (voir p. 18)

[39] D. SILVER, A. HUANG, C. J. MADDISON, A. GUEZ, L. SIFRE, G. VAN DEN DRIESSCHE, J. SCHRITTWIESER, I. ANTONOGLOU, V. PANNEERSHELVAM, M. LANCTOT et coll. Mastering the game of Go with deep neural networks and tree search. *Nature*, **529** : 484–489, 2016. (voir p. 19)

[40] R. VAILLANT, C. MONROCQ et Y. LECUN. Original approach for the localisation of objects in images. *IEEE Proceedings of Vision Image and Signal Processing*, **141** : 245–250, 1994. (voir p. 20)

[41] S. LAWRENCE, C. L. GILES, A. C. TSOI et A. D. BACK. Face recognition : A convolutional neural-network approach. *IEEE Transactions on Neural Networks*, **8** : 98–113, 1997. (voir p. 20)

[42] D. CIREŞAN, U. MEIER, J. MASCI et J. SCHMIDHUBER. Multi-column deep neural network for traffic sign classification. *Neural Networks*, **32** : 333–338, 2012. (voir p. 20)

[43] J. TOMPSON, R. GOROSHIN, A. JAIN, Y. LECUN et C. BREGLER. "Efficient object localization using convolutional networks" *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015. 648–656 (voir p. 20)

[44] Y. TAIGMAN, M. YANG, M. A. RANZATO et L. WOLF. "Deepface : Closing the gap to human-level performance in face verification" *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014. 1701–1708 (voir p. 20)

[45] D. H. HUBEL et T. N. WIESEL. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, **160** : 106–154, 1962. (voir p. 20)

[46] D. J. FELLEMAN et D. C. VAN ESSEN. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex*, **1** : 1–47, 1991. (voir p. 20)

[47] W. MAASS. Networks of spiking neurons : the third generation of neural network models. *Neural networks*, **10** : 1659–1671, 1997. (voir p. 21)

[48] S. B. FURBER. Brain-inspired computing. *IET Computers & Digital Techniques*, 2016. (voir p. 21, 31)

[49] A. L HODGKIN et A. F. HUXLEY. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology*, **117** : 500, 1952. (voir p. 21)

[50] D. HANSEL, G. MATO et C. MEUNIER. Phase Dynamics for Weakly Coupled Hodgkin-Huxley Neurons. *Europhysics Letters*, **23** : 367, 1993. (voir p. 22)

[51] L. LAPICQUE. Recherches quantitatives sur l'excitation électrique des nerfs traitée comme une polarisation. *J. Physiol. Pathol. Gen*, **9** : 620–635, 1907. (voir p. 21)

[52] R. B. STEIN. A theoretical analysis of neuronal variability. *Biophysical Journal*, **5** : 173, 1965. (voir p. 21)

[53] W. GERSTNER, R. RITZ et J. L. VAN HEMMEN. Why spikes ? Hebbian learning and retrieval of time-resolved excitation patterns. *Biological cybernetics*, **69** : 503–515, 1993. (voir p. 23)

[54] R. BRETTE. Philosophy of the Spike : Rate-Based vs. Spike-Based Theories of the Brain. *Frontiers in systems neuroscience*, **9** : 2015. (voir p. 23)

[55] S. THORPE, D. FIZE, C. MARLOT et coll. Speed of processing in the human visual system. *Nature*, **381** : 520–522, 1996. (voir p. 23)

[56] S. THORPE, A. DELORME et R. VAN RULLEN. Spike-based strategies for rapid processing. *Neural networks*, **14** : 715–725, 2001. (voir p. 24)

[57] G.-Q. BI et M.-M. POO. Synaptic modifications in cultured hippocampal neurons : dependence on spike timing, synaptic strength, and postsynaptic cell type. *The Journal of neuroscience*, **18** : 10464–10472, 1998. (voir p. 24, 103)

[58] T. MASQUELIER et S. J. THORPE. Unsupervised learning of visual features through spike timing dependent plasticity. *PLoS Computational Biology*, **3** : e31, 2007. (voir p. 25)

[59] A. VANARSE, A. OSSEIRAN et A. RASSAU. A Review of Current Neuromorphic Approaches for Vision, Auditory, and Olfactory Sensors. *Frontiers in Neuroscience*, **10** : 115, 2016. DOI : 10.3389/fnins.2016.00115 (voir p. 25)

[60] J. HASLER et H. MARR. Finding a roadmap to achieve large neuromorphic hardware systems. *Frontiers in Neuroscience*, **7** : 118, 2013. DOI : 10.3389/fnins.2013.00118 (voir p. 25)

[61] K. A. BOAHEN. Point-to-point connectivity between neuromorphic chips using address events. *IEEE Transactions on Circuits and Systems II : Analog and Digital Signal Processing*, **47** : 416–434, 2000. DOI : 10.1109/82.842110 (voir p. 25, 107)

[62] P. LICHTSTEINER, C. POSCH et T. DELBRÜCK. A 128× 128 120 dB 15 μs latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits*, **43** : 566–576, 2008. DOI : 10.1109/JSSC.2007.914337 (voir p. 25, 26, 101, 102, 108, 146, 161)

[63] C. POSCH, T. SERRANO-GOTARREDONA, B. LINARES-BARRANCO et T. DELBRÜCK. Retinomorphic Event-Based Vision Sensors : Bioinspired Cameras With Spiking Output. *Proceedings of the IEEE*, **102** : 1470–1484, 2014. DOI : 10.1109/JPROC.2014.2346153 (voir p. 25)

[64] C. BRANDLI, T. MANTEL, M. HUTTER, M. HÄFLINGER, R. BERNER, R. SIEGWART et T. DELBRÜCK. Adaptive pulsed laser line extraction for terrain reconstruction using a dynamic vision sensor. *Frontiers in Neuroscience*, **7** : 275, 2014. DOI : 10.3389/fnins.2013.00275 (voir p. 26)

[65] D. DRAZEN, P. LICHTSTEINER, P. HÄFLIGER, T. DELBRÜCK et A. JENSEN. Toward real-time particle tracking using an event-based dynamic vision sensor. *Experiments in Fluids*, **51** : 1465–1469, 2011. DOI : 10.1007/s00348-011-1207-y (voir p. 26)

[66] T. SERRANO-GOTARREDONA et B. LINARES-BARRANCO. A 128×128 1.5% Contrast Sensitivity 0.9% FPN 3 μs Latency 4 mW Asynchronous Frame-Free Dynamic Vision Sensor Using Transimpedance Preamplifiers. *IEEE Journal of Solid-State Circuits*, **48** : 827–838, 2013. DOI : 10.1109/JSSC.2012.2230553 (voir p. 26)

[67] S.-C. LIU, A. van SCHAIK, B. A. MINCH et T. DELBRÜCK. Asynchronous Binaural Spatial Audition Sensor With 2×64×4 Channel Output. *IEEE Transactions on Biomedical Circuits and Systems*, **8** : 453–464, 2014. DOI : 10.1109/TBCAS.2013.2281834 (voir p. 26)

[68] C.-H. LI, T. DELBRÜCK et S.-C. LIU. "Real-time speaker identification using the AEREAR2 event-based silicon cochlea" *IEEE International Symposium on Circuits and Systems*. IEEE 2012. 1159–1162 DOI : 10.1109/ISCAS.2012.6271438 (voir p. 26)

[69] M. ABDOLLAHI et S.-C. LIU. "Speaker-independent isolated digit recognition using an AER silicon cochlea" *IEEE Biomedical Circuits and Systems Conference*. 2011. 269–272 DOI : 10.1109/BioCAS.2011.6107779 (voir p. 26)

[70] S.-W. CHIU et K.-T. TANG. Towards a Chemiresistive Sensor-Integrated Electronic Nose : A Review. *Sensors*, **13** : 14214, 2013. DOI : 10.3390/s131014214 (voir p. 26)

[71] H. Y. HSIEH et K. T. TANG. VLSI Implementation of a Bio-Inspired Olfactory Spiking Neural Network. *IEEE Transactions on Neural Networks and Learning Systems*, **23** : 1065–1073, 2012. DOI : 10.1109/TNNLS.2012.2195329 (voir p. 26)

[72] K. T. NG, B. GUO, A. BERMAK, D. MARTINEZ et F. BOUSSAID. "Characterization of a logarithmic spike timing encoding scheme for a 4× 4 TiN oxide gas sensor array" *Sensors*. IEEE 2009. 731–734 DOI : 10.1109/ICSENS.2009.5398548 (voir p. 26)

[73] V. CHAN, C. JIN et A. van SCHAIK. Neuromorphic Audio-Visual Sensor Fusion on a Sound-Localising Robot. *Frontiers in Neuroscience*, **6** : 21, 2012. DOI : 10.3389/fnins.2012.00021 (voir p. 27)

[74] P. O'CONNOR, D. NEIL, S.-C. LIU, T. DELBRÜCK et M. PFEIFFER. Real-time classification and sensor fusion with a spiking deep belief network. *Frontiers in Neuroscience*, **7** : 178, 2013. DOI : 10.3389/fnins.2013.00178 (voir p. 27)

[75] B. A. OLSHAUSEN et D. J. FIELD. Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, **14** : 481–487, 2004. DOI : 10.1016/j.conb.2004.07.007 (voir p. 27)

[76] H.-S. P. WONG et S. SALAHUDDIN. Memory leads the way to better computing. *Nature Nanotechnology*, **10** : 191–194, 2015. (voir p. 28, 51–54)

[77] D. SOUDRY, D. DI CASTRO, A. GAL, A. KOLODNY et S. KVATINSKY. Memristor-based multilayer neural networks with online gradient descent training. *IEEE Transactions on Neural Networks and Learning Systems*, **26** : 2408–2421, 2015. (voir p. 28, 29)

[78] J.-S. SEO, B. LIN, M. KIM, P.-Y. CHEN, D. KADETOTAD, Z. XU, A. MOHANTY, S. VRUDHULA, S. YU, J. YE et coll. On-chip sparse learning acceleration with CMOS and resistive synaptic devices. *IEEE Transactions on Nanotechnology*, **14** : 969–979, 2015. (voir p. 29)

[79] T. GOKMEN et Y. VLASOV. Acceleration of Deep Neural Network Training with Resistive Cross-Point Devices : Design Considerations. *Frontiers in Neuroscience*, **10** : 2016. (voir p. 29, 180)

[80] P. GYSEL, M. MOTAMEDI et S. GHIASI. Hardware-oriented Approximation of Convolutional Neural Networks. *arXiv preprint arXiv :1604.03168*, 2016. (voir p. 30, 31)

[81] Z. LIN, M. COURBARIAUX, R. MEMISEVIC et Y. BENGIO. Neural networks with few multiplications. *arXiv preprint arXiv :1510.03009*, 2015. (voir p. 30–32)

[82] M. KIM et P. SMARAGDIS. Bitwise neural networks. *arXiv preprint arXiv :1601.06071*, 2016. (voir p. 30–32)

[83] M. COURBARIAUX, I. HUBARA, D. SOUDRY, R. EL-YANIV et Y. BENGIO. Binarized Neural Networks : Training Deep Neural Networks with Weights and Activations Constrained to + 1 or - 1. *arXiv preprint arXiv : 1602.02830 v3*, 2016. (voir p. 30, 31)

[84] M. RASTEGARI, V. ORDONEZ, J. REDMON et A. FARHADI. XNOR-Net : ImageNet Classification Using Binary Convolutional Neural Networks. *arXiv preprint arXiv :1603.05279*, 2016. (voir p. 30, 31)

[85] M. HOROWITZ. "Computing's energy problem (and what we can do about it)" *IEEE International Solid-State Circuits Conference Digest of Technical Papers*. IEEE 2014. 10–14 (voir p. 30)

[86] S. K. ESSER, R. APPUSWAMY, P. MEROLLA, J. V. ARTHUR et D. S. MODHA. "Backpropagation for energy-efficient neuromorphic computing" *Advances in Neural Information Processing Systems*. 2015. 1117–1125 (voir p. 31, 33, 44)

[87] W. SENN et S. FUSI. Convergence of stochastic learning in perceptrons with binary synapses. *Physical Review E*, **71** : 061907, 2005. DOI : 10.1103/PhysRevE.71.061907 (voir p. 32, 104)

[88] M. SURI, D. QUERLIOZ, O. BICHLER, G. PALMA, E. VIANELLO, D. VUILLAUME, C. GAMRAT et B. DESALVO. Bio-inspired stochastic computing using binary CBRAM synapses. *IEEE Transactions on Electron Devices*, **60** : 2402–2409, 2013. (voir p. 32, 53, 70, 73, 100, 104, 105, 114)

[89] S. YU, B. GAO, Z. FANG, H. YU, J. KANG et H. P. WONG. Stochastic learning in oxide binary synaptic device for neuromorphic computing. *Frontiers in Neuroscience*, **7** : 2013. DOI : 10.3389/fnins.2013.00186 (voir p. 32, 51)

[90] S. W. MOORE, P. J. FOX, S. J. T. MARSH, A. T. MARKETTOS et A. MUJUMDAR. "Bluehive - A field-programable custom computing machine for extreme-scale real-time neural network simulation" *IEEE Annual International Symposium on Field-Programmable Custom Computing Machines*. 2012. 133–140 DOI : 10.1109/FCCM.2012.32 (voir p. 34, 45)

[91] E. M. IZHIKEVICH. Simple model of spiking neurons. *IEEE Transactions on Neural Networks*, **14** : 1569–1572, 2003. (voir p. 34)

[92] E. M. IZHIKEVICH. Which model to use for cortical spiking neurons ? *IEEE Transactions on Neural Networks*, **15** : 1063–1070, 2004. (voir p. 34)

[93] K. CHEUNG, S. R. SCHULTZ et W. LUK. NeuroFlow : A General Purpose Spiking Neural Network Simulation Platform using Customizable Processors. *Frontiers in Neuroscience*, **9** : 516, 2016. DOI : 10.3389/fnins.2015.00516 (voir p. 35, 45)

[94] A. DAVISON, D. BRÜDERLE, J. EPPLER, J. KREMKOW, E. MULLER, D. PECEVSKI, L. PERRINET et P. YGER. PyNN : a common interface for neuronal network simulators. *Frontiers in Neuroinformatics*, **2** : 11, 2009. DOI : 10.3389/neuro.11.011.2008 (voir p. 35, 43)

[95] R. BRETTE et W. GERSTNER. Adaptive exponential integrate-and-fire model as an effective description of neuronal activity. *Journal of Neurophysiology*, **94** : 3637–3642, 2005. DOI : 10.1152/jn.00686.2005 (voir p. 35, 41, 42, 46)

[96] S. B. FURBER, F. GALLUPPI, S. TEMPLE et L. A. PLANA. The SpiNNaker Project. *Proceedings of the IEEE*, **102** : 652–665, 2014. DOI : 10.1109/JPROC.2014.2304638 (voir p. 36, 45, 48)

[97] E. STROMATIAS, F. GALLUPPI, C. PATTERSON et S. B. FURBER. "Power analysis of large-scale, real-time neural networks on SpiNNaker" *International Joint Conference on Neural Networks*. 2013. 1–8 DOI : 10.1109/IJCNN.2013.6706927 (voir p. 37)

[98] F. GALLUPPI, K. BROHAN, S. DAVIDSON, T. SERRANO-GOTARREDONA, J.-A. P. CARRASCO, B. LINARES-BARRANCO et S. B. FURBER. *A Real-Time, Event-Driven Neuromorphic System for Goal-Directed Attentional Selection*. sous la dir. de T. HUANG, Z. ZENG, C. LI et C. S. LEUNG Berlin, Heidelberg : Springer Berlin Heidelberg, 2012. 226–233 DOI : 10.1007/978-3-642-34481-7_28 (voir p. 37)

[99] G. ORCHARD, X. LAGORCE, C. POSCH, S. B. FURBER, R. BENOSMAN et F. GALLUPPI. "Real-time event-driven spiking neural network object recognition on the SpiNNaker platform" *IEEE International Symposium on Circuits and Systems*. 2015. 2413–2416 DOI : 10.1109/ISCAS.2015.7169171 (voir p. 37)

[100] R. SARPESHKAR. Analog Versus Digital : Extrapolating from Electronics to Neurobiology. *Neural Computation*, **10** : 1601–1638, 1998. DOI : 10.1162/089976698300017052 (voir p. 38)

[101] B. V. BENJAMIN, P. GAO, E. MCQUINN, S. CHOUDHARY, A. R. CHANDRASEKARAN, J. M. BUSSAT, R. ALVAREZ-ICAZA, J. V. ARTHUR, P. A. MEROLLA et K. BOAHEN. Neurogrid : A Mixed-Analog-Digital Multichip System for Large-Scale Neural Simulations. *Proceedings of the IEEE*, **102** : 699–716, 2014. DOI : 10.1109/JPROC.2014.2313565 (voir p. 38, 45)

[102] T. YU, J. PARK, S. JOSHI, C. MAIER et G. CAUWENBERGHS. "65k-neuron integrate-and-fire array transceiver with address-event reconfigurable synaptic routing" *IEEE Biomedical Circuits and Systems Conference*. 2012. 21–24 DOI : 10.1109/BioCAS.2012.6418479 (voir p. 39)

[103] T. YU et G. CAUWENBERGHS. "Log-Domain Time-Multiplexed Realization of Dynamical Conductance-Based Synapses" *IEEE International Symposium on Circuits and Systems*. 2010. 2558–2561 DOI : 10.1109/ISCAS.2010.5537114 (voir p. 39)

[104] J. PARK, T. YU, S. JOSHI, C. MAIER et G. CAUWENBERGHS. Hierarchical Address Event Routing for Reconfigurable Large-Scale Neuromorphic Systems. *IEEE Transactions on Neural Networks and Learning Systems*, **PP** : 1–15, 2016. DOI : 10.1109/TNNLS.2016.2572164 (voir p. 39, 45)

[105] J. M. BRADER, W. SENN et S. FUSI. Learning real-world stimuli in a neural network with spike-driven synaptic dynamics. *Neural computation*, **19** : 2881–2912, 2007. DOI : 10.1162/neco.2007.19.11.2881 (voir p. 41)

[106] J. SCHEMMEL, D. BRIIDERLE, A. GRIIBL, M. HOCK, K. MEIER et S. MILLNER. "A wafer-scale neuromorphic hardware system for large-scale neural modeling" *IEEE International Symposium on Circuits and Systems*. 2010. 1947–1950 DOI : 10.1109/ISCAS.2010.5536970 (voir p. 41, 45, 48)

[107] J. SCHEMMEL, A. GRUBL, K. MEIER et E. MUELLER. "Implementing Synaptic Plasticity in a VLSI Spiking Neural Network Model" *IEEE International Joint Conference on Neural Network Proceedings*. 2006. 1–6 DOI : 10.1109/IJCNN.2006.246651 (voir p. 42)

[108] J. SCHEMMEL, D. BRUDERLE, K. MEIER et B. OSTENDORF. "Modeling Synaptic Plasticity within Networks of Highly Accelerated I F Neurons" *IEEE International Symposium on Circuits and Systems*. 2007. 3367–3370 DOI : 10.1109/ISCAS.2007.378289 (voir p. 42)

[109] P. MEROLLA, J. ARTHUR, F. AKOPYAN, N. IMAM, R. MANOHAR et D. S. MODHA. "A digital neurosynaptic core using embedded crossbar memory with 45pJ per spike in 45nm" *IEEE Custom Integrated Circuits Conference*. 2011. 1–4 DOI : 10.1109/CICC.2011.6055294 (voir p. 43)

[110] A. S. CASSIDY, P. MEROLLA, J. V. ARTHUR, S. K. ESSER, B. JACKSON, R. ALVAREZ-ICAZA, P. DATTA, J. SAWADA, T. M. WONG, V. FELDMAN et coll. "Cognitive computing building block : A versatile and efficient digital neuron model for neurosynaptic cores" *International Joint Conference on Neural Networks*. IEEE 2013. 1–10 (voir p. 43)

[111] L. DENG. The MNIST database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 2012. DOI : 10.1109/MSP.2012.2211477 (voir p. 44)

[112] E. STROMATIAS, D. NEIL, F. GALLUPPI, M. PFEIFFER, S.-C. LIU et S. FURBER. "Scalable energy-efficient, low-latency implementations of trained spiking deep belief networks on SpiNNaker" *International Joint Conference on Neural Networks*. IEEE 2015. 1–8 DOI : 10.1109/IJCNN.2015.7280625 (voir p. 44)

[113] S. K. ESSER, P. A. MEROLLA, J. V. ARTHUR, A. S. CASSIDY, R. APPUSWAMY, A. ANDREOPOULOS, D. J. BERG, J. L. MCKINSTRY, T. MELANO, D. R. BARCH et coll. Convolutional Networks for Fast, Energy-Efficient Neuromorphic Computing. *arXiv preprint arXiv :1603.08270*, 2016. (voir p. 44)

[114] A. AMIR, P. DATTA, W. P. RISK, A. S. CASSIDY, J. A. KUSNITZ, S. K. ESSER, A. ANDREOPOULOS, T. M. WONG, M. FLICKNER, R. ALVAREZ-ICAZA, E. MCQUINN, B. SHAW, N. PASS et D. S. MODHA. "Cognitive computing programming paradigm : A Corelet Language for composing networks of neurosynaptic cores" *International Joint Conference on Neural Networks*. 2013. 1–10 DOI : 10.1109/IJCNN.2013.6707078 (voir p. 44)

[115] N. QIAO, H. MOSTAFA, F. CORRADI, M. OSSWALD, F. STEFANINI, D. SUMISLAWSKA et G. INDIVERI. A reconfigurable on-line learning spiking neuromorphic processor comprising 256 neurons and 128K synapses. *Frontiers in Neuroscience*, **9** : 141, 2015. DOI : 10.3389/fnins.2015.00141 (voir p. 45, 48)

[116] D. B. STRUKOV, G. S. SNIDER, D. R. STEWART et R. S. WILLIAMS. The missing memristor found. *Nature*, **453** : 80–83, 2008. DOI : 10.1038/nature06932 (voir p. 48)

[117] L. CHUA. Memristor-The missing circuit element. *IEEE Transactions on Circuit Theory*, **18** : 507–519, 1971. DOI : 10.1109/TCT.1971.1083337 (voir p. 48)

[118] L. CHUA. Resistance switching memories are memristors. *Applied Physics A*, **102** : 765–783, 2011. DOI : 10.1007/s00339-011-6264-9 (voir p. 49)

[119] B. LINARES-BARRANCO et T. SERRANO-GOTARREDONA. Memristance can explain Spike-Time-Dependent-Plasticity in Neural Synapses. *Nature Precedings*, 2009. DOI : 10.1038/npre.2009.3010.1 (voir p. 49)

[120] T. SERRANO-GOTARREDONA, T. MASQUELIER, T. PRODROMAKIS, G. INDIVERI et B. LINARES-BARRANCO. STDP and STDP variations with memristors for spiking neuromorphic learning systems. *Frontiers in Neuroscience*, **7** : 2, 2013. DOI : 10.3389/fnins.2013.00002 (voir p. 49, 105, 125, 156)

[121] R. WASER et M. AONO. Nanoionics-based resistive switching memories. *Nature materials*, **6** : 833–840, 2007. DOI : 10.1038/nmat2023 (voir p. 51)

[122] H. S. P. WONG, H. Y. LEE, S. YU, Y. S. CHEN, Y. WU, P. S. CHEN, B. LEE, F. T. CHEN et M. J. TSAI. Metal-Oxide RRAM. *Proceedings of the IEEE*, **100** : 1951–1970, 2012. DOI : 10.1109/JPROC.2012.2190369 (voir p. 51)

[123] B. PRINCE. *Vertical 3D memory technologies*. John Wiley & Sons, 2014. (voir p. 51)

[124] S. YU, Y. WU, R. JEYASINGH, D. KUZUM et H. P. WONG. An Electronic Synapse Device Based on Metal Oxide Resistive Switching Memory for Neuromorphic Computation. *IEEE Transactions on Electron Devices*, **58** : 2729–2737, 2011. DOI : 10.1109/TED.2011.2147791 (voir p. 51, 104)

[125] F. ALIBART, E. ZAMANIDOOST et D. B. STRUKOV. Pattern classification by memristive crossbar circuits using ex situ and in situ training. *Nature Communications*, **4** : 2013. DOI : 10.1038/ncomms3072 (voir p. 52)

[126] 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*, **521** : 61–64, 2015. (voir p. 52)

[127] L. GAO, I.-T. WANG, P.-Y. CHEN, S. VRUDHULA, J.-S. SEO, Y. CAO, T.-H. HOU et S. YU. Fully parallel write/read in resistive synaptic array for accelerating on-chip learning. *Nanotechnology*, **26** : 455204, 2015. (voir p. 52)

[128] R. WASER, R. DITTMANN, G. STAIKOV et K. SZOT. Redox-based resistive switching memories–nanoionic mechanisms, prospects, and challenges. *Advanced Materials*, **21** : 2632–2663, 2009. (voir p. 52, 125)

[129] J. ZAHURAK, K. MIYATA, M. FISCHER, M. BALAKRISHNAN, S. CHHAJED, D. WELLS, H. LI, A. TORSI, J. LIM, M. KORBER, K. NAKAZAWA, S. MAYUZUMI, M. HONDA, S. SILLS, S. YASUDA, A. CALDERONI, B. COOK, G. DAMARLA, H. TRAN, B. WANG, C. CARDON, K. KARDA, J. OKUNO, A. JOHNSON, T. KUNIHIRO, J. SUMINO, M. TSUKAMOTO, K. ARATANI, N. RAMASWAMY, W. OTSUKA et K. PRALL. "Process integration of a 27nm, 16Gb Cu ReRAM" *IEEE International Electron Devices Meeting*. 2014. 6.2.1–6.2.4 DOI : 10.1109/IEDM.2014.7046994 (voir p. 52)

[130] S. H. JO, T. CHANG, I. EBONG, B. B. BHADVIYA, P. MAZUMDER et W. LU. Nanoscale Memristor Device as Synapse in Neuromorphic Systems. *Nano Letters*, **10** : PMID : 20192230, 1297–1301, 2010. DOI : 10.1021/nl904092h eprint : http://pubs.acs.org/doi/pdf/10.1021/nl904092h (voir p. 52, 104, 125)

[131] C. DU, W. MA, T. CHANG, P. SHERIDAN et W. D. LU. Biorealistic Implementation of Synaptic Functions with Oxide Memristors through Internal Ionic Dynamics. *Advanced Functional Materials*, **25** : 4290–4299, 2015. (voir p. 52, 125, 136)

[132] M. SURI, O. BICHLER, D. QUERLIOZ, G. PALMA, E. VIANELLO, D. VUILLAUME, C. GAMRAT et B. DESALVO. "CBRAM devices as binary synapses for low-power stochastic neuromorphic systems : auditory (cochlea) and visual (retina) cognitive processing applications" *IEEE International Electron Devices Meeting*. IEEE 2012. 10–3 (voir p. 53, 100, 105)

[133] S. LA BARBERA, D. VUILLAUME et F. ALIBART. Filamentary Switching : Synaptic Plasticity through Device Volatility. *ACS Nano*, **9** : PMID : 25581249, 941–949, 2015. DOI : 10.1021/nn506735m eprint : 10.1021/nn506735m (voir p. 53, 125, 126, 128, 131, 157)

[134] S. RAOUX, G. W. BURR, M. J. BREITWISCH, C. T. RETTNER, Y. C. CHEN, R. M. SHELBY, M. SALINGA, D. KREBS, S. H. CHEN, H. L. LUNG et C. H. LAM. Phase-change random access memory : A scalable technology. *IBM Journal of Research and Development*, **52** : 465–479, 2008. DOI : 10.1147/rd.524.0465 (voir p. 53)

[135] G. W. BURR, M. J. BREITWISCH, M. FRANCESCHINI, D. GARETTO, K. GOPALAKRISHNAN, B. JACKSON, B. KURDI, C. LAM, L. A. LASTRAS, A. PADILLA, B. RAJENDRAN, S. RAOUX et R. S. SHENOY. Phase change memory technology. *Journal of Vacuum Science & Technology B*, **28** : 223–262, 2010. DOI : 10.1116/1.3301579 (voir p. 53)

[136] H. S. P. WONG, S. RAOUX, S. KIM, J. LIANG, J. P. REIFENBERG, B. RAJENDRAN, M. ASHEGHI et K. E. GOODSON. Phase Change Memory. *Proceedings of the IEEE*, **98** : 2201–2227, 2010. DOI : 10.1109/JPROC.2010.2070050 (voir p. 53)

[137] O. BICHLER, M. SURI, D. QUERLIOZ, D. VUILLAUME, B. DESALVO et C. GAMRAT. Visual Pattern Extraction Using Energy-Efficient 2-PCM Synapse Neuromorphic Architecture. *IEEE Transactions on Electron Devices*, **59** : 2206–2214, 2012. DOI : 10.1109/TED.2012.2197951 (voir p. 54, 100)

[138] M. SURI, O. BICHLER, D. QUERLIOZ, O. CUETO, L. PERNIOLA, V. SOUSA, D. VUILLAUME, C. GAMRAT et B. DESALVO. "Phase change memory as synapse for ultra-dense neuromorphic systems : Application to complex visual pattern extraction" *International Electron Devices Meeting*. IEEE 2011. 4–4 (voir p. 54, 100, 104)

[139] G. W. BURR, R. M. SHELBY, S. SIDLER, C. DI NOLFO, J. JANG, I. BOYBAT, R. S. SHENOY, P. NARAYANAN, K. VIRWANI, E. U. GIACOMETTI et coll. Experimental demonstration and tolerancing of a large-scale neural network (165 000 Synapses) using phase-change memory as the synaptic weight element. *IEEE Transactions on Electron Devices*, **62** : 3498–3507, 2015. (voir p. 54)

[140] M. Y. ZHURAVLEV, R. F. SABIRIANOV, S. S. JASWAL et E. Y. TSYMBAL. Giant Electroresistance in Ferroelectric Tunnel Junctions. *Physical Review Letters*, **94** : 246802, 2005. DOI : 10.1103/PhysRevLett.94.246802 (voir p. 54)

[141] A. GRUVERMAN, D. WU, H. LU, Y. WANG, H. W. JANG, C. M. FOLKMAN, M. Y. ZHURAVLEV, D. FELKER, M. RZCHOWSKI, C.-B. EOM et coll. Tunneling electroresistance effect in ferroelectric tunnel junctions at the nanoscale. *Nano Letters*, **9** : 3539–3543, 2009. (voir p. 54)

[142] A. CHANTHBOUALA, A. CRASSOUS, V. GARCIA, K. BOUZEHOUANE, S. FUSIL, X. MOYA, J. ALLIBE, B. DLUBAK, J. GROLLIER, S. XAVIER et coll. Solid-state memories based on ferroelectric tunnel junctions. *Nature nanotechnology*, **7** : 101–104, 2012. (voir p. 54, 55)

[143] A. CHANTHBOUALA, V. GARCIA, R. O. CHERIFI, K. BOUZEHOUANE, S. FUSIL, X. MOYA, S. XAVIER, H. YAMADA, C. DERANLOT, N. D. MATHUR et coll. A ferroelectric memristor. *Nature materials*, **11** : 860–864, 2012. (voir p. 54)

[144] G. LECERF, J. TOMAS, S. BOYN, S. GIROD, A. MANGALORE, J. GROLLIER et S. SAÏGHI. "Silicon neuron dedicated to memristive spiking neural networks" *IEEE International Symposium on Circuits and Systems*. IEEE 2014. 1568–1571 (voir p. 54, 145)

[145] D. S. JEONG, R. THOMAS, R. S. KATIYAR, J. F. SCOTT, H. KOHLSTEDT, A. PETRARU et C. S. HWANG. Emerging memories : resistive switching mechanisms and current status. *Reports on Progress in Physics*, **75** : 076502, 2012. (voir p. 55)

[146] V. GARCIA, S. FUSIL, K. BOUZEHOUANE, S. ENOUZ-VEDRENNE, N. D. MATHUR, A. BARTHELEMY et M. BIBES. Giant tunnel electroresistance for non-destructive readout of ferroelectric states. *Nature*, **460** : 81–84, 2009. (voir p. 55)

[147] J. C. SLONCZEWSKI. Current-driven excitation of magnetic multilayers. *Journal of Magnetism and Magnetic Materials*, **159** : L1 –L7, 1996. DOI : 10.1016/0304-8853(96)00062-5 (voir p. 55, 74)

[148] L. BERGER. Current-induced oscillations of a Bloch wall in magnetic thin films. *Journal of Magnetism and Magnetic Materials*, **162** : 155 –161, 1996. DOI : 10.1016/S0304-8853(96)00286-7 (voir p. 55)

[149] N. D. RIZZO, D. HOUSSAMEDDINE, J. JANESKY, R. WHIG, F. B. MANCOFF, M. L. SCHNEIDER, M. DEHERRERA, J. SUN, K. NAGEL, S. DESHPANDE et coll. A Fully Functional 64 Mb DDR3 ST-MRAM Built on 90 nm CMOS Technology. *IEEE Transactions on Magnetics*, **49** : 4441–4446, 2013. (voir p. 55, 68, 71)

[150] X. FONG, Y. KIM, R. VENKATESAN, S. H. CHODAY, A. RAGHUNATHAN et K. ROY. Spin-Transfer Torque Memories : Devices, Circuits, and Systems. *Proceedings of the IEEE*, **PP** : 1–40, 2016. DOI : 10.1109/JPROC.2016.2521712 (voir p. 55, 68, 78)

[151] C. AUGUSTINE, N. N. MOJUMDER, X. FONG, S. H. CHODAY, S. P. PARK et K. ROY. Spin-transfer torque MRAMs for low power memories : Perspective and prospective. *IEEE Sensors Journal*, **12** : 756–766, 2012. DOI : 10.1109/JSEN.2011.2124453 (voir p. 61)

[152] S. YUASA, T. NAGAHAMA, A. FUKUSHIMA, Y. SUZUKI et K. ANDO. Giant room-temperature magnetoresistance in single-crystal Fe/MgO/Fe magnetic tunnel junctions. *Nature materials*, **3** : 868–871, 2004. DOI : 10.1038/nmat1257 (voir p. 62, 64, 67)

[153] M. JULLIÈRE. Tunneling between ferromagnetic films. *Physics Letters A*, **54** : 225–226, 1975. (voir p. 62, 64)

[154] S. IKEDA, J. HAYAKAWA, Y. ASHIZAWA, Y. M. LEE, K. MIURA, H. HASEGAWA, M. TSUNODA, F. MATSUKURA et H. OHNO. Tunnel magnetoresistance of 604% at 300 K by suppression of Ta diffusion in CoFeB/MgO/CoFeB pseudo-spin-valves annealed at high temperature. *Applied Physics Letters*, **93** : 2508, 2008. (voir p. 64)

[155] A. BARTHELEMY, A. FERT, J.-P. CONTOUR, M. BOWEN, V. CROS, J.-M. DE TERESA, A. HAMZIC, J. C. FAINI, J.-M. GEORGE, J. GROLLIER et coll. Magnetoresistance and spin electronics. *Journal of Magnetism and Magnetic Materials*, **242** : 68–76, 2002. (voir p. 64)

[156] T. MCGUIRE et R. L. POTTER. Anisotropic magnetoresistance in ferromagnetic 3d alloys. *IEEE Transactions on Magnetics*, **11** : 1018–1038, 1975. (voir p. 64)

[157] Y. ZHANG, W. S. ZHAO, Y. LAKYS, J.-O. KLEIN, J.-V. KIM, D. RAVELOSONA et C. CHAPPERT. Compact modeling of perpendicular-anisotropy CoFeB/MgO magnetic tunnel junctions. *IEEE Transactions on Electron Devices*, **59** : 819–826, 2012. DOI : 10.1109/TED.2011.2178416 (voir p. 64, 65, 72, 73, 78)

[158] S. ZHANG, P. M. LEVY, A. C. MARLEY et S. PARKIN. Quenching of magnetoresistance by hot electrons in magnetic tunnel junctions. *Physical Review Letters*, **79** : 3744, 1997. DOI : 10.1103/PhysRevLett.79.3744 (voir p. 64)

[159] D. C. RALPH et M. D. STILES. Spin Transfer Torques. *ArXiv e-prints*, 2007. arXiv : 0711.4608 (voir p. 67)

[160] Y. GANG, W. S. ZHAO, J.-O. KLEIN, C. CHAPPERT et P. MAZOYER. A high-reliability, low-power magnetic full adder. *IEEE Transactions on Magnetics*, **47** : 4611–4616, 2011. (voir p. 68)

[161] S. SENNI, L. TORRES, G. SASSATELLI, A. GAMATIE et B. MUSSARD. Exploring MRAM Technologies for Energy Efficient Systems-On-Chip. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, **6** : 279–292, 2016. DOI : 10.1109/JETCAS.2016.2547680 (voir p. 68)

[162] E. KITAGAWA, S. FUJITA, K. NOMURA, H. NOGUCHI, K. ABE, K. IKEGAMI, T. DAIBOU, Y. KATO, C. KAMATA, S. KASHIWADA, N. SHIMOMURA, J. ITO et H. YODA. "Impact of ultra low power and fast write operation of advanced perpendicular MTJ on power reduction for high-performance mobile CPU" *IEEE International Electron Devices Meeting*. 2012. 29.4.1–29.4.4 DOI : 10.1109/IEDM.2012.6479129 (voir p. 68, 72)

[163] W. S. ZHAO, M. MOREAU, E. DENG, Y. ZHANG, J.-M. PORTAL, J.-O. KLEIN, M. BOCQUET, H. AZIZA, D. DELERUYELLE, C. MULLER et coll. Synchronous non-volatile logic gate design based on resistive switching memories. *IEEE Transactions on Circuits and Systems I : Regular Papers*, **61** : 443–454, 2014. (voir p. 68)

[164] B. JOVANOVIĆ, R. M. BRUM et L. TORRES. Evaluation of hybrid MRAM/CMOS cells for "normally-off and instant-on" computing. *Analog Integrated Circuits and Signal Processing*, **81** : 607–621, 2014. DOI : 10.1007/s10470-014-0427-5 (voir p. 69)

[165] W. S. ZHAO, E. BELHAIRE, C. CHAPPERT et P. MAZOYER. Spin transfer torque (STT)-MRAM–based runtime reconfiguration FPGA circuit. *ACM Transactions on Embedded Computing Systems*, **9** : 14, 2009. (voir p. 69)

[166] N. LOCATELLI, V. CROS et J. GROLLIER. Spin-torque building blocks. *Natural Materials*, **13** : 11–20, 2014. DOI : 10.1038/nmat3823 (voir p. 69)

[167] Y. HIGO, K. YAMANE, K. OHBA, H. NARISAWA, K. BESSHO, M. HOSOMI et H. KANO. Thermal activation effect on spin transfer switching in magnetic tunnel junctions. *Applied Physics Letters*, **87** : 082502, 082502, 2005. DOI : 10.1063/1.2011795 (voir p. 69, 90)

[168] R. HEINDL, W. H. RIPPARD, S. E. RUSSEK, M. R. PUFALL et A. B. KOS. Validity of the thermal activation model for spin-transfer torque switching in magnetic tunnel junctions. *Journal of Applied Physics*, **109** : 073910–073910, 2011. (voir p. 69)

[169] A. FUKUSHIMA, T. SEKI, K. YAKUSHIJI, H. KUBOTA, H. IMAMURA, S. YUASA et K. ANDO. Spin dice : A scalable truly random number generator based on spintronics. *Applied Physics Express*, **7** : 083001, 2014. (voir p. 69)

[170] A. FUKUSHIMA, K. YAKUSHIJI, H. KUBOTA et S. YUASA. "Spin dice (physical random number generator using spin torque switching) and its thermal response" *IEEE Magnetics Conference*. IEEE 2015. 1–1 (voir p. 69)

[171] J. S. FRIEDMAN, L. E. CALVET, P. BESSIÈRE, J. DROULEZ et D. QUERLIOZ. Bayesian Inference With Muller C-Elements. *IEEE Transactions on Circuits and Systems I : Regular Papers*, **63** : 895–904, 2016. DOI : 10.1109/TCSI.2016.2546064 (voir p. 69)

[172] T. DEVOLDER, J. HAYAKAWA, K. ITO, H. TAKAHASHI, S. IKEDA, P. CROZAT, N. ZEROUNIAN, J.-V. KIM, C. CHAPPERT et H. OHNO. Single-shot time-resolved measurements of nanosecond-scale spin-transfer induced switching : Stochastic versus deterministic aspects. *Physical review letters*, **100** : 057206, 2008. (voir p. 70, 85)

[173] M. Marins de CASTRO, R. C. SOUSA, S. BANDIERA, C. DUCRUET, A. CHAVENT, S. AUFFRET, C. PAPUSOI, I. L. PREJBEANU, C. PORTEMONT, L. VILA et coll. Precessional spin-transfer switching in a magnetic tunnel junction with a synthetic antiferromagnetic perpendicular polarizer. *Journal of Applied Physics*, **111** : 07C912–07C912, 2012. (voir p. 70)

[174] A. MIZRAHI, N. LOCATELLI, R. LEBRUN, V. CROS, A. FUKUSHIMA, H. KUBOTA, S. YUASA, D. QUERLIOZ et J. GROLLIER. Controlling the phase locking of unstable magnetic bits for ultra-low power computation. *arXiv preprint arXiv :1605.06932*, 2016. (voir p. 70)

[175] S. IKEDA, J. HAYAKAWA, Y. M. LEE, R. SASAKI, T. MEGURO, F. MATSUKURA et H. OHNO. Dependence of Tunnel Magnetoresistance in MgO Based Magnetic Tunnel Junctions on Ar Pressure during MgO Sputtering. *Japanese Journal of Applied Physics*, **44** : L1442–L1445, 2005. DOI : 10.1143/jjap.44.l1442 (voir p. 70)

[176] D. DJAYAPRAWIRA, K. TSUNEKAWA, M. NAGAI, H. MAEHARA, S. YAMAGATA, N. WATANABE, S. YUASA, Y. SUZUKI et K. ANDO. 230% room-temperature magnetoresistance in CoFeB/ MgO/ CoFeB magnetic tunnel junctions. *Applied Physics Letters*, **86** : 092502, 2005. DOI : 10.1063/1.1871344 (voir p. 70)

[177] J.-G. ZHU et C. PARK. Magnetic tunnel junctions. *Materials Today*, **9** : 36 –45, 2006. DOI : 10.1016/S1369-7021(06)71693-5 (voir p. 70)

[178] N. LOCATELLI, A. MIZRAHI, A. ACCIOLY, R. MATSUMOTO, A. FUKUSHIMA, H. KUBOTA, S. YUASA, V. CROS, L. G. PEREIRA, D. QUERLIOZ et coll. Noise-enhanced synchronization of stochastic magnetic oscillators. *Physical Review Applied*, **2** : 034009, 2014. DOI : 10.1103/PhysRevApplied.2.034009 (voir p. 71)

[179] S. IKEDA, K. MIURA, H. YAMAMOTO, K. MIZUNUMA, H. D. GAN, M. ENDO, S. L. KANAI, J. HAYAKAWA, F. MATSUKURA et H. OHNO. A perpendicular-anisotropy CoFeB-MgO magnetic tunnel junction. *Nature materials*, **9** : 721–724, 2010. (voir p. 71)

[180] K. IKEGAMI, H. NOGUCHI, C. KAMATA, M. AMANO, K. ABE, K. KUSHIDA, E. KITAGAWA, T. OCHIAI, N. SHIMOMURA, S. ITAI, D. SAIDA, C. TANAKA, A. KAWASUMI, H. HARA, J. ITO et S. FUJITA. "Low power and high density STT-MRAM for embedded cache memory using advanced perpendicular MTJ integrations and asymmetric compensation techniques" *IEEE International Electron Devices Meeting*. 2014. 28.1.1–28.1.4 DOI : 10.1109/IEDM.2014.7047123 (voir p. 71, 77)

[181] T. ANDRE, S. M. ALAM, D. GOGL, C. K. SUBRAMANIAN, H. LIN, W. MEADOWS, X. ZHANG, N. D. RIZZO, J. JANESKY, D. HOUSSAMEDDINE et J. M. SLAUGHTER. "ST-MRAM fundamentals, challenges, and applications" *Custom Integrated Circuits Conference, IEEE*. 2013. 1–8 DOI : 10.1109/CICC.2013.6658449 (voir p. 71)

[182] R. BEACH, T. MIN, C. HORNG, Q. CHEN, P. SHERMAN, S. LE, S. YOUNG, K. YANG, H. YU, X. LU et coll. "A statistical study of magnetic tunnel junctions for high-density spin torque transfer-MRAM (STT-MRAM)" *IEEE International Electron Devices Meeting*. IEEE 2008. 1–4 (voir p. 71, 114, 116, 118, 120)

[183] S. CHUNG, K.-M. RHO, S.-D. KIM, H.-J. SUH, D.-J. KIM, H. J. KIM, S. H. LEE, J.-H. PARK, H.-M. HWANG, S.-M. HWANG et coll. "Fully integrated 54 nm STT-RAM with the smallest bit cell dimension for high density memory application" *IEEE International Electron Devices Meeting*. IEEE 2010. 12–7 (voir p. 71)

[184] D. C. WORLEDGE, G. HU, P. L. TROUILLOUD, D. W. ABRAHAM, S. BROWN, M. C. GAIDIS, J. NOWAK, E. J. O'SULLIVAN, R. P. ROBERTAZZI, J. Z. SUN et coll. "Switching distributions and write reliability of perpendicular spin torque MRAM" *IEEE International Electron Devices Meeting*. IEEE 2010. 12–5 (voir p. 71, 114, 116, 118, 120)

[185] H. NOGUCHI, K. KUSHIDA, K. IKEGAMI, K. ABE, E. KITAGAWA, S. KASHIWADA, C. KAMATA, A. KAWASUMI, H. HARA et S. FUJITA. "A 250-MHz 256b-I/O 1-Mb STT-MRAM with advanced perpendicular MTJ based dual cell for nonvolatile magnetic caches to reduce active power of processors" *Symposium on VLSI Technology*. 2013. C108–C109 (voir p. 72)

[186] Z. DIAO, Z. LI, S. WANG, Y. DING, A. PANCHULA, E. CHEN, L.-C. WANG et Y. HUAI. Spin-transfer torque switching in magnetic tunnel junctions and spin-transfer torque random access memory. *Journal of Physics : Condensed Matter*, **19** : 165209, 2007. (voir p. 72–75, 85, 87, 90)

[187] J. J. NOWAK, R. P. ROBERTAZZI, J. Z. SUN, G. HU, J. H. PARK, J. LEE, A. J. ANNUNZIATA, G. P. LAUER, R. KOTHANDARAMAN, E. J. O'SULLIVAN, P. L. TROUILLOUD, Y. KIM et D. C. WORLEDGE. Dependence of Voltage and Size on Write Error Rates in Spin-Transfer Torque Magnetic Random-Access Memory. *IEEE Magnetics Letters*, **7** : 1–4, 2016. DOI : 10.1109/LMAG.2016.2539256 (voir p. 72, 175)

[188] J. H. KIM, W. C. LIM, U. H. PI, J. M. LEE, W. K. KIM, J. H. KIM, K. W. KIM, Y. S. PARK, S. H. PARK, M. A. KANG et coll. "Verification on the extreme scalability of STT-MRAM without loss of thermal stability below 15 nm MTJ cell" *Symposium on VLSI Technology : Digest of Technical Papers*. IEEE 2014. 1–2 (voir p. 72, 78)

[189] L. THOMAS, G. JAN, J. ZHU, H. LIU, Y.-J. LEE, S. LE, R.-Y. TONG, K. PI, Y.-J. WANG, D. SHEN et coll. Perpendicular spin transfer torque magnetic random access memories with high spin torque efficiency and thermal stability for embedded applications. *Journal of Applied Physics*, **115** : 172615, 2014. (voir p. 72)

[190] W. S. ZHAO, C. CHAPPERT, V. JAVERLIAC et J.-P. NOZIERE. High Speed, High Stability and Low Power Sensing Amplifier for MTJ/CMOS Hybrid Logic Circuits. *IEEE Transactions on Magnetics*, **45** : 3784–3787, 2009. DOI : 10.1109/TMAG.2009.2024325 (voir p. 72, 155, 169)

[191] Y. LAKYS, W. S. ZHAO, T. DEVOLDER, Y. ZHANG, J.-O. KLEIN, D. RAVELOSONA et C. CHAPPERT. Self-Enabled Error-Free Switching Circuit for Spin Transfer Torque MRAM and Logic. *IEEE Transactions on Magnetics*, **48** : 2403–2406, 2012. (voir p. 72)

[192] S. LEE, S. LEE, H. SHIN et D. KIM. Advanced HSPICE Macromodel for Magnetic Tunnel Junction. *Japanese Journal of Applied Physics*, **44** : 2696, 2005. (voir p. 72)

[193] M. E. BARAJI, V. JAVERLIAC, W. GUO, G. PRENAT et B. DIENY. Dynamic compact model of thermally assisted switching magnetic tunnel junctions. *Journal of Applied Physics*, **106** : 123906, 123906, 2009. DOI : 10.1063/1.3259373 (voir p. 72, 73)

[194] W. GUO, G. PRENAT, V. JAVERLIAC, M. E. BARAJI, N. de MESTIER, C. BARADUC et B. DIENY. SPICE modelling of magnetic tunnel junctions written by spin-transfer torque. *Journal of Physics D : Applied Physics*, **43** : 215001, 2010. (voir p. 72, 73)

[195] S. LEE, H. LEE, S. KIM, S. LEE et H. SHIN. A novel macro-model for spin-transfer-torque based magnetic-tunnel-junction elements. *Solid-State Electronics*, **54** : 497 –503, 2010. DOI : 10.1016/j.sse.2010.01.002 (voir p. 72)

[196] W. S. ZHAO, J. DUVAL, J.-O. KLEIN et C. CHAPPERT. A compact model for magnetic tunnel junction (MTJ) switched by thermally assisted Spin transfer torque (TAS + STT). *Nanoscale Research Letters*, **6** : 1–4, 2011. DOI : 10.1186/1556-276X-6-368 (voir p. 72, 73)

[197] Y. ZHANG, W. S. ZHAO, G. PRENAT, T. DEVOLDER, J.-O. KLEIN, C. CHAPPERT, B. DIENY et D. RAVELOSONA. Electrical Modeling of Stochastic Spin Transfer Torque Writing in Magnetic Tunnel Junctions for Memory and Logic Applications. *IEEE Transactions on Magnetics*, **49** : 4375–4378, 2013. DOI : 10.1109/TMAG.2013.2242257 (voir p. 72, 73, 91)

[198] P. WANG, W. ZHANG, R. JOSHI, R. KANJ et Y. CHEN. "A thermal and process variation aware MTJ switching model and its applications in soft error analysis" *IEEE/ACM International Conference on Computer-Aided Design*. IEEE 2012. 720–727 (voir p. 72, 73)

[199] G. D. PANAGOPOULOS, C. AUGUSTINE et K. ROY. Physics-Based SPICE-Compatible Compact Model for Simulating Hybrid MTJ/CMOS Circuits. *IEEE Transactions on Electron Devices*, **60** : 2808–2814, 2013. (voir p. 72, 73)

[200] C. CHAPPERT, A. FERT et F. N. VAN DAU. The emergence of spin electronics in data storage. *Nature materials*, **6** : 813–823, 2007. (voir p. 72)

[201] T. KAWAHARA, K. ITO, R. TAKEMURA et H. OHNO. Spin-transfer torque RAM technology : Review and prospect. *Microelectronics Reliability*, **52** : Advances in non-volatile memory technology, 613 –627, 2012. DOI : 10.1016/j.microrel.2011.09.028 (voir p. 72)

[202] L. NÉEL. Théorie du traînage magnétique des ferromagnétiques en grains fins avec applications aux terres cuites. *Annales de géophysique*, **5** : 99–136, 1949. (voir p. 73, 83)

[203] W. F. BROWN. Thermal Fluctuations of a Single-Domain Particle. *Physical Review*, **130** : 1677–1686, 1963. DOI : 10.1103/PhysRev.130.1677 (voir p. 73, 74, 83)

[204] J. Z. SUN. Spin-current interaction with a monodomain magnetic body : A model study. *Physical Review B*, **62** : 570–578, 2000. DOI : 10.1103/PhysRevB.62.570 (voir p. 73, 76, 82, 85)

[205] D. BEDAU, H. LIU, J. Z. SUN, J. A. KATINE, E. E. FULLERTON, S. MANGIN et A. D. KENT. Spin-transfer pulse switching : From the dynamic to the thermally activated regime. *Applied Physics Letters*, **97** : 262502, 2010. DOI : 10.1063/1.3532960 (voir p. 73, 90)

[206] Y. ZHANG, W. S. ZHAO, J.-O. KLEIN, W. KANG, D. QUERLIOZ, C. CHAPPERT et D. RAVELOSONA. "Multi-level cell Spin Transfer Torque MRAM based on stochastic switching" *IEEE Conference on Nanotechnology*. 2013. 233–236 DOI : 10.1109/NANO.2013.6720849 (voir p. 73)

[207] W. WU, X. ZHU, S. KANG, K. YUEN et R. GILMORE. Probabilistically Programmed STT-MRAM. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, **2** : 42–51, 2012. DOI : 10.1109/JETCAS.2012.2187401 (voir p. 73)

[208] J. Z. SUN, R. P. ROBERTAZZI, J. NOWAK, P. L. TROUILLOUD, G. HU, D. W. ABRAHAM, M. C. GAIDIS, S. L. BROWN, E. J. O'SULLIVAN, W. J. GALLAGHER et D. C. WORLEDGE. Effect of subvolume excitation and spin-torque efficiency on magnetic switching. *Physical Review B*, **84** : 064413, 2011. DOI : 10.1103/PhysRevB.84.064413 (voir p. 74, 75, 95)

[209] J. L. GARCÍA-PALACIOS et F. J. LÁZARO. Langevin-dynamics study of the dynamical properties of small magnetic particles. *Physical Review B*, **58** : 14937–14958, 1998. DOI : 10.1103/PhysRevB.58.14937 (voir p. 74, 82, 86)

[210] M. D. STILES et J. MILTAT. "Spin-Transfer Torque and Dynamics" *Spin Dynamics in Confined Magnetic Structures III* sous la dir. de B. HILLEBRANDS et A. THIAVILLE. t. 101 Topics in Applied Physics Springer Berlin Heidelberg, 2006. 225–308 DOI : 10.1007/10938171_7 (voir p. 74, 91)

[211] J. Z. SUN. Spin angular momentum transfer in current-perpendicular nanomagnetic junctions. *IBM Journal of Research and Development*, **50** : 81–100, 2006. (voir p. 75, 76)

[212] R. MATSUMOTO, A. CHANTHBOUALA, J. GROLLIER, V. CROS, A. FERT, K. NISHIMURA, Y. NAGAMINE, H. MAEHARA, K. TSUNEKAWA, A. FUKUSHIMA et coll. Spin-torque diode measurements of MgO-based magnetic tunnel junctions with asymmetric electrodes. *Applied physics express*, **4** : 063001, 2011. (voir p. 75)

[213] A. V. KHVALKOVSKIY, D. APALKOV, S. WATTS, R. CHEPULSKII, R. S. BEACH, A. ONG, X. TANG, A. DRISKILL-SMITH, W. H. BUTLER, P. B. VISSCHER et coll. Basic principles of STT-MRAM cell operation in memory arrays. *Journal of Physics D : Applied Physics*, **46** : 74001–74020, 2013. (voir p. 77, 86)

[214] K. BERNERT, V. SLUKA, C. FOWLEY, J. LINDNER, J. FASSBENDER et A. M. DEAC. Phase diagrams of MgO magnetic tunnel junctions including the perpendicular spin-transfer torque in different geometries. *Physical Review B*, **89** : 134415, 2014. DOI : 10.1103/PhysRevB.89.134415 (voir p. 77)

[215] J. Z. SUN. "Spin-transfer torque switched magnetic tunnel junctions in magnetic random access memory" *Proceedings of the SPIE*. t. 9931 2016. 993113–993113–13 DOI : 10.1117/12.2238712 (voir p. 77)

[216] M. GAJEK, J. J. NOWAK, J. Z. SUN, P. L. TROUILLOUD, E. J. O'SULLIVAN, D. W. ABRAHAM, M. C. GAIDIS, G. HU, S. BROWN, Y. ZHU et coll. Spin torque switching of 20 nm magnetic tunnel junctions with perpendicular anisotropy. *Applied Physics Letters*, **100** : 132408, 2012. DOI : 10.1063/1.3694270 (voir p. 78)

[217] C. SERPICO, I.D. MAYERGOYZ et G. BERTOTTI. Numerical technique for integration of the Landau-Lifshitz equation. *Journal of Applied Physics*, **89** : 6991–6993, 2001. DOI : 10.1063/1.1358818 (voir p. 78)

[218] L. BANAS. "Numerical Methods for the Landau-Lifshitz-Gilbert Equation" *Numerical Analysis and Its Applications* sous la dir. de Z. LI, L. VULKOV et J. WANIEWSKI. t. 3401 Lecture Notes in Computer Science Springer Berlin Heidelberg, 2005. 158–165 DOI : 10.1007/978-3-540-31852-1_17 (voir p. 78)

[219] J. A. OSBORN. Demagnetizing factors of the general ellipsoid. *Physical Review*, **67** : 351–357, 1945. (voir p. 82)

[220] D. M. APALKOV et P. B. VISSCHER. Spin-torque switching : Fokker-Planck rate calculation. *Physical Review B*, **72** : 180405, 2005. DOI : 10.1103/PhysRevB.72.180405 (voir p. 82, 83)

[221] T. DEVOLDER. Scalability of magnetic random access memories based on an in-plane magnetized free layer. *Appl. Phys. Express*, **4** : 093001, 2011. (voir p. 82)

[222] Z. LI et S. ZHANG. Thermally assisted magnetization reversal in the presence of a spin-transfer torque. *Physical Review B*, **69** : 134416, 2004. DOI : 10.1103/PhysRevB.69.134416 (voir p. 83, 90)

[223] A. EINSTEIN. Investigations on the theory of the Brownian movement. *Annalen der Physik*, 1905. (voir p. 86)

[224] G. E. UHLENBECK et L. S. ORNSTEIN. On the Theory of the Brownian Motion. *Physical Review*, **36** : 823–841, 1930. DOI : 10.1103/PhysRev.36.823 (voir p. 86)

[225] H. TOMITA, T. NOZAKI, T. SEKI, T. NAGASE, K. NISHIYAMA, E. KITAGAWA, M. YOSHIKAWA, T. DAIBOU, M. NAGAMINE, T. KISHI, S. IKEGAWA, N. SHIMOMURA, H. YODA et Y. SUZUKI. High-Speed Spin-Transfer Switching in GMR Nano-Pillars With Perpendicular Anisotropy. *IEEE Transactions on Magnetics*, **47** : 1599–1602, 2011. DOI : 10.1109/TMAG.2011.2105860 (voir p. 91)

[226] H. ZHAO, Y. ZHANG, P. K. AMIRI, J. A. KATINE, J. LANGER, H. JIANG, I. N. KRIVOROTOV, K. L. WANG et J.-P. WANG. Spin-Torque Driven Switching Probability Density Function Asymmetry. *IEEE Transactions on Magnetics*, **48** : 3818–3820, 2012. (voir p. 91)

[227] W. S. ZHAO, Y. ZHANG, T. DEVOLDER, J.-O. KLEIN, D. RAVELOSONA, C. CHAPPERT et P. MAZOYER. Failure and reliability analysis of STT-MRAM. *Microelectronics Reliability*, **52** : 1848 –1852, 2012. DOI : 10.1016/j.microrel.2012.06.035 (voir p. 91)

[228] H.-K. CHO, K. P. BOWMAN et G. R. NORTH. A comparison of gamma and lognormal distributions for characterizing satellite rain rates from the tropical rainfall measuring mission. *Journal of Applied Meteorology*, **43** : 1586–1597, 2004. (voir p. 91)

[229] A. ABDI et M. KAVEH. "On the utility of gamma PDF in modeling shadow fading (slow fading)" *IEEE Vehicular Technology Conference*. t. 3 IEEE 1999. 2308–2312 DOI : 10.1109/VETEC.1999.778479 (voir p. 91)

[230] A. B. Z. SALEM et T. D. MOUNT. A convenient descriptive model of income distribution : the gamma density. *Econometrica : Journal of the Econometric Society*, 1115–1127, 1974. DOI : 10.2307/1914221 (voir p. 91)

[231] M. SHARAD, D. FAN et K. ROY. "Ultra Low Power Associative Computing with Spin Neurons and Resistive Crossbar Memory" *Proceedings of the Annual Design Automation Conference*. DAC '13 New York, NY, USA : ACM, 2013. 107 :1–107 :6 DOI : 10.1145/2463209.2488866 (voir p. 96)

[232] W. S. ZHAO, J.-O. KLEIN, Z. WANG, Y. ZHANG, N. BEN ROMDHANE, D. QUERLIOZ, D. RAVELOSONA et C. CHAPPERT. "Spin-electronics based logic fabrics" *IFIP/IEEE International Conference on Very Large Scale Integration*. 2013. 174–179 DOI : 10.1109/VLSI-SoC.2013.6673271 (voir p. 96)

[233] N. LOCATELLI, A. F. VINCENT, S. GALDIN-RETAILLEAU, J.-O. KLEIN et D. QUERLIOZ. "Approximate programming of magnetic memory elements for energy saving" *International Conference on Memristive Systems*. 2015. 1–2 DOI : 10.1109/MEMRISYS.2015.7378400 (voir p. 96)

[234] W. S. ZHAO, S. CHAUDHURI, C. ACCOTO, J.-O KLEIN, C. CHAPPERT et P. MAZOYER. Cross-Point Architecture for Spin-Transfer Torque Magnetic Random Access Memory. *IEEE Transactions on Nanotechnology*, **11** : 907–917, 2012. DOI : 10.1109/TNANO.2012.2206051 (voir p. 102, 160)

[235] G. LECERF, J. TOMAS et S. SAÏGHI. "Excitatory and Inhibitory Memristive Synapses for Spiking Neural Networks" *IEEE International Symposium on Circuits and Systems*. 2013. 1616–1619 DOI : 10.1109/ISCAS.2013.6572171 (voir p. 102, 104, 163)

[236] S. MITRA, S. FUSI et G. INDIVERI. Real-Time Classification of Complex Patterns Using Spike-Based Learning in Neuromorphic VLSI. *IEEE Transactions on Biomedical Circuits and Systems*, **3** : 32–42, 2009. DOI : 10.1109/TBCAS.2008.2005781 (voir p. 103)

[237] Y. WANG et S.-C. LIU. Active Processing of Spatio-Temporal Input Patterns in Silicon Dendrites. *IEEE Transactions on Biomedical Circuits and Systems*, **7** : 307–318, 2013. DOI : 10.1109/TBCAS.2012.2199487 (voir p. 103)

[238] J. ARTHUR et K. BOAHEN. Learning in silicon : Timing is everything. *Advances in neural information processing systems*, **18** : 75, 2006. (voir p. 103)

[239] H. MARKRAM, J. LUBKE, M. FROTSCHER et B. SAKMANN. Regulation of Synaptic Efficacy by Coincidence of Postsynaptic APs and EPSPs. *Science*, **275** : 213–215, 1997. DOI : 10.1126/science.275.5297.213 (voir p. 103, 125)

[240] G. S. SNIDER. Self-organized computation with unreliable, memristive nanodevices. *Nanotechnology*, **18** : 365202, 2007. DOI : 10.1088/0957-4484/18/36/365202 (voir p. 104)

[241] G. S. SNIDER. "Spike-timing-dependent learning in memristive nanodevices" *Proceedings of the IEEE International Symposium on Nanoscale Architectures*. 2008. 85–92 DOI : 10.1109/NANOARCH.2008.4585796 (voir p. 104)

[242] B. LINARES-BARRANCO et T. SERRANO-GOTARREDONA. "Exploiting memristance in adaptive asynchronous spiking neuromorphic nanotechnology systems" *Proceedings of the IEEE Conference on Nanotechnology*. 2009. 601–604 (voir p. 104)

[243] F. ALIBART, S. PLEUTIN, O. BICHLER, C. GAMRAT, T. SERRANO-GOTARREDONA, B. LINARES-BARRANCO et D. VUILLAUME. A Memristive Nanoparticle/Organic Hybrid Synapstor for Neuroinspired Computing. *Advanced Functional Materials*, **22** : 609–616, 2012. DOI : 10.1002/adfm.201101935 (voir p. 104, 125)

[244] K. SEO, I. KIM, S. JUNG, M. JO, S. PARK, J. PARK, J. SHIN, K. P. BIJU, J. KONG, K. LEE, B. LEE et H. HWANG. Analog memory and spike-timing-dependent plasticity characteristics of a nanoscale titanium oxide bilayer resistive switching device. *Nanotechnology*, **22** : 254023, 2011. DOI : 10.1088/0957-4484/22/25/254023 (voir p. 104)

[245] A. AFIFI, A. AYATOLLAHI et F. RAISSI. "Implementation of biologically plausible spiking neural network models on the memristor crossbar-based CMOS/nano circuits" *European Conference on Circuit Theory and Design*. 2009. 563–566 DOI : 10.1109/ECCTD.2009.5275035 (voir p. 104)

[246] D. QUERLIOZ, O. BICHLER et C. GAMRAT. "Simulation of a memristor-based spiking neural network immune to device variations" *International Joint Conference on Neural Networks*. IEEE 2011. 1775–1781 (voir p. 104)

[247] B. NESSLER, M. PFEIFFER, L. BUESING et W. MAASS. Bayesian Computation Emerges in Generic Cortical Microcircuits through Spike-Timing-Dependent Plasticity. *PLoS Computational Biology*, **9** : 2013. DOI : 10.1371/journal.pcbi.1003037 (voir p. 104, 105)

[248] D. QUERLIOZ, O. BICHLER, A. F. VINCENT et C. GAMRAT. Bioinspired Programming of Memory Devices for Implementing an Inference Engine. *Proceedings of the IEEE*, **103** : 1398–1416, 2015. DOI : 10.1109/JPROC.2015.2437616 (voir p. 104, 122, 193)

[249] *jAER datasets* 2005 (voir p. 108)

[250] O. BICHLER, D. QUERLIOZ, S. J. THORPE, J.-P. BOURGOIN et C. GAMRAT. Extraction of temporally correlated features from dynamic vision sensors with spike-timing-dependent plasticity. *Neural Networks*, **32** : 339–348, 2012. (voir p. 113)

[251] H. LIM, I. KIM, J.-S. KIM, C. S. HWANG et D. S. JEONG. Short-term memory of TiO2-based electrochemical capacitors : empirical analysis with adoption of a sliding threshold. *Nanotechnology*, **24** : 384005, 2013. (voir p. 124)

[252] L. F. ABBOTT, J. A. VARELA, K. SEN et S. B. NELSON. Synaptic depression and cortical gain control. *Science*, **275** : 221–224, 1997. (voir p. 124)

[253] R. BERDAN, C. LIM, A. KHIAT, C. PAPAVASSILIOU et T. PRODROMAKIS. A memristor SPICE model accounting for volatile characteristics of practical ReRAM. *IEEE Electron Device Letters*, **35** : 135–137, 2014. (voir p. 124)

[254] T. OHNO, T. HASEGAWA, T. TSURUOKA, K. TERABE, J. K. GIMZEWSKI et M. AONO. Short-term plasticity and long-term potentiation mimicked in single inorganic synapses. *Nature materials*, **10** : 591–595, 2011. (voir p. 125)

[255] T. CHANG, S.-H. JO et W. LU. Short-term memory to long-term memory transition in a nanoscale memristor. *ACS Nano*, **5** : 7669–7676, 2011. (voir p. 125)

[256] R. M. SHIFFRIN et R. C. ATKINSON. Storage and retrieval processes in long-term memory. *Psychological Review*, **76** : 179, 1969. (voir p. 125)

[257] D. KUZUM, R. G. D. JEYASINGH, B. LEE et H.-S. P. WONG. Nanoelectronic programmable synapses based on phase change materials for brain-inspired computing. *Nano letters*, **12** : 2179–2186, 2011. (voir p. 125)

[258] M. SURI, O. BICHLER, D. QUERLIOZ, B. TRAORÉ, O. CUETO, L. PERNIOLA, V. SOUSA, D. VUILLAUME, C. GAMRAT et B. DESALVO. Physical aspects of low power synapses based on phase change memory devices. *Journal of Applied Physics*, **112** : 054904, 2012. (voir p. 125)

[259] S. KIM, C. DU, P. SHERIDAN, W. MA, S. H. CHOI et W. D. LU. Experimental demonstration of a second-order memristor and its ability to biorealistically implement synaptic plasticity. *Nano letters*, **15** : 2203–2211, 2015. (voir p. 125)

[260] F. ZENKE, E. J. AGNES et W. GERSTNER. Diverse synaptic plasticity mechanisms orchestrated to form and retrieve memories in spiking neural networks. *Nature Communications*, **6** : 2015. (voir p. 125)

[261] G. CHECHIK, I. MEILIJSON et E. RUPPIN. Neuronal regulation : A mechanism for synaptic pruning during brain maturation. *Neural Computation*, **11** : 2061–2080, 1999. (voir p. 132)

[262] R. SETIONO. A penalty-function approach for pruning feedforward neural networks. *Neural Computation*, **9** : 185–204, 1997. (voir p. 132)

[263] L. F. ABBOTT et S. B. NELSON. Synaptic plasticity : taming the beast. *Nature Neuroscience*, **3** : 1178–1183, 2000. (voir p. 133)

[264] S. SAÏGHI, C. G. MAYR, T. SERRANO-GOTARREDONA, H. SCHMIDT, G. LECERF, J. TOMAS, J. GROLLIER, S. BOYN, A. F. VINCENT, D. QUERLIOZ et coll. Plasticity in memristive devices for spiking neural networks. *Frontiers in neuroscience*, **9** : 51, 2015. (voir p. 148)

[265] A. SERB, W. REDMAN-WHITE, C. PAPAVASSILIOU, R. BERDAN et T. PRODROMAKIS. "Limitations and precision requirements for read-out of passive, linear, selectorless RRAM arrays" *IEEE International Symposium on Circuits and Systems*. 2015. 189–192 DOI : 10.1109/ISCAS.2015.7168602 (voir p. 160)

[266] M. S. QURESHI, M. PICKETT, F. MIAO et J. P. STRACHAN. "CMOS interface circuits for reading and writing memristor crossbar array" *IEEE International Symposium on Circuits and Systems*. 2011. 2954–2957 DOI : 10.1109/ISCAS.2011.5938211 (voir p. 162)

[267] L. ZHANG, W. S. ZHAO, Y. ZHUANG, J. BAO, G. WANG, H. TANG, C. LI et B. XU. A 16 Kb Spin-Transfer Torque Random Access Memory With Self-Enable Switching and Precharge Sensing Schemes. *IEEE Transactions on Magnetics*, **50** : 1–7, 2014. DOI : 10.1109/TMAG.2013.2291222 (voir p. 162, 169)

[268] B. RAJENDRAN, Y. LIU, J.-S. SEO, K. GOPALAKRISHNAN, L. CHANG, D. J. FRIEDMAN et M. B. RITTER. Specifications of nanoscale devices and circuits for neuromorphic computational systems. *IEEE Transactions on Electron Devices*, **60** : 246–253, 2013. (voir p. 162)

[269] R. DORRANCE, F. REN, Y. TORIYAMA, A. A. HAFEZ, C. K. K. YANG et D. MARKOVIC. Scalability and Design-Space Analysis of a 1T-1MTJ Memory Cell for STT-RAMs. *IEEE Transactions on Electron Devices*, **59** : 878–887, 2012. DOI : 10.1109/TED.2011.2182053 (voir p. 168)

[270] S. A. WOLF, J. LU, M. R. STAN, E. CHEN et D. M. TREGER. The Promise of Nanomagnetics and Spintronics for Future Logic and Universal Memory. *Proceedings of the IEEE*, **98** : 2155–2168, 2010. DOI : 10.1109/JPROC.2010.2064150 (voir p. 168)

[271] U. K. KLOSTERMANN, M. ANGERBAUER, U. GRUNING, F. KREUPL, M. RUHRIG, F. DAHMANI, M. KUND et G. MULLER. "A Perpendicular Spin Torque Switching based MRAM for the 28 nm Technology Node" *IEEE International Electron Devices Meeting*. 2007. 187–190 DOI : 10.1109/IEDM.2007.4418898 (voir p. 168)

[272] S. YU, P. Y. CHEN, Y. CAO, L. XIA, Y. WANG et H. WU. "Scaling-up resistive synaptic arrays for neuro-inspired architecture : Challenges and prospect" *IEEE International Electron Devices Meeting*. 2015. 17.3.1–17.3.4 DOI : 10.1109/IEDM.2015.7409718 (voir p. 180)

[273] S. ZULOAGA, R. LIU, P. Y. CHEN et S. YU. "Scaling 2-layer RRAM cross-point array towards 10 nm node : A device-circuit co-design" *IEEE International Symposium on Circuits and Systems*. 2015. 193–196 DOI : 10.1109/ISCAS.2015.7168603 (voir p. 180)

[274] S. B. ERYILMAZ, D. KUZUM, S. YU et H. S. P. WONG. "Device and system level design considerations for analog-non-volatile-memory based neuromorphic architectures" *IEEE International Electron Devices Meeting*. 2015. 4.1.1–4.1.4 DOI : 10.1109/IEDM.2015.7409622 (voir p. 180)

[275] S. PARKIN, M. HAYASHI et L. THOMAS. Magnetic domain-wall racetrack memory. *Science*, **320** : 190–194, 2008. (voir p. 193)

[276] X. WANG, Y. CHEN, H. XI, H. LI et D. DIMITROV. Spintronic memristor through spin-torque-induced magnetization motion. *IEEE Electron Device Letters*, **30** : 294–297, 2009. (voir p. 193)

[277] S. LEQUEUX, J. SAMPAIO, V. CROS, K. YAKUSHIJI, A. FUKUSHIMA, R. MATSUMOTO, H. KUBOTA, S. YUASA et J. GROLLIER. Inside the perpendicular spin-torque memristor. *arXiv preprint arXiv :1605.07460*, 2016. (voir p. 193)

Listes de publications et d'interventions

Articles de journaux avec comité de lecture

✠ S. La Barbera[†], **A. F. Vincent**[†], D. Vuillaume, D. Querlioz et F. Alibart, « Interplay of Multiple Synaptic Plasticity Features in Filamentary Memristive Devices for Neuromorphic Computing », *Sci. Rep.*, 2016. [†] : co-premiers auteurs. *Article accepté.*

✠ **A. F. Vincent**, J. Larroque, N. Locatelli, N. Ben Romdhane, O. Bichler, C. Gamrat, W. S. Zhao, J.-O. Klein, S. Galdin-Retailleau et D. Querlioz, « Spin-Transfer Torque Magnetic Memory as a Stochastic Memristive Synapse for Neuromorphic Systems », *IEEE Trans. Biomedical Circuits and Systems*, vol. 9, n° 2, p. 166–174, 2015. doi:10.1109/TBCAS.2015.2414423

✠ **A. F. Vincent**, N. Locatelli, J.-O. Klein, W.S. Zhao, S. Galdin-Retailleau et D. Querlioz, « Analytical Macrospin Modeling of the Stochastic Switching Time of Spin Transfer-Torque Devices », *IEEE Trans. Electron Dev.*, vol. 62, n° 1, p. 164–170, 2015. doi:10.1109/TED.2014.2372475

✠ D. Querlioz, O. Bichler, **A. F. Vincent** et C. Gamrat, « Bioinspired Programming of Memory Devices for Implementing an Inference Engine », *Proceedings of the IEEE*, vol. 103, n° 8, p. 1398, 2015. doi:10.1109/JPROC.2015.2437616

✠ S. Saighi, C. G Mayr, T. Serrano-Gotarredona, H. Schmidt, G. Lecerf, J. Tomas, J. Grollier, S. Boyn, **A. F. Vincent**, D. Querlioz, S. La Barbera, F. Alibart, D. Vuillaume, O. Bichler, C. Gamrat et B. Linares-Barranco, « Plasticity in Memristive Devices for Spiking Neural Networks », *Front. Neurosci. – Neuromorphic Engineering*, vol. 9, n° 51, 2015. doi:10.3389/fnins.2015.00051

Conférences internationales avec actes

✠ C. H. Bennett, S. La Barbera, **A. F. Vincent**, J.-O. Klein, F. Alibart et D. Querlioz, « Exploiting the Short-term to Long-term Plasticity Transition in Memristive Nanodevice Learning Architectures », *IJCNN*, 2016.

✠ N. LOCATELLI, **A. F. VINCENT**, S. GALDIN-RETAILLEAU, J.-O. KLEIN et D. QUERLIOZ, « Approximate Programming of Magnetic Memory Elements for Energy Saving », *Procs. of MEMRYSIS*, 2015. doi:10.1109/MEMRISYS.2015.7378400

✠ S. LA BARBERA, **A. F. VINCENT**, D. VUILLAUME, D. QUERLIOZ et F. ALIBART, « Short-term to Long-term Plasticity Transition in Filamentary Switching for Memory Applications », *Procs. of MEMRYSIS*, 2015. doi:10.1109/MEMRISYS.2015.7378402

✠ D. QUERLIOZ, **A. F. VINCENT**, A. MIZRAHI, N. LOCATELLI, J. S. FRIEDMAN et D. VODENICAREVIC, « Computational Techniques for the Design of Bioinspired Systems that Employ Nanodevices », *Procs. of IWCE*, 2015. https://nanohub.org/resources/22988

✠ N. LOCATELLI, **A. F. VINCENT**, A. MIZRAHI, J. S. FRIEDMAN, D. VODENICAREVIC, J.-V. KIM, J.-O. KLEIN, W. S. ZHAO, J. GROLLIER et D. QUERLIOZ, « Spintronic Devices as Key-elements for Energy-efficient Neuro-inspired Architectures », *Procs. of DATE*, p. 994, 2015.

✠ **A. F. VINCENT**, J. LARROQUE, W. S. ZHAO, N. BEN ROMDHANE, O. BICHLER, C. GAMRAT, J.-O. KLEIN, S. GALDIN-RETAILLEAU et D. QUERLIOZ, « Spin-Transfer Torque Magnetic Memory as a Stochastic Memristive Synapse », *Procs. of ISCAS*, p. 1074, 2014. doi:10.1109/ISCAS.2014.6865325

✠ **A. F. VINCENT**, W.S. ZHAO, J.-O. KLEIN, S. GALDIN-RETAILLEAU et D. QUERLIOZ, « Monte-Carlo Simulations of Magnetic Tunnel Junctions : from Physics to Application », *Procs. of IWCE*, 2014. *Intervention orale.* doi:10.1109/IWCE.2014.6865815

✠ M. DIEZ-GARCIA, **A. F. VINCENT**, N. IZARD et D. QUERLIOZ, « Monte Carlo Simulations of Carbon Nanotube Networks for Optoelectronic Applications », *Procs. of ACM/IEEE Nanoarch*, p. 135, 2014. doi:10.1145/2770287.2770319

Conférences et communications internationales sans actes

✠ S. LA BARBERA, **A. F. VINCENT**, D. VUILLAUME, D. QUERLIOZ et F. ALIBART, « Multiple Plasticity Mechanisms in Synaptic-like Memory Devices », *European Material Research Society Spring Meeting*, 2016. *Intervention orale.*

✠ **A. F. VINCENT**, N. LOCATELLI, J. LARROQUE, W. S. ZHAO, J.-O. KLEIN, S. GALDIN-RETAILLEAU et D. QUERLIOZ, « Stochastic Memristive Synapses from Spin-Transfer Torque Magnetic Tunnel Junctions », *IEEE International Magnetics Conference*, 2015. *Intervention orale.*

✠ **A. F. VINCENT**, J. LARROQUE, N. BEN ROMDHANE, O. BICHLER, C. GAMRAT, W. S. ZHAO, J.-O. KLEIN, S. GALDIN-RETAILLEAU et D. QUERLIOZ, « Spin-Transfer Torque Magnetic Memory as Stochastic Memristive Synapses », *2nd International Workshop on Neuromorphic and Brain-Based Computing Systems*, 2015. *Poster.*

Événements nationaux

✠ **A. F. VINCENT**, J. LARROQUE, N. BEN ROMDHANE, O. BICHLER, C. GAMRAT, W. S. ZHAO, J.-O. KLEIN, S. GALDIN-RETAILLEAU et D. QUERLIOZ, « Spin-Transfer Torque Magnetic Memory as Stochastic Memristive Synapses », *Colloque national du GdR SoC-SiP*, 2016. *Poster*.

✠ **A. F. VINCENT**, J. LARROQUE, N. BEN ROMDHANE, O. BICHLER, C. GAMRAT, W. S. ZHAO, J.-O. KLEIN, S. GALDIN-RETAILLEAU et D. QUERLIOZ, « Spin-Transfer Torque Magnetic Memory as Stochastic Memristive Synapses », *Journées du GdR BioComp*, 2015. *Poster*.

✠ **A. F. VINCENT**, J. LARROQUE, N. BEN ROMDHANE, O. BICHLER, C. GAMRAT, W. S. ZHAO, J.-O. KLEIN, S. GALDIN-RETAILLEAU et D. QUERLIOZ, « Spin-Transfer Torque Magnetic Memory as Stochastic Memristive Synapses », *Journées NeuroSTIC*, 2015. *Poster*.

✠ **A. F. VINCENT**, J. LARROQUE, W. S. ZHAO, N. BEN ROMDHANE, O. BICHLER, C. GAMRAT, J.-O. KLEIN, S. GALDIN-RETAILLEAU et D. QUERLIOZ, « Nanosynapses binaires magnétiques : lorsque le hasard devient un atout », *Journées NeuroSTIC*, 2014. *Poster*.

✠ **A. F. VINCENT**, J. LARROQUE, W. S. ZHAO, N. BEN ROMDHANE, O. BICHLER, C. GAMRAT, J.-O. KLEIN, S. GALDIN-RETAILLEAU et D. QUERLIOZ, « Utilisation de mémoires magnétiques à transfert de spin comme synapses memristives stochastiques », *Journées Nationales du Réseau Doctoral en Micro-nanoélectronique*, 2014. *Intervention orale*.

Chapitres d'ouvrage

✠ **A. F. VINCENT**, N. LOCATELLI et D. QUERLIOZ, « Reinterpretation of Magnetic Tunnel Junctions as Stochastic Memristive Devices », *Springer*, 2017. À paraître dans *Advances in Neuromorphic Hardware using Emerging Nanodevices*, édité par M. SURI.

✠ D. QUERLIOZ, O. BICHLER, **A. F. VINCENT** et C. GAMRAT, « Theoretical Analysis of Spike Timing Dependent Plasticity Learning with Memristive Devices », *Springer*, 2017. À paraître dans *Advances in Neuromorphic Hardware using Emerging Nanodevices*, édité par M. SURI.

université PARIS-SACLAY

ÉCOLE DOCTORALE
Physique et ingénierie : électrons, photons, sciences du vivant (EOBE)

Titre : Vers une utilisation synaptique de composants mémoires innovants pour l'électronique neuro-inspirée

Mots clefs : électronique neuro-inspirée, systèmes neuromorphiques, synapses artificielles, jonctions tunnel magnétiques, cellules électrochimiques métalliques.

Résumé : Les réseaux de neurones artificiels, dont le concept s'inspire du fonctionnement des cerveaux biologiques et de leurs capacités d'apprentissage, sont une approche prometteuse pour répondre aux nouveaux usages informatiques dits « cognitifs », tels que la reconnaissance d'images ou l'interaction en langage naturel. Néanmoins, leur mise en œuvre par des ordinateurs conventionnels est peu efficace. Une solution à ce problème est le développement de puces d'accélération matérielle spécialisées qui comportent :

– des *neurones*, unités de traitement de l'information, pour lesquelles des circuits électroniques efficaces existent ;
– des *synapses*, reliant les neurones mais aussi support matériel de l'apprentissage, par le biais de la modulation de leur conductance électrique (qualifiée de « plasticité synaptique »). Réaliser des synapses artificielles intégrables densément et capables d'apprendre *in situ* reste aujourd'hui un défi majeur.

Ces travaux de thèse portent sur l'utilisation synaptique de nanocomposants mémoires innovants, dont certains comportements plastiques riches et intrinsèques sont analogues aux fonctionnalités que nous recherchons.
Nous nous intéressons tout d'abord aux jonctions tunnel magnétiques à transfert de spin, développées dans l'industrie pour concevoir de nouvelles mémoires informatiques non volatiles. Nous montrons qu'il est aussi possible d'en faire des synapses artificielles binaires. Après la modélisation analytique de leur comportement naturellement stochastique, nous présentons comment exploiter ce dernier pour faciliter la mise en œuvre *in situ* d'une règle d'apprentissage probabiliste. À l'aide d'outils de simulation développés au laboratoire, nous étudions l'influence du régime de programmation sur la robustesse d'un système à la variabilité de telles synapses et sur leur consommation énergétique.
Nous nous tournons ensuite vers des cellules électrochimiques métalliques Ag_2S, d'autres nanocomposants mémoires innovants fabriqués et étudiés par des collaborateurs de l'Université de Lille I, qui y ont déjà observé plusieurs comportements plastiques. Nous avons découvert une plasticité supplémentaire, proche d'un comportement observé en neurosciences. Grâce à un modèle analytique simple permettant de comprendre les relations entre les différentes plasticités, nous montrons en simulation une preuve de concept d'apprentissage non supervisé qui repose sur l'interaction de ces multiples comportements.
Pour finir, nous soulevons des pistes de réflexion sur les défis posés par les circuits nécessaires au bon fonctionnement d'un système utilisant comme synapses artificielles les nanocomposants étudiés, notamment lors de la lecture ou de l'écriture de ces derniers.
Les résultats de cette thèse ouvrent la voie à la conception de systèmes neuro-inspirés capables d'apprendre en s'appuyant sur la richesse de comportements plastiques offerte par les nanocomposants mémoires innovants.

Title: Toward using innovative memory devices as artificial synapses in neuro-inspired electronics

Keywords: neuro-inspired electronics, neuromorphic systems, artificial synapses, magnetic tunnel junctions, electrochemical metalization cells.

Abstract: Artificial neural networks, which take some inspiration from the behavior of biological brains and their learning capabilities, are promising tools to address emerging computing uses known as "cognitive" tasks like classifying images or natural language interaction. However, implementing them on conventional computers is poorly efficient. A solution to this problem is to develop specialized acceleration chips which feature:

- *neurons*, the information processing units, which can be implemented efficiently with current electronic technologies;
- *synapses*, the connections between the neurons which also support the learning process by adjusting their electrical conductance ("synaptic plasticity"). Implementing artificial synapses with high integration and on-line learning capabilities is still a challenge.

This thesis explores the use of innovative memory nanodevices as artificial synapses: some of their rich plastic behaviors naturally implement features that are difficult to access with other devices. First, we investigate spin-transfer torque magnetic tunnel junctions, that are currently developed in industry as a new non volatile memory technology. We show that they can also be used as binary artificial synapses. After modeling their intrinsic stochastic behavior analytically, we describe how to harness this behavior to facilitate the implementation of an on-line probabilistic learning rule. With simulation tools developed in the laboratory, we detail the impact of the programming regime on the resilience of a system that uses such synapses, as well as on the system's power consumption
We then investigate Ag_2S electrochemical metalization cells, another type of innovative memory nanodevices fabricated and characterized by collaborators from Université de Lille I, who had already observed the existence of several plastic behaviors. We discovered an additional plasticity, close to a behavior known in neurosciences. With a simple analytical model that allows a better understanding of the relationships between theses plasticities, we show by simulations means a proof of concept of an unsupervised learning that relies on the interaction of the plastic behaviors theses nanodevices feature.
Finally, we consider the challenges arising from the circuits that are required to read and write such artificial synapses in a neuro-inspired system.
The results of this Ph.D. work pave the way for the design of neuro-inspired systems that can learn by harnessing the rich plastic behaviors that are featured by innovative memory nanodevices.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128, 91190 Saint-Aubin, France