Neurone analogique robuste et technologies émergentes pour les architectures neuromorphiques

Antoine Joubert

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Nanoélectronique et nanotechnologie**

Arrêté ministériel : 7 août 2006

Présentée par

Antoine JOUBERT

Thèse dirigée par **Dominique DAVID**

préparée au sein du **CEA-LETI, LISAN**
et de l'**École Doctorale Electronique, Electrotechnique, Automatique et Traitement du Signal (EEATS)**

Neurone Analogique Robuste et Technologies Émergentes pour les Architectures Neuromorphiques

Thèse soutenue publiquement le **26 mars 2013**,
devant le jury composé de :

Christian JUTTEN
INP Grenoble, Président
Sylvain SAIGHI
Université Bordeaux 1, Rapporteur
Ian O'CONNOR
EC-Lyon, Rapporteur
Olivier TEMAM
INRIA Saclay, Examinateur
Cyril LAHUEC
Telecom Brest, Examinateur
Dominique DAVID
CEA-LETI Grenoble, Directeur de thèse
Rodolphe HELIOT
CEA-LETI Grenoble, Co-Encadrant

Résumé

Les récentes évolutions en microélectronique nécessitent une attention particulière lors de la conception d'un circuit. Depuis les nœuds technologiques de quelques dizaines de nanomètres, les contraintes de consommation deviennent prépondérantes. Pour répondre à ce problème, les concepteurs se penchent aujourd'hui sur l'utilisation d'architectures multi-cœurs hétérogènes incluant des accélérateurs matériels dotés d'une grande efficacité énergétique. Le maintien des spécifications d'un circuit apparait également essentiel à l'heure où sa fabrication est de plus en plus sujette à la variabilité et aux défauts. Il existe donc un réel besoin pour des accélérateurs robustes.

Les architectures neuromorphiques, et notamment les réseaux de neurones à impulsions, offrent une bonne tolérance aux défauts, de part leur parallélisme massif, et une aptitude à exécuter diverses applications à faible coût énergétique.

La thèse défendue se présente sous deux aspects. Le premier consiste en la conception d'un neurone analogique robuste et à son intégration dans un accélérateur matériel neuro-inspiré à des fins calculatoires. Cet opérateur mathématique à basse consommation a été dimensionné puis dessiné en technologie 65 nm. Intégré au sein de deux circuits, il a pu être caractérisé dans l'un d'entre eux et ainsi démontrer la faisabilité d'opérations mathématiques élémentaires. Le second objectif est d'estimer, à plus long terme, l'impact des nouvelles technologies sur le développement de ce type d'architecture. Ainsi, les axes de recherches suivis ont permis d'étudier un passage vers un nœud technologique très avancé, les opportunités procurées par des Through-Silicon-Vias ou encore, l'utilisation de mémoires résistives à changement de phase ou à filament conducteur.

Abstract

Due to the latest evolutions in microelectronic field, a special care has to be given to circuit designs. In aggressive technology nodes down to dozen of nanometres, a recent need of high energy efficiency has emerged. Consequently designers are currently exploring heterogeneous multi-cores architectures based on accelerators. Besides this problem, variability has also become a major issue. It is hard to maintain a specification without using an overhead in term of surface and/or power consumption. Therefore accelerators should be robust against fabrication defects.

Neuromorphic architectures, especially spiking neural networks, address robustness and power issues by their massively parallel and hybrid computation scheme. As they are able to tackle a broad scope of applications, they are good candidates for next generation accelerators.

This PhD thesis will present two main aspects. Our first and foremost objectives were to specify and design a robust analog neuron for computational purposes. It was designed and simulated in a 65 nm process. Used as a mathematical operator, the neuron was afterwards integrated in two versatile neuromorphic architectures. The first circuit has been characterized and performed some basic computational operators. The second part explores the impact of emerging devices in future neuromorphic architectures. The starting point was a study of the scalability of the neuron in advanced technology nodes ; this approach was then extended to several technologies such as Through-Silicon-Vias or resistive memories.

Remerciements

Je souhaite tout d'abord remercier les membres de mon jury de thèse : Christian Jutten, Sylvain Saighi, Ian O'Connor, Olivier Temam et Cyril Lahuec pour l'intérêt qu'ils ont porté à mon travail.

Je remercie également mon directeur de thèse, Dominique David, qui m'a donné des conseils judicieux tout au long de ses trois ans. Je suis extrêmement reconnaissant envers Rodolphe Héliot, mon encadrant, qui m'aura procuré l'environnement de travail idéal pour mes travaux de thèse. Je tiens également à le remercier pour la confiance et la liberté qu'il m'a accordées, ainsi que pour son extrême disponibilité.

Je remercie tous ceux qui ont été impliqués dans les projets auxquels j'ai participé pendant ces trois ans, et plus particulièrement Bilel Belhadj pour ses contributions à la réalisation des circuits. Je remercie également les autres personnes du laboratoire LISAN avec qui j'ai pu échanger et passer de bons moments. Je leur souhaite une très bonne continuation tant professionnelle que personnelle.

Je remercie mes amis docteurs pour leurs conseils et je souhaite bonne chance et bonne soutenance à tous les doctorants que j'ai côtoyés pendant ces années.

Enfin, je tiens à remercier Audrey, ma famille et mes amis pour leur soutien, leur aide, et plus simplement pour le temps que l'on a passé ensemble.

Merci à tous,

Antoine

Table des matières

Table des figures **viii**

Liste des tableaux **xi**

I Introduction **1**

A. Notions de neuromorphisme 2

B. Le neuromorphique, une approche *more-than-Moore* 18

C. Objectifs de cette thèse 29

II Intégration d'un neurone robuste pour des applications computationnelles **33**

A. Quelques notions de conception en microélectronique 34

B. Conception d'un neurone LIF robuste 44

C. Circuits réalisés 68

III Études sur les technologies avancées **85**

A. Quel avenir pour un neurone analogique ? 86

B. Notions de mémoire résistive 89

C. Cœur du neurone : élément capacitif 95

D. Implémentation des connexions synaptiques 106

IV Conclusion & perspectives **119**

A. Contributions de cette thèse 120

B. Des perspectives à court terme. 121

C. Les architectures neuromorphiques, une technologie en devenir 123

Références **129**

V Annexes **137**

A. *Adress Event Representation* - AER 137

B. Carte de test 139

Table des figures

I.1 Représentation d'un neurone biologique 4
I.2 Synapse : schéma et mécanismes 5
I.3 Synapse : règle d'apprentissage 6
I.4 Cartographie des interconnexions au sein du cerveau du macaque . . . 7
I.5 Schéma électronique équivalent du modèle d'Hodgkin-Huxley 8
I.6 Schéma électronique équivalent du modèle LIF 10
I.7 Schéma électronique équivalent du modèle Fitzhugh-Nagumo 11
I.8 Architecture de Von Neumann 19
I.9 Effets de la réduction de la longueur du canal du transistor MOS . . . 20
I.10 Loi de Moore 22
I.11 Approche "more Moore" et concept "more than Moore" 24
I.12 Évolution des processeurs durant les dernières décennies 25
I.13 Concept d'accélérateur matériel et d'optimisation énergétique 26
I.14 Comparaison de systèmes bio-inspirés 28
I.15 Chaine de programmation 30

II.1 Types de capacités disponibles 35
II.2 Front-end/Back-end 37
II.3 Phénomènes de couplage 38
II.4 Impact des procédés de fabrication sur les neurones 40
II.5 Implémentation de neurone d'après *Van Schaik* (85) 41
II.6 Implémentation de neurone d'après *Vogelstein* (86) 42
II.7 Implémentation de neurone d'après *Wijekoon* (87) 42
II.8 Implémentation de neurone d'après *Livi* (49) 43
II.9 Fonctionnement mixte des neurones 46
II.10 Schéma bloc du neurone LIF 46

II.11 Stratégie de conception . . . 47
II.12 Effets de la variabilité sur un MUX 2 :1 . . . 50
II.13 Schéma bloc modifié du neurone LIF . . . 50
II.14 Caractérisation des courants de fuite des transistors . . . 51
II.15 Interfaçace analogique/numérique . . . 52
II.16 Comparateur à bascule . . . 53
II.17 Identification des courants de fuite . . . 54
II.18 Schéma du neurone analogique réalisé. . . . 56
II.19 Structure du DAC . . . 57
II.20 Chronogramme des signaux de contrôle du DAC . . . 59
II.21 Banc de test mixte pour neurone analogique . . . 60
II.22 Simulation comportementale du neurone analogique réalisé . . . 61
II.23 Impact des procédés de fabrication sur les neurones . . . 63
II.24 Consommation énergétique du neurone . . . 65
II.25 Layout du neurone . . . 67
II.26 Tuile de 12 neurones . . . 68
II.27 Architecture du circuit *Reptile* . . . 69
II.28 Implémentation de la partie analogique du circuit *Reptile* . . . 70
II.29 Schéma de l'amplificateur Miller . . . 71
II.30 Photographies du circuit . . . 72
II.31 Carte de caractérisation . . . 73
II.32 Environnement de caractérisation . . . 74
II.33 Comportement d'un neurone . . . 75
II.34 Caractérisation énergétique . . . 76
II.35 Comportement de fuite d'un neurone . . . 77
II.36 Caractérisation de la variabilité des neurones . . . 79
II.37 Layout de *Spider* . . . 81
II.38 Alimentation des tuiles . . . 83

III.1 Implémentation numérique du neurone LIF. . . . 87
III.2 Passage à l'échelle de neurones analogiques et numériques . . . 88
III.3 RLC et M . . . 90
III.4 Classification des phénomènes physiques des mémoires résistives . . . 93

TABLE DES FIGURES

III.5 Structure d'une mémoire à changement de phase 94
III.6 Fonctionnement thermique d'une PCM 95
III.7 Séquences de fonctionnement d'une CBRAM 95
III.8 Concept d'implémentation d'un futur neurone analogique 98
III.9 Structure du TSV et modèle équivalent employé. 99
III.10 Intégration du TSV dans l'environnement de simulation Cadence. . . . 100
III.11 Potentiel de membrane - neurone classique vs TSV 101
III.12 Illustrations de neurones 2D et 3D avec ou sans TSV 102
III.13 Schéma du neurone IF à base de PCM 103
III.14 Comportement du neurone IF basé sur PCM 104
III.15 Schéma du neurone pour LTP. 106
III.16 Exemple de LTP 108
III.17 Architecture implémentée avec CBRAM 110
III.18 Détails de l'architecture implémentée 112
III.19 Chronogramme de la topologie basée sur des CBRAM 114
III.20 Layout de la structure de test des CBRAMS 115
III.21 Impacts des technologies émergentes sur des neurones 117

IV.1 Réalisation d'opérations 122
IV.2 Temps de conception et développement 124
IV.3 Comparaison énergétique processeur/architecture neuromorphique . . 125

V.1 Principe d'encodage de l'AER 137
V.2 Architecture AER décrit dans (86) 138

Liste des tableaux

I.1 Notations utilisées pour modéliser un neurone 8
I.2 Quelques exemples d'implémentations silicium de l'état de l'art. 17

II.1 Flot de conception utilisé. 34
II.2 Analogique vs numérique 39
II.3 Capacités disponibles 48
II.4 Constantes de fuite implémentées pour le neurone 54
II.5 Signaux numériques du neurone : 9 entrées, 1 sortie 55
II.6 Valeurs analogiques nominales de polarisation 55
II.7 Tableau des principaux temps nécessaires à la réalisation du DAC . . . 58
II.8 Caractéristiques énergétiques de neurones 76
II.9 Caractéristiques de la fuite du neurone 78

III.1 Implémentations d'un neurone LIF : analogique VS numérique 87
III.2 Capacités innovantes 97
III.3 Caractéristiques mesurées d'un TSV. 99
III.4 Gains potentiels apportés par les TSVs 102
III.5 Paramètres de simulation 105
III.6 Cœur de stockage d'un neurone : capacité vs PCM 105
III.7 Caractéristiques des CBRAMS prises en compte pour la conception . . . 109
III.8 Caractéristiques du générateur programmable d'impulsions 111
III.9 Récapitulatif de l'impact des technologies émergentes 116

IV.1 Différentes caractéristiques de neurones 120

Glossaire

ADE : Analog Design Environment
ADN : Acide DésoxyriboNucléique
AER : Adress Event Representation
CAN : Convertisseur Analogique/Numérique
CBRAM : Conducting Bridge RAM
CMOS : Complementary Metal Oxide Semiconductor
CNA : Convertisseur Numérique/Analogique
CPU : Central Processing Unit
DAC : Digital to Analog Conversion
DK : Design Kit
DRAM : Dynamic Random Access Memory
DRC : Design Rule Check
ECM : Electro Chemical Metallization
FFT : Fast Fourier Transform
FLL : Frequency Locked Loop
FPGA : Field-Programmable Gate Array
GPU : Graphics Processing Unit
GST : Germanium-Antimony-Tellurium
IF : Integrate-and-Fire
LIF : Leaky Integrate-and-Fire
LTD : Long Terme Depression
LTP : Long Terme Potentiation
LVS : Layout Versus Schematic
MEMS : Microelectromechanical Systems
MIM : Metal Insulator Metal

MOS : Metal Oxide Semiconductor
NEMS : Nano Electro-Mechanical Systems
NOC : Network-On-Chip
OTA : Operational Trans Amplifier
PA : Potentiel d'Action
PCM : Phase Change Memory
PI : Pulsed Injection
PVT : Process-Voltage-Temperature
RAM : Random Access Memory
SNR : Signal to Noise Ratio
SOC : System On Chip
SOI : Silicon On Insulator
SPEC : Standard Performance Evaluation Corporation
SRAM : Static Random Access Memory
STDP : Spike-Time-Dependent Plasticity
TSV : Through-Silicon-Vias
VLSI : Very-Large-Scale Integration

I

Introduction

Celui qui copie la nature est impuissant, celui qui l'interprète est ridicule, celui qui l'ignore n'est rien du tout.

René Barjavel, *Colomb de la lune*, 1959.

Sommaire

A. Notions de neuromorphisme 2
1. Quelques rappels de biologie 2
2. Modèles mathématiques et électroniques de neurone 6
3. Exemples de réalisations de neurones 12
4. Comparatifs des implémentations de réseaux de neurones sur silicium 16
5. Quels sont les objectifs de ces implémentations ? 16
B. Le neuromorphique, une approche *more-than-Moore* 18
1. Historique du transistor : miniaturisation et hausse de la fréquence 19
2. Attentes et évolutions de l'industrie microélectronique . . . 21
3. Les limites physiques du transistor MOS en passe d'être atteintes 23
4. Un environnement propice au développement du neuromorphique 23
C. Objectifs de cette thèse 29
1. Le projet Arch^{2}Neu : 29
2. Contributions au projet Arch^{2}Neu. 30
3. Etudes des technologies avancées. 30

I. INTRODUCTION

- Résumé - Ce chapitre est une introduction aux systèmes neuro-inspirés autrement appelés systèmes neuromorphiques. On y présentera succinctement le neurone biologique et son environnement pour comprendre les diverses démarches qui ont historiquement permis l'élaboration des différents modèles de neurones. La miniaturisation des composants intégrés en micro-électronique a rendu possible l'intégration silicium de ces différents modèles de neurones. Ces neurones électroniques ont été conçus et réalisés dans l'optique de fournir un outil pour les biologistes. Cependant, il est fort probable que les systèmes neuromorphiques auront un rôle à jouer dans les architectures de calcul de demain.

A. Notions de neuromorphisme

Dès les années 1950, différents travaux de recherche (29, 68) ont initié le développement du concept de neuromorphisme. Il a ensuite été approfondi par Carver Mead à la fin des années 1980. En s'inspirant du fonctionnement du cerveau, il a émis l'idée de le reproduire sur silicium à l'aide d'une intégration à très grande échelle (VLSI) (53).

Dans cette optique, une approche bottom-up du cerveau est logiquement privilégiée. Ainsi, on étudie le fonctionnement de la cellule de base, c'est à dire un neurone qui sera caractérisé, modélisé puis implémenté, à grande échelle, pour former un réseau.

Le prochain paragraphe a pour objectif de présenter les notions de biologie nécessaires à la suite du manuscrit.

1. Quelques rappels de biologie

Une cellule est un organisme de structuration et de régulation pour tout être vivant. Chaque cellule est spécifique à son environnement mais partage cependant quelques éléments en commun avec l'ensemble des cellules, comme par exemple, une membrane. Contrairement aux cellules procaryotes, les cellules eucaryotes possèdent toutes un noyau. La membrane sert à isoler les milieux intra et extra cellulaire. Le noyau contient des informations nécessaires au bon fonctionnement de la cellule ainsi qu'à sa reproduction. En constant échange avec l'extérieur et entre elles, les cellules interagissent à l'aide de messagers chimiques, de contraintes mécaniques ou encore de signaux électriques. Nous nous intéresserons maintenant plus particulièrement aux cellules nerveuses.

a) Neurone biologique

Le neurone est une des cellules de base du système nerveux. Il transmet une activité électrique appelée influx nerveux sous forme de séquences de potentiels d'action. On peut voir sur la figure I.1 le schéma d'un neurone biologique et identifier 3 grandes parties :

- l'arbre dendritique : il est constitué des dendrites du neurone. Les signaux en provenance des différentes terminaisons axonales des précédents neurones y sont regroupés. Ces signaux sont appelés potentiels d'action ou plus simplement impulsions. Ils traversent une synapse que l'on détaillera dans la partie I-A.1.b). Le transfert d'un potentiel d'action se fera donc toujours d'un neurone présynaptique à un neurone postsynaptique.
- le soma : la membrane du neurone est constituée d'une bicouche lipidique. Elle sert à isoler mécaniquement et électriquement l'intérieur du neurone (noyau et cytoplasme). Les espèces d'ions nécessaires à la propagation des potentiels d'action sont majoritairement potassiques (K^+) à l'intérieur du neurone et sodiques (Na^+) à l'extérieur. On peut noter la présence d'un courant de fuite responsable du retour au potentiel de repos du soma. En effet la membrane n'est pas un diélectrique parfait et contient des sites d'échanges passifs spécifiques à un type d'ion que l'on appelle canaux ioniques.
- l'axone : il permet la propagation du signal vers les différents neurones destinataires par ses terminaisons axonales. La propagation du potentiel d'action peut être accélérée par l'intermédiaire des nœuds de Ranvier, endroits situés le long de l'axone où s'amincit la paroi isolante appelée gaine de myéline. Les échanges d'ions y sont localement favorisés et permettent la régénération du signal impulsionnel. En sautant de nœud en nœud, la propagation du potentiel d'action est accélérée.

D'un point de vue fonctionnel, le neurone combine plusieurs entrées au niveau du soma. Le résultat, la génération d'un potentiel d'action, se propage via l'axone, vers d'autres neurones par l'intermédiaire de synapses.

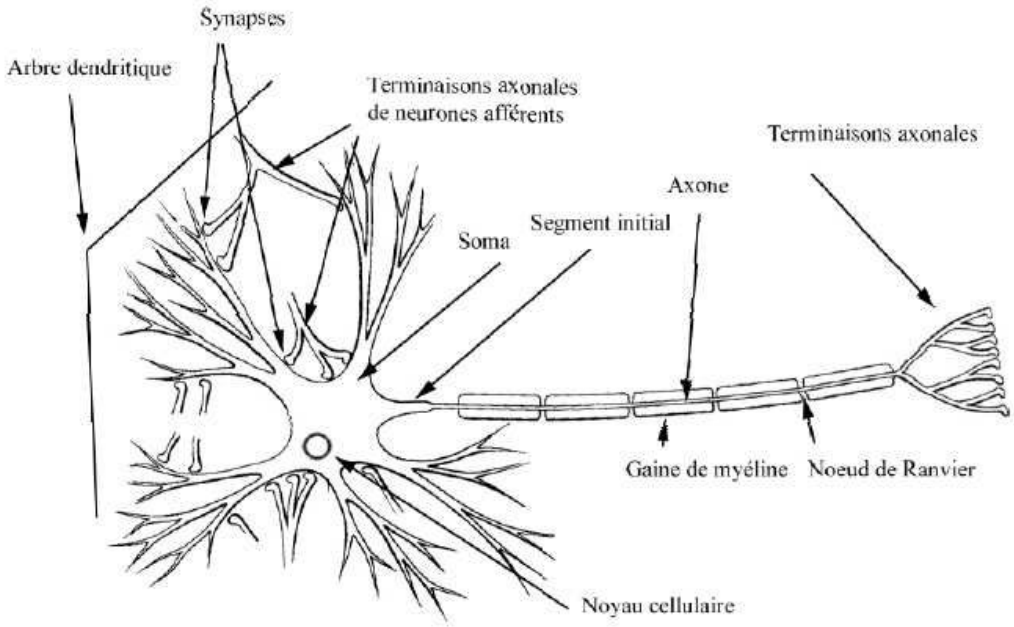


Figure I.1: Représentation d'un neurone biologique d'après (21).

b) La synapse

La synapse a pour rôle la transmission de l'information du neurone présynaptique au postsynaptique et nécessite par conséquent au moins deux neurones. En outre, il a été montré par (19) que l'environnement extra-cellulaire joue un rôle sur les capacités d'une synapse.

La synapse présente une forme et un fonctionnement électrico-chimique particulier visible sur la figure I.2. Un potentiel d'action électrique stimule le bouton présynaptique qui relâche des messagers chimiques. Ces derniers vont venir modifier l'ouverture des canaux ioniques au niveau de l'arbre dendritique du neurone postsynaptique. En laissant passer des ions calciques (Ca^{2+}), sodiques (Na^{+}) ou potassiques (K^{+}), le potentiel de membrane du neurone postsynaptique est alors modifié. En parallèle, les stocks de messagers chimiques consommés se reconstruisent progressivement.

La synapse peut ainsi être plus ou moins inhibitrice ou excitatrice, ce qui définit son poids synaptique. Dans le premier cas, elle va diminuer le potentiel de membrane et donc retarder la génération du prochain potentiel d'action. A l'inverse, une synapse excitatrice provoque l'augmentation du potentiel de membrane et permet éventuellement

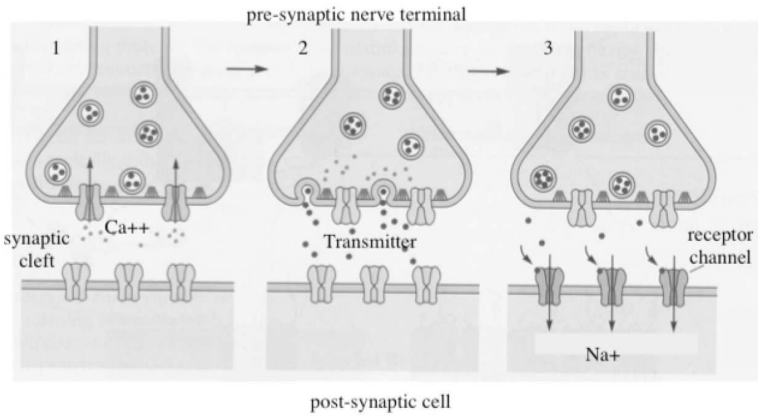


Figure I.2: Synapse chimique d'après (75) : (1) l'arrivée d'un potentiel d'action modifie l'équilibre électrique dans le neurone présynaptique. (2) Celui-ci relâche des vésicules de neurotransmetteur. (3) Ce neurotransmetteur va activer les canaux sodiques du neurone postsynaptique. Cette activation peut se faire par le biais d'une autre molécule (synapse à action directe ou indirecte).

la génération immédiate d'un potentiel d'action.

Les synapses sont de véritables mémoires évolutives. Au cours du temps, elles s'affirment dans un caractère inhibiteur ou excitateur à long terme : LTD pour *Long Term Depression* ou LTP pour *Long Term Potentiation*. Ceci leur permet de renforcer ou d'affaiblir la transmission d'informations au sein d'un réseau de neurones. La STDP (*Spike-Time-Dependent Plasticity*) (77) est une théorie expliquant l'évolution sur le long terme du poids synaptique. Elle repose sur un ajustement du poids synaptique en fonction de la réponse du neurone, le poids évoluant selon la fonction tracée sur le figure I.3. Lorsqu'une synapse excitée contribue à la génération d'un potentiel d'action dans une fenêtre de temps, $\delta t > 0$, son poids est incrémenté de δw. A contrario la stimulation d'une synapse excitatrice après la génération d'un potentiel d'action du neurone postsynaptique diminue son poids ($\delta t < 0$ et décrémentation de δw). Ces évolutions semblent être à la base du développement cérébral et des phénomènes d'apprentissage.

c) Le cerveau, réseau de neurones

Grâce aux synapses, les neurones forment un réseau fortement connecté au sein du cerveau. En effet, on estime qu'un cerveau humain est constitué de 10^{10} neurones et 10^{14} synapses (53). D'une part, ceci induit le traitement d'un grand nombre d'informations.

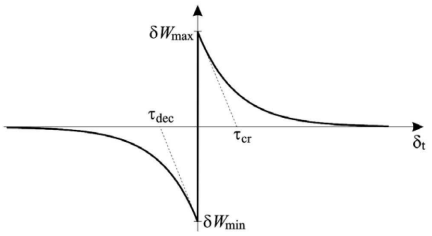


Figure I.3: Évolution du poids synaptique selon l'écart entre le temps d'arrivée du potentiel d'action présynaptique et la génération de celui postsynaptique. D'après (66).

Ceci permet d'autre part une certaine redondance du traitement de données dans les différents cortex, et par conséquent induit une robustesse lorsqu'une cellule s'altère ou meurt.

On peut également constater l'efficacité énergétique du cerveau. Celui-ci consomme seulement une vingtaine de watts (53) alors qu'il traite parallèlement des informations en provenance de nombreux capteurs (ouïe, vue, ...). A celles-ci viennent s'ajouter la pensée ou les réflexes qui ajoutent à la complexité du système. Une représentation de la connectivité du système nerveux d'un singe est présentée sur la figure I.4. On y retrouve les différentes régions spécialisées du cerveau et leurs interconnexions. En contenant plusieurs accélérateurs spécifiques, le fonctionnement de certains circuits intégrés est finalement assez similaire. L'implémentation d'accélérateurs optimisés pour une tâche puis interconnectés permet la réalisation de fonctions plus complexes.

Les capacités hors-normes du cerveau ont suscité l'intérêt de nombreux chercheurs qui ont tenté de comprendre son fonctionnement en modélisant son unité de base : le neurone.

2. Modèles mathématiques et électroniques de neurone

La première modélisation d'un neurone biologique fut développée par Hodgkin et Huxley en 1952 (29). En vue de son usage à des fins de simulations pour comprendre le système nerveux ou pour réaliser du traitement de l'information, et face à la complexité du neurone biologique, il a rapidement fallu développer d'autres modèles. Ils ont permis d'améliorer l'efficacité en termes de temps de simulation ou en termes de facilité d'intégration. Avant de présenter par ordre chronologique les principaux modèles de

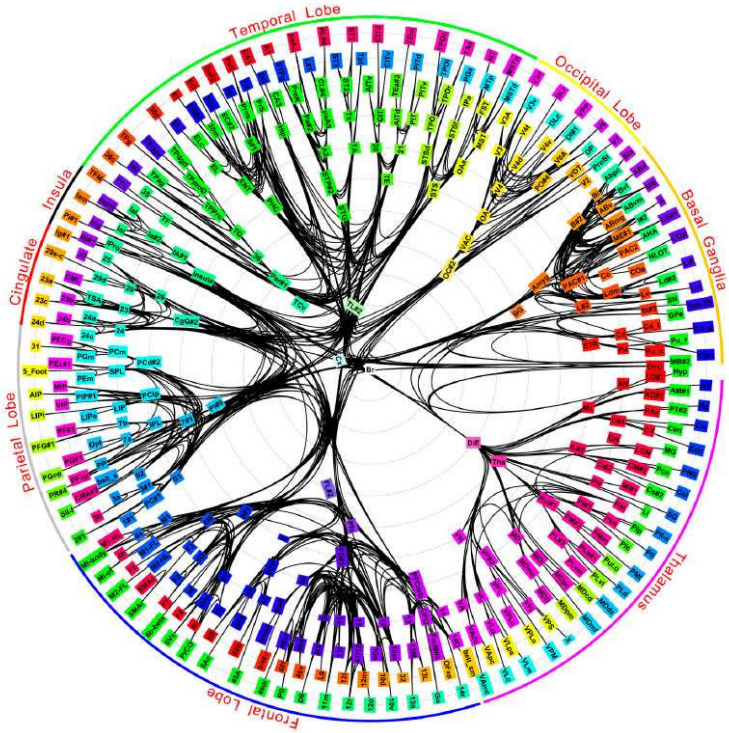


Figure I.4: Modélisation des interconnexions entre les régions du cerveau du macaque. D'après (55).

neurones à impulsions, on précise les notations utilisées dans le tableau I.1. Les unités renseignées sont celles généralement utilisées lors d'implémentations électroniques.

a) Hodgkin-Huxley

Ce modèle de référence, qui se veut proche de la biologie, est basé sur la conductance des canaux ioniques, noté g_i, qui relient les milieux intra et extra-cellulaire. A chaque canal est associé un potentiel d'équilibre noté V_i. L'équation du modèle se résume alors à :

$$\Sigma(g_i(t) * (V_m - V_i)) + I_{leak} = C_m * \frac{dV_m}{dt} \tag{I.1}$$

Le schéma électrique équivalent est montré sur le figure I.5 sur lequel on peut parfaitement identifier les différentes caractéristiques du neurone biologique telles que son potentiel de repos, sa capacité et les courants ioniques formés par un générateur de tension et une conductance variable.

Nom	Description	Unité
X ou I	entrée	(V)
Y	sortie	(V)
A ou W	poids des entrées	S.U. ou (S)
V_{th}	tension seuil	(V)
V_m	tension membranaire	(V)
C_m	capacité membranaire	(V)
S	fonction sigmoïde	S.U.
δ	impulsion de dirac	S.U.
V_{rest} ou V_{leak}	tension de repos	(V)

Table I.1: Notations utilisées pour modéliser un neurone

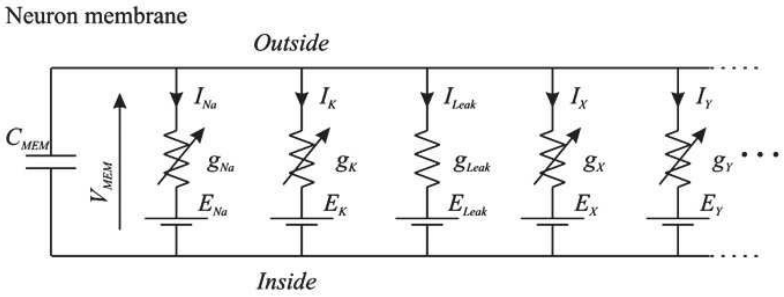


Figure I.5: Schéma électronique équivalent du modèle d'Hodgkin-Huxley d'après (66). Les milieux intra et extracellulaires sont séparés par la capacité et les différents canaux ioniques sont modélisés par une source de tension et une conductance variable. L'ajout de courants supplémentaires, I_x ou I_y sur le schéma, peut être utilisé pour décrire plus précisément le fonctionnement du neurone.

Pour des fins computationnelles, ce modèle n'est pas optimal puisqu'il fait appel à une dérivée temporelle ainsi qu'à des conductances fonctions du temps. Par conséquent, il a donné naissance à de nombreux autres modèles plus ou moins proches de la biologie ou au contraire plus ou moins formels.

b) Perceptron

Le modèle du perceptron a été décrit par Rosenblatt en 1958 (68), quelques années après le modèle d'Hodgkin-Huxley. Il s'agit du modèle le plus simple dont le compor-

tement est décrit par l'équation suivante :

$$Y = S(\Sigma(A_i * X_i) - V_{th}) \tag{I.2}$$

Les entrées sont combinées et la sortie passe à 1 lorsque le seuil est dépassé. Un perceptron peut être vu comme un réseau mono-couche et marque, par conséquent, le début d'assemblage de neurones en réseaux.

c) Integrate and fire (IF)

C'est une version simplifié du model d'Hodgkin & Huxley présenté précédemment. Les conductances des différents canaux ioniques sont ici supposées constantes au cours du temps. Les entrées sont intégrées sur la capacité membranaire :

$$\Sigma(W_i * X_i) = C_m * \frac{dV_m}{dt} \tag{I.3}$$

Lorsque le potentiel membranaire V_m atteint la tension de seuil V_{th}, un potentiel d'action est émis, accompagné d'une retour de V_m à une valeur de repos V_{rest} :

$$Y = \delta(V_m - V_{th}) \tag{I.4}$$

$$V_m = V_{rest} \tag{I.5}$$

d) Leaky integrate and fire (LIF)

Pour ce modèle, présenté sur la figure I.6, on ajoute une fuite à l'équation I.3. Elle est notée g_{leak}, conductance de fuite constante, dans l'équation suivante :

$$\Sigma(W_i * X_i) - I_{leak} = C_m * \frac{dV_m}{dt} \tag{I.6}$$

$$avec\ I_{leak} = g_{leak}(V_m - V_{rest}) \tag{I.7}$$

Ces deux derniers modèles ont été largement utilisés pour l'émulation de réseaux d'un grand nombre de neurones. En simplifiant leur fonctionnement, la surface de silicium nécessaire pour implémenter un neurone diminue de façon significative.

On verra par la suite, que la contrainte de surface se trouvera alors sur l'implémentation de l'ensemble des synapses. En effet, si le nombre de neurones augmentent d'un ordre n, les synapses devront augmenter en n^2 pour conserver le même degré de connectivité inter-neurones.

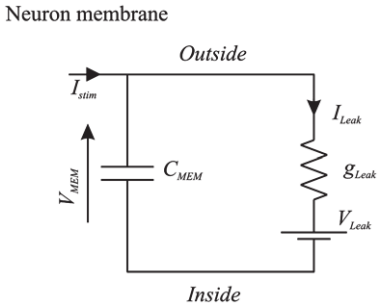


Figure I.6: Schéma électronique équivalent du modèle LIF d'après (66) : les milieux intra et extracellulaires sont séparés par la capacité et reliés par un courant de fuite. On notera cependant l'absence de remise au potentiel de repos V_{rest} sur ce schéma.

e) FitzHugh-Nagumo

Ce modèle a été décrit par Richard FitzHugh en 1961 (26) dans un souci de simplification du modèle d'Hodgkin Huxley. Il est question d'obtenir un comportement semblable au neurone biologique sans faire une réplique de ses constituants. Le fonctionnement du neurone est alors assimilé à un système oscillatoire non-linéaire général. Il est décrit à l'aide d'un système de deux équations différentielles couplées :

$$x' = c(y + x - x^3/3 + z) \tag{I.8}$$

$$y' = -(x - a + by)/c \tag{I.9}$$

J. Nagumo (59) propose rapidement une implémentation électronique qui, en adaptant les équations, se réécrivent alors :

$$j = C\frac{dv}{d\tau} - i - f(e) \tag{I.10}$$

$$L\frac{di}{d\tau} + Ri = -v = e - E_0 \tag{I.11}$$

ou f(e) est l'équation de courant d'une diode à effet tunnel. Le schéma montré sur le figure I.7, précise les autres variables des équations précédentes.

f) Izhikevich

Dans la même optique, Izhikevich propose dans (37) un système de deux équations différentielles et d'une condition de remise à zéro. Le modèle est également simple au

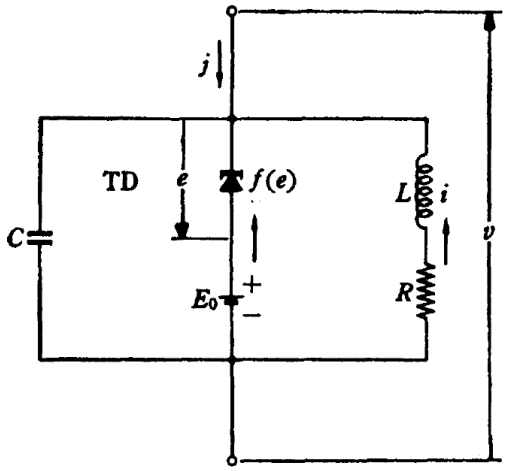


Figure I.7: Schéma électronique équivalent du modèle de Fitzhugh et implémenté par Nagumo dans (59).

niveau computationnel et décrit jusqu'à vingt modes d'émission du potentiel d'action. Il n'est pas bio-représentatif mais purement comportemental puisqu'aucun des canaux ioniques n'est explicitement modélisé. Dans les équations I.12 et I.13, v et u représentent le potentiel de membrane et son évolution (modélisation du comportement des canaux ioniques). Les quatre paramètres a, b, c et d permettent de décrire l'ensemble des régimes d'émissions et représentent respectivement : une échelle de temps pour la variable u, l'interaction entre u et v, le potentiel initial de la membrane équivalent à V_{rest} et la valeur initiale de u.

$$v' = 0.04 * v^2 + 5 * v + 140 - u + I \tag{I.12}$$

$$u' = a * (b * v - u) \tag{I.13}$$

$$et\ si\ v > 30\ mV, \left\{ \begin{array}{rcl} v & = & c \\ u & = & u + d \end{array} \right\} \tag{I.14}$$

g) Adaptative exponential - Adex

Toujours dans la même veine, le modèle développé par Wulfram Gurstner et Romain Brette dans (9) s'écrit également à l'aide d'un système de deux équations. La dynamique de potentiel de membrane est cette fois-ci décrit à l'aide d'une évolution exponentielle

et s'écrit :

$$C_m \frac{dV_m}{dt} = -g_{leak}(V_m - V_{rest}) + g_{leak}\Delta_T exp(\frac{V_m - V_{th}}{\Delta_T}) - w + I \quad \text{(I.15)}$$

$$\tau_w \frac{dw}{dt} = a(V_m - V_{rest}) - w \quad \text{(I.16)}$$

$$et\ si\ V_m > 30\ mV, \left\{ \begin{array}{ccc} v & = & V_{rest} \\ w & = & w + b \end{array} \right\} \quad \text{(I.17)}$$

Où Δ_T, w, τ_w et a sont respectivement un facteur de pente, la variable d'adaptation, sa constante de temps, et son couplage avec les tensions du neurone.

Dans ces trois derniers modèles, on notera que le point de départ est un oscillateur décrit à l'aide d'un système différentiel de deux équations. Ils se différencient par l'évolution du potentiel de membrane décrite successivement par un terme en cube, en carré et enfin en exponentiel.

3. Exemples de réalisations de neurones

Il existe différentes approches afin de simuler un réseau de neurones biologiques à impulsion. La première est la simulation logicielle de l'ensemble des neurones sur un réseau de microprocesseurs (CPU ou GPU). La seconde est l'implémentation matérielle, numérique ou analogique, du neurone. Ces différentes réalisations poursuivent des objectifs précis et parfois distincts. Cette spécialisation leur confère des atouts mais également des inconvénients propres à chacun.

a) Simulations logicielles

Il existe actuellement trois projets majeurs de simulation logicielle pour les neurosciences :

- *Manchester University, Angleterre* - Steve Furber et al. - Projet SpiNNaker :

Ce premier projet a pour but de simuler une architecture neuromorphique à l'aide d'une architecture spécifique massivement parallèle (41). A terme, elle devrait se composer de 65536 puces comprenant chacunes 18 cœurs ARM968 pour simuler un réseau représentant tout juste 1% de notre cerveau. L'objectif sous-jacent est le développement d'une plateforme de simulation de réseaux de neurones facilement accessible. Des algorithmes de diverses équipes de chercheurs pourront alors être testés et ceci constituera

le principal atout de ce projet. Les différents types de neurones ainsi que les jeux de connectivité pourront être facilement paramétrés à l'aide de la l'interface de programmation PyNN, extension du langage de programmation Python pour les réseaux de neurones.

- *EPFL, Lausanne, Suisse* - Henry Markram - Projet Blue Brain :

Ce projet basé à l'EPFL a pour objectif de simuler, à l'horizon 2023, le fonctionnement du cerveau (51). A l'heure actuelle, il a déjà pour vocation l'étude de certaines maladies cérébrales en simulant le fonctionnement des neurones à une échelle moléculaire. Pour descendre à un tel niveau de modélisation, les simulations sont effectuées sur un super-calculateur Blue Gene. Elles représentent, aujourd'hui, la propagation de signaux dans plusieurs colonnes corticales, macrostructures transverses aux couches du cerveau. Les motivations sont de trouver des soins aux maladies cérébrales, d'essayer de comprendre la formation et le fonctionnement de l'esprit humain et pourquoi pas de créer des machines dotées d'une certaine intelligence.

- *San Diego USA* - Eugene Izhikevich et Allen Gruber - Brain Corporation :

Cette start-up américaine fondée en 2009 a deux objectifs majeurs. La premier est de créer des applications pour lesquelles l'exécution sur une architecture neuromorphique semble être plus appropriée. Les derniers résultats communiqués, courant été 2012, montrent l'utilisation d'une application de détection visuelle de forme incluant une part d'apprentissage dans un environnement réel. Le deuxième est la simulation du cerveau entier avec un niveau d'abstraction plus élevé que le projet précédent. Les neurones sont en effet décrits simplement à l'aide du modèle comportemental d'Izhikevich.

b) Implémentations matérielles numériques

Cette approche se révèle intéressante pour une étude du comportement globale d'un réseau puisque le temps de développement d'un neurone matériel est relativement court. Il est en effet utilisable de nouveau lors d'un passage à l'échelle à chaque nœud technologique. On pourra ainsi réfléchir à la gestion du trafic des potentiels d'action et à son optimisation au sein du réseau.

Cette approche a, par exemple, été privilégiée dans (22) pour l'étude d'une structure de type Network-On-Chip (NOC) dans laquelle un modèle LIF a été implémenté en

technologie 130 nm. Des travaux plus récents ont été réalisés par IBM et HRL, au sein du projet SyNAPSE, dont les objectifs seraient la création d'un nouveau type de processeur neuro-inspiré. En utilisant les cellules numériques standards fournies par un fondeur, l'implémentation d'un neurone ne sera toutefois pas optimisée en termes de surface et de consommation.

c) Implémentations matérielles analogiques

De part leur nature analogique, les neurones se modélisent simplement grâce aux caractéristiques intrinsèques des composants (24). Nous présentons ici, de manière détaillée, différents exemples d'implémentations analogiques de neurones à impulsions. La théorie de leur intégration à grande échelle a été décrite par Carver Mead en 1989 (53) qui introduit le concept de *neuromorphic engineering*. Deux ans plus tard, la première implémentation sur silicium a été décrite par Misha Mahowald et Rodney Douglas (50). Il s'agit d'un neurone basé sur un modèle simplifié d'Hodgkin-Huxley en vue d'une intégration à grande échelle puisque la taille du neurone est inférieure à 0,1 mm^2 sur une technologie de l'ordre du μm. Ce choix de modèle se justifiait par le fait qu'il leur semblait judicieux de conserver de nombreuses caractéristiques du neurone biologique pour que le réseau puisse interagir avec l'environnement réel. Un point intéressant est l'idée, déjà présente, de construire une architecture neuromorphique pour réaliser des opérations à une vitesse très supérieure à celle des neurones biologiques. Dans la majorité des cas, leurs interactions se feront alors de manière numérique. Après cette première réalisation des groupes de recherches sur la thématique du neuromorphique ont essaimé sur tout les continents.

- *IMS Bordeaux, France* - S. Renaud, S. Saighi, G. Le Masson et al. :

L'objectif de ce groupe est de concevoir des réseaux de neurones électroniques ayant un comportement semblable aux neurones biologiques. Dans cette optique de bio-mimétisme, le modèle d'Hodgkin-Huxley est utilisé car il possède de nombreux paramètres réalistes et permet de simuler différents types et comportements de neurones. Les circuits réalisés peuvent atteindre plusieurs mm^2 en fonction de la technologie utilisée ainsi que de l'intégration, ou non, de certains composants dans la puce comme la capacité C_m (à l'extérieur dans (46)). Un atout majeur de leur approche est de pouvoir

interagir avec des neurones biologiques et ainsi observer ses réponses en fonction de stimuli variés.

- *UC San Diego, John Hopkins University Baltimore & Stanford University, USA* - G. Cauwenberghs et al. & Boahen et al. :

Ces deux approches sont ici radicalement différentes : l'émulation d'un unique neurone proche de la réalité n'est pas leurs objectifs respectifs. Les premières réalisations (12) mettent en évidence l'attrait de ces groupes pour un réseau de neurones. Le modèle IF est choisi car plus compact et il est alors possible d'avoir un nombre de neurones plus important par circuit intégré. Le faible nombre de neurones (seulement 6) provient du nombre limité de synapses implémentables. Il faut déjà 36 synapses pour obtenir une connectivité totale entre les 6 neurones. L'arrivée de l'*Adress Event Representation* (cf. Annexe V-A.) en 2000 (8), résoudra ce problème : on peut désormais implémenter une synapse au neurone de destination. Elle sera paramétrable suivant le neurone source (27). Ceci permet une souplesse au niveau programmation des liaisons interneurones et des poids synaptiques. Récemment et grâce à l'évolution de la technologie, l'équipe de Cauwenberghs s'est tournée vers un modèle plus poussé du neurone, proche de Hodgkin-Huxley (86). Ce modèle permet de conserver un rapport avec la biologie au niveau du neurone et des applications de type traitement d'image au niveau réseau. Dans l'optique d'une simulation proche de la biologie, Kwabena Boahen s'est tournée vers une implémentation matérielle afin de simuler le fonctionnement d'un million de neurones (*Projet Neurogrid*).

- *ETH et INI Zurich, Suisse* R. Douglas, G. Indiveri, S-C. Liu, T. Delbruck et al. :

Ce groupe de chercheurs a publié de nombreux articles dans le domaine et s'intéresse particulièrement à l'interaction d'une puce avec le monde ambiant. Dans cet optique, ils ont développé une rétine et une cochlée artificielles (13, 47). Les prototypes de circuits neuromorphiques ont, quant à eux, également profité de l'*AER* pour augmenter considérablement le nombre de neurones par puce (11, 14, 34, 35). Ils ont cependant gardé l'implémentation de chacune des synapses (contrairement à (86)) et c'est seulement le transfert des potentiels d'actions qui est géré par le biais du protocole *AER*. De cette façon les paramètres des différentes synapses sont conservés, là où ils étaient

continuellement changés en fonction du neurone source dans (86), et peuvent bénéficier de tout les mécanismes d'apprentissage et d'adaptation.

- *Heidelberg, Allemagne* - K. Meier et al. :

Le projet européen FACETS, aujourd'hui terminé, aura permis de nombreuses avancées dans le domaine neuromorphique. Son but aura été de créer une architecture pour la simulation par le biais de neurones analogiques programmables. Il a également permis le développement de l'interface de programmation PyNN et celui du modèle AdEx grâce à des mesures in-vivo. Il est maintenant repris par le projet BrainScaleS dont l'objectif est de réaliser un réseau de neurones sur plusieurs wafers de silicium. Les constantes de temps mises en jeu dans ces réseaux intégrés sont plus courtes qu'en biologie et permettent des simulations plus rapide.

- *Manchester, Angleterre* - Wijekoon :

C'est la première implémentation silicium d'un neurone basé sur le modèle analytique d'Izhikevich (87). Il utilise un faible nombre de transistors et permet la reproduction des différents régimes d'émission des potentiels d'action. Cependant l'utilisation d'un réseau basé sur ce type de neurone n'a pas encore été montrée.

4. Comparatifs des implémentations de réseaux de neurones sur silicium

Les différentes implémentations silicium de réseau de neurones sont résumées dans le tableau I.2. Pour chaque publication, il est présenté le modèle implémenté, la surface du circuit, la consommation de celui-ci et la technologie utilisée. Dans certains cas, lorsque le nombre de neurones est indiqué, une estimation de la taille de ceux-ci ainsi que celle des synapses ont été calculées.

On observe de grandes disparités en termes de modèle utilisé, de nombre de neurones implémentés ou encore de technologie employée. Celles-ci s'expliquent en partie par la poursuite d'objectifs différents que l'on peut regrouper en plusieurs approches.

5. Quels sont les objectifs de ces implémentations ?

Le premier objectif, proche de la biologie, vise soit à s'interfacer avec un neurone biologique soit à le simuler le plus fidèlement possible. Dans le cas d'une stimulation

Modèle	Référence	Nb de neurones par puce	Surface circuit, mm^2 (neurone, synapse(s) μm^2)	Remarques	Consommation (μW) du circuit (neurone)	Techno. en μm
Integrate & Fire	(12)	6	$4,84$ (6,2.10^5, 8,1.10^5)	$6 * 6$ synapses	1200	2
	(27)	1024	2,25 (420, $1{,}8.10^3$)	AER, mémoire RAM $128k * 16$	-	$0,5$
	(52)	16384	- (-, $2{,}0.10^3$)	AER, 4 synapses adaptatives + mémoires analogiques	-	0,13
	(14, 35)	32	$1,6$ ($2{,}5.10^3$, $3{,}3.10^4$)	AER, STDP, mécanisme d'apprentissage, $32 * 8$ synapses	- ([20...100])	$0,8$
	(34)	16	$6,1$ ($5{,}0.10^2$, $3{,}8.10^5$)	AER, STDP, mécanisme d'apprentissage, $16 * 8$ synapses	- ([20...100])	$0,35$
	(11)	128	$69,0$ ($2{,}4.10^3$, $4{,}1.10^5$)	AER, adaptation de la fréquence d'impulsion, $128 * 128$ synapses plastiques bistables	-	$0,35$
	(72)	384	25 (-, $4{,}7.10^4$)	neurones LIF accéléré avec $(384 * 256$ synapses en conductance	-	$0,18$
	(22)	-	- (-, $1{,}2.10^3$)	Neurone numérique, architecture NOC	-	$0,13$
Hodgin-Huxley	(45, 46)	1	3 ($8{,}0.10^5$)	Système hybride	-	$1,2$
	(2)	1	11 (-, -)	Simulation biologique 2 synapses	-	0,8
	(74)	1	4 (-, -)	Système hybride, neurone à 6 conductances	-	-
	(65, 70)	2	14,5 ($1{,}5.10^6$, $6{,}0.10^6$)	4*2 synapses, 5 courants ioniques	-	0,35
	(82)	5	10,5 (-, -)	-	-	0,35
Simplified HH	(50)	1	- ($< 1{,}0.10^5$,-)	-	(60)	$> 1,5$
	(40)	6	0,6 (-, -)	Synapses : 6 excitatrices, 4 toniques, 8 inhibitrices	-	2
	(86)	2400	9 (3750, -)	AER, DAC et RAM externes	645 (0,26) @ 100.10^3 MHz	0,5
	(30)	7200	11,5 (774, -)	-	-	0,25
Perceptron	(84)	4	20,8 (5,2.10^6)	8*4 synapses		1,5
Izhikevich	(87)	202	- (2800, -)		- ([8...40])	0,35

Table I.2: Quelques exemples d'implémentations silicium de l'état de l'art.

hybride biologique/électronique, le neurone doit fonctionner avec des grandeurs réelles (tension, courant et temps). Ceci permet de caractériser le neurone selon les stimuli appliqués en entrée.

Le second objectif vise une simulation à l'échelle d'un réseau. Le neurone peut alors fonctionner de manière accélérée pour permettre une implémentation plus rapide et plus efficace en termes de surface. Les tensions et les courants sont alors choisis afin de simplifier l'intégration sur silicium. L'intérêt d'un réseau ayant un nombre important de neurones consiste à approcher un fonctionnement global similaire à certaines parties du cerveau. Ce dernier traite des flux d'informations en provenance du monde réel et ce sont donc ces applications qui seront naturellement adaptées au traitement des architectures neuromorphiques. Conforté par les développements de rétines et de cochlées artificielles ((13, 47), plusieurs chercheurs testent leurs réseaux logiciels ou matériels grâce aux trains d'impulsions générés par ces deux circuits.

B. Le neuromorphique, une approche *more-than-Moore*

- Résumé - Dans un contexte nécessitant l'exploration d'alternatives aux paradigmes de calcul existants, le neuromorphique est une piste qui permettrait de répondre à deux problèmes majeurs en microélectronique, à savoir la robustesse et la consommation.

On assiste aujourd'hui à des changements critiques dans le domaine de la fabrication des circuits intégrés. La loi de Moore prévoyait une augmentation du nombre de transistors présents dans une puce grâce à la réduction de la longueur du canal, constituant la dimension critique des transistors. Cette approche, qualifiée de *more Moore*, établit que la réduction de la surface permet soit une hausse de la capacité de calcul soit une diminution de la puissance consommée.

Les processeurs actuels, généralement de type Von Neumann, sont composés d'une mémoire, d'un cœur de calcul et d'entrées-sorties montrés sur la figure I.8. Les progrès des dernières décennies, miniaturisation des composants en tête, ont permis à ce type d'architectures de devenir de plus en plus puissantes. Bientôt, on risque d'assister au ralentissement de la diminution de la surface des transistors qui semble indiquer des limites physiques en passe d'être atteintes.

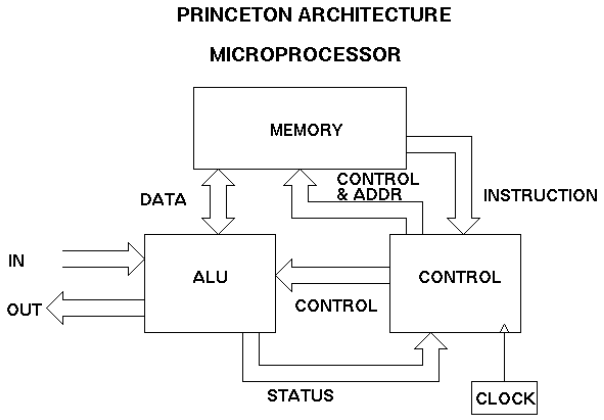


Figure I.8: Architecture de processeur de type Von Neumann appelée également architecture de Princeton. D'après (67).

De fait, l'amélioration d'un système ne se base plus uniquement sur la miniaturisation des composants mais privilégie également la diversification des technologies utilisées ou encore l'emploi d'architectures radicalement différentes. Les architectures neuromorphiques, comme rappelées dans l'International Technology Roadmap for Semiconductors (2011) (36), proposent un paradigme de calcul radicalement différent de celui proposé dans les machines de type Van Neumann, susceptible de correspondre à certains besoins futurs de la micro-électronique.

1. Historique du transistor : miniaturisation et hausse de la fréquence

L'intégration du nombre de transistors MOS par puce suit la loi énoncée par Gordon Moore en 1965 (56). Il avait prédit leur augmentation dans une puce d'un facteur deux tout les deux ans (figure I.10). Pour ce faire, la longueur du canal du transistor a été grandement réduite jusqu'à atteindre aujourd'hui quelques dizaines de nm.

Cependant, la réduction de la taille du canal de grille a eu d'autres conséquences lors la course à la miniaturisation (figure I.9). L'augmentation du dopage du substrat va empêcher que les zones de déplétion entrent en contact. Par conséquent la commande du transistor se fait toujours par la tension de grille et le courant I_{ds} est faiblement dépendant de la tension V_{ds} en saturation.

Afin d'éviter les phénomènes de claquage, la tension d'alimentation V_{dd} doit être réduite. Parallèlement, la tension à appliquer pour modifier le potentiel de surface du

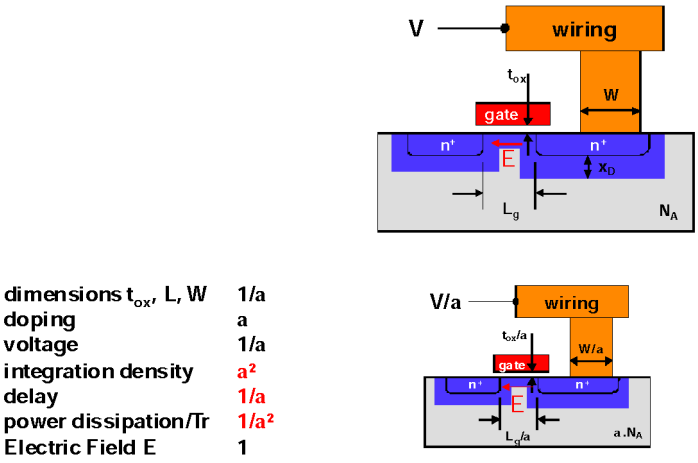


Figure I.9: Évolutions du transistors MOS. Quelques effets de la réduction par un facteur a de la longueur de grille L à champs électrique constant.

canal augmente (réduction de L à champ constant). Le changement de l'état du canal à champ constant est obtenu par la diminution de l'épaisseur de l'oxyde de grille au prix d'un courant de fuite I_g plus important.

A la multiplication du nombre de transistors intégrés par puce s'est ajoutée une hausse de la fréquence de fonctionnement des circuits numériques. En effet, la diminution de la surface du transistor à permis l'augmentation de sa fréquence maximale intrinsèque (f_{max} ou f_t).

Récemment, un problème est apparu dès lors que la puissance consommée par transition est diminuée par le même facteur que la surface. Le système atteint ses limites si l'on utilise les transistors à leur fréquence maximale de fonctionnement. La puissance dynamique dissipée par unité de surface est plus grande et donne lieu à des problèmes de dissipation thermique qui peuvent aboutir à la fonte du circuit, provoquer un vieillissement accéléré ou encore générer des fautes. En effet certains paramètres, comme la mobilité des électrons, dépendent directement de la température et une hausse trop importante de celle-ci peut altérer les spécifications temporelles des portes logiques. Ceci explique en partie le pallier atteint par la fréquence autour de l'année 2005 et que nous verrons sur la figure I.12.

2. Attentes et évolutions de l'industrie microélectronique

Dans la course à l'optimisation des systèmes embarqués, trois paramètres sont à prendre en compte lors de la réalisation d'un circuit :

- La hausse de la densité d'intégration : une augmentation du nombre de composants intégrés permet d'étendre la capacité de calcul de l'architecture
- La faible consommation : la recherche d'une autonomie accrue est une contrainte majeure pour tout système embarqué.
- La robustesse face aux défauts de fabrication : un rendement élevé et une faible variabilité vis à vis des procédés de fabrication sont autant d'atouts pour une architecture dont la technologie est de plus en plus soumise à des phénomènes statistiques.

Jusqu'à aujourd'hui, c'est grâce à l'amélioration physique du transistor, notamment la diminution de sa surface, que les gains en puissance de calcul ont été si importants. Plus récemment, l'impératif d'une faible consommation électrique est apparu pour répondre aux contraintes de dissipation thermique d'une part mais également d'autonomie pour les dispositifs mobiles d'autre part.

Si dans les nœuds technologiques actuels, la variabilité des composants fait partie intégrante du flot de conception, elle se fait au prix de temps de simulations élevés. En effet, les très faibles dimensions mises en jeu lors des procédés de fabrication donnent naissance à des phénomènes aléatoires. Leur impact peut être estimé par des méthodes de types Monte-Carlo ou encore des analyses temporelles statiques statistiques.

Des procédés de fabrication mal maitrisés peuvent mener à des performances moindres qu'espérées en termes de fréquence maximale ou de consommation, voire dans le pire des cas à un circuit défectueux. Plusieurs méthodes ont été adoptées pour que le composant reste fonctionnel et pallier ainsi les erreurs provoquées par ces défauts. A titre d'exemple, on peut envisager de fonctionner à différentes fréquences, de doubler les chemins critiques ou encore de désactiver de manière volontaire une partie du circuit. Si les deux dernières options doivent être considérées avant la conception du composant, elles requièrent toutes une surface de silicium redondante et donc un surcoût pour les puces fabriquées.

Ces limites (densité d'intégration, consommation, variabilité) ont été sans cesse repoussées depuis les années 60. Si, comme on l'a précédemment énoncée, la miniatu-

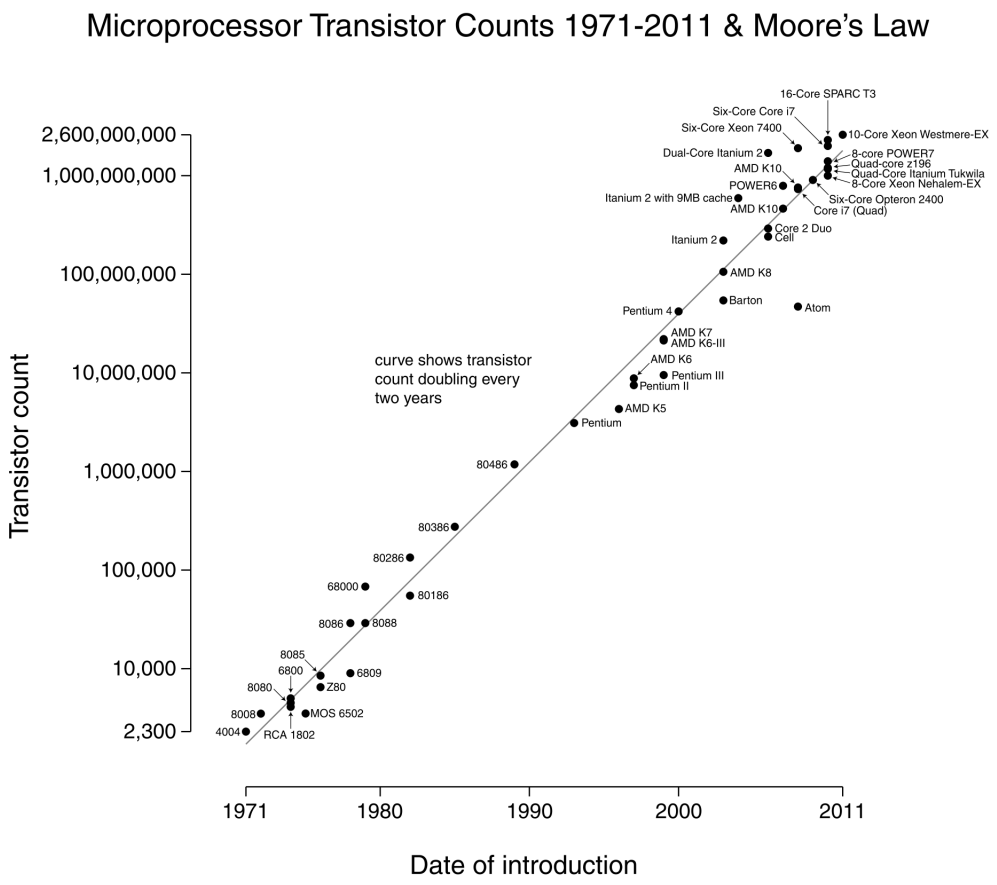


Figure I.10: Loi de Moore et processeurs : tout les deux ans, le nombre de transistors double au sein d'une puce.

risation des transistors est la source principale des progrès de la micro-électronique, on peut également souligner les avancées des domaines connexes tels que l'informatique, les architectures ou encore la physique. En prenant exemple sur une architecture de calcul générique, on voit que ces disciplines ont permis d'en améliorer les performances globales, et ceci à différents niveaux :

- L'optimisation du logiciel : parallélisation de l'exécution des tâches, exploitation de la technologie multi-cœur pour un seul logiciel...
- La conception de l'architecture : utilisation de logique asynchrone, développement de système hybride NEMS/MEMS et d'architectures multi-cœur (SOC et NOC)...
- La maitrise de la réalisation physique : augmentation de la densité d'intégration,

maitrise de la variabilité...

Nous allons nous intéresser dans les prochaines parties aux améliorations liées à la conception de l'architecture ainsi qu'à l'utilisation de technologies émergentes pour les architectures neuromorphiques. On notera que, dans certains cas, le développement de nouvelles technologies vont de pair avec la conceptualisation d'un nouveau type d'architecture.

3. Les limites physiques du transistor MOS en passe d'être atteintes

La diminution de la largeur de grille, longueur critique dans un circuit électronique, est devenu un véritable défi à chaque nœud technologique. La compréhension et surtout le contrôle des phénomènes physicochimiques impliqués dans chaque étape de fabrication devient de plus en plus difficile. L'ITRS (2010) prévoyait en effet la fin du transistor MOS plan sur substrat silicium dans les année 2015 avec une longueur $L = 20\ nm$.

Pour permettre une plus grande densité d'intégration, deux alternatives sont actuellement envisagées. En plus de répondre aux problèmes de variabilité, le silicium sur isolant (SOI, ST) dont le canal est en déplétion partielle ou totale apparait comme un candidat permettant de descendre dans les nœuds plus agressifs (36). Le contrôle électrostatique du canal, affiné par le procédé, est amélioré ; ceci permet de réduire les courants de fuite et donc la puissance statique.

Les transistors à plusieurs grilles ou à grille enrobante ont également les mêmes qualités et sont la stratégie d'autres groupes (Intel). Cependant la mise en œuvre est moins directe car les procédés de fabrication de l'empilement de la grille doivent être adaptés au passage en 3D du transistor. Aujourd'hui, la longueur du canal de ces transistors est de 22 nm (38) dans certains processeurs. Néanmoins, les premiers circuits en 14 nm sont en cours de développement et devraient sortir courant 2013.

Ces 2 options sont des alternatives au CMOS plan sur substrat massif et sont des solutions de type *"more Moore"*. Il s'agit en effet de la transposition de fonctions existantes vers des dimensions toujours plus réduites. Le concept de *"more than Moore"* propose d'améliorer les systèmes existants en diversifiant les fonctions intégrées (figure I.11).

4. Un environnement propice au développement du neuromorphique

La figure I.12 mets en évidence le changement brutal qui atteint les processeurs. Si le nombre de transistors total est toujours en augmentation, on observe la stagnation de

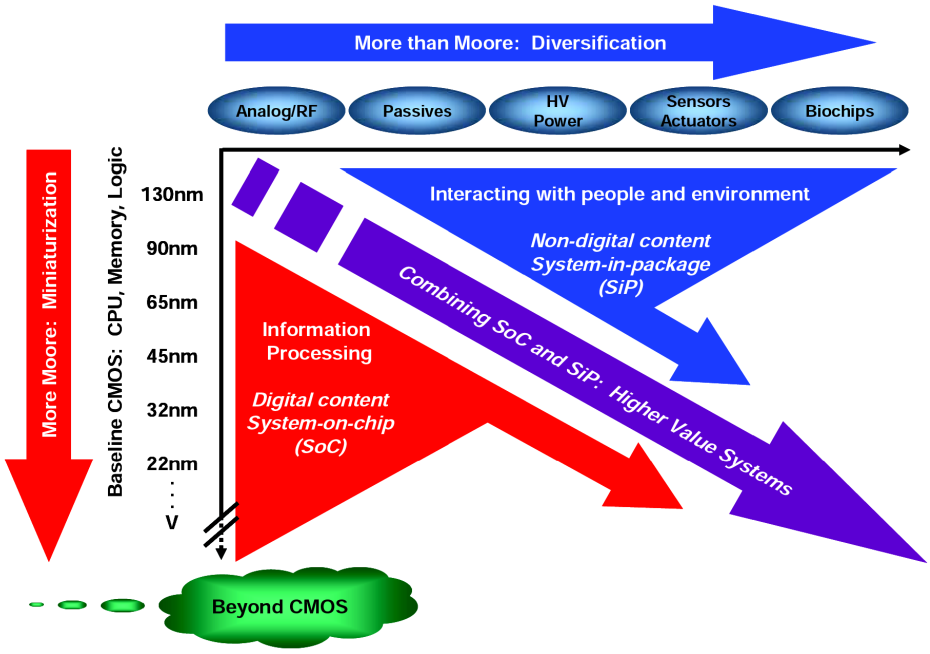


Figure I.11: La diversification : en plus de la miniaturisation toujours plus poussée (*"more Moore"*), on étend les domaines de compétences employés dans le développement d'un circuit. Les solutions proposées se trouvent hors des solutions usuels de la microélectronique (*"more than Moore"*). D'après (36).

la consommation comme de la fréquence. Parallèlement, on assiste à une augmentation du nombre de cœurs afin d'exécuter plusieurs tâches simultanément mais sans pour autant augmenter leur performance individuelle.

a) Une modification du type d'applications

Les processeurs ont été longtemps utilisés pour des applications scientifiques déterministes complexes et requérant une grande précision. On peut citer à titre d'exemples la modélisation physique, la cryptographie ou le décodage de séquence ADN. Ces différentes applications typiques sont regroupées dans des jeux de tests qui permettent de fournir une figure de mérite d'un processeur.

SPEC est l'un d'entre eux (78) ; il recense plusieurs algorithmes couramment utilisés comme A* (473.astar) qui permet de trouver le plus court chemin entre deux points, ou encore la compression de donnée (401.bzip2). Il permet de tester des processeurs mais également des GPU, des serveurs de messagerie, des machines virtuelles Java, etc.

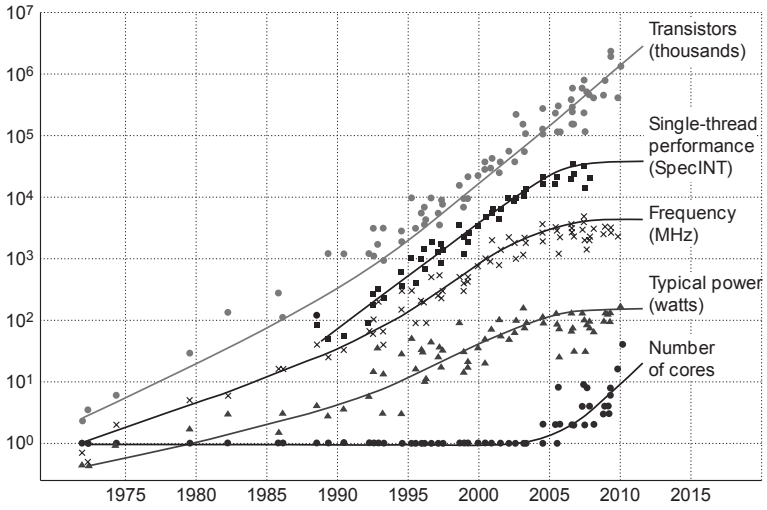


Figure I.12: Évolution du nombre de transistors, de la performance d'exécution d'une tâche, de la fréquence, de la consommation et du nombre de cœurs pour les microprocesseurs sur les 45 dernières années. D'après (43).

Conjointement à la multiplication des cœurs dans les processeurs, les besoins applicatifs se sont radicalement modifiés. La majorité d'entre eux ne servent plus à faire du calcul scientifique, la performance calculatoire n'est plus l'impératif unique. Les applications utilisées quotidiennement dans la plupart des appareils sont en interaction directe avec notre environnement et donc avec les êtres humains. Les futurs architectures doivent répondre à des impératifs tels que la consommation et l'exécution en temps réel sans nécessiter forcément une grande précision.

De nouveaux protocoles de test ont vu le jour pour permettre d'évaluer ces nouvelles architectures hautement parallèles. Ainsi, PARSEC (6) a été créé conjointement par l'université de Parse et Intel pour compléter les tests SPEC et propose un nouvel ensemble de tests en cohérence avec les besoins applicatifs actuels.

Pour ces nouveaux calculs approchés, le cerveau surpasse les machines dans l'environnement réel. Par exemple, il peut reconnaitre un objet rapidement, quelque soit sa forme ou sa position. Cette capacité découle, entre autres, de l'extrême parallélisme du fonctionnement du cerveau.

Si l'on veut être optimal d'un point de vue énergétique dans le traitement d'informations du monde réel, sujet aux aléas et au bruit, celui-ci n'a pas besoin d'être réalisé de manière précise et/ou accélérée (23).

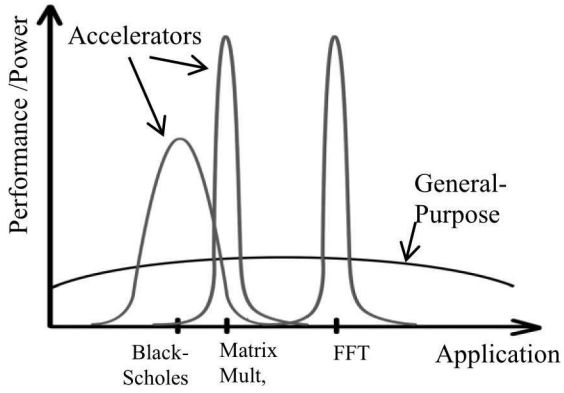


Figure I.13: En visant une application, un accélérateur matériel accroit ses performances énergétiques. D'après (89).

b) Un accélérateur matériel au sein d'un système sur puce

Les accélérateurs matériels sont des circuits dédiés à certaines applications. Ils possèdent de meilleures performances, en termes d'énergie, comme présenté sur la figure I.13, mais également de surface, de vitesse, etc. Ces points forts se font au prix d'un nombre réduit d'applications qu'ils peuvent executer. Par conséquent, l'intégration d'un accélérateur pour chaque application d'un processeur n'est pas envisageable puisque limitée par la surface du circuit.

Comme la surface de silicium a un coût, on préfère la limiter en procurant une certaine programmabilité au circuit au détriment de son efficacité énergétique. Ainsi un processeur pourra effectuer un large spectre d'applications mais verra ses performances diminuées. Ceci s'explique par l'utilisation d'instructions ou, à plus haut niveau, d'un langage de programmation qui servent d'interface entre une application visée et une unité de calcul générique.

Dans un futur proche (58), l'échauffement d'un circuit ne permettra plus d'utiliser simultanément tous les transistors intégrés. Dans ce contexte, les accélérateurs permettront soit d'utiliser de manière optimale l'énergie disponible soit de conserver une surface inactive améliorant le refroidissement d'une autre partie du circuit. Par conséquent, les concepteurs de circuits étudient aujourd'hui la réalisation d'architecture multi-cœurs hétérogènes.

c) Une architecture neuromorphique au sein d'un système sur puce

L'introduction d'une architecture neuromorphique au sein d'un système sur puce permet de résoudre à la fois les problèmes de variabilité et de consommation (28). Ces architectures à base de neurones répondent également aux besoins applicatifs actuels de traitement du signal.

La robustesse face aux défauts de fabrications est acquise puisque l'architecture est massivement distribuée. La mémoire, souvent identifiée aux poids synaptique, est ainsi distribuée à l'instar des cœurs de calculs, les neurones. En rapprochant physiquement le cœur de calcul et sa mémoire, on envisage également de réduire l'énergie nécessaire à la réalisation d'une application. Les réseaux de neurones peuvent ainsi résoudre de manière efficace une majeure partie des applications précédemment citées à savoir le traitement en temps réels de données.

d) Quelle architecture neuromorphique ?

La figure I.14 nous interroge sur l'architecture neuromorphique idéale en se référant au développement aéronautique. La nature fait bien les choses, certes, mais elles n'a pas les mêmes moyens ni les mêmes impératifs. Un oiseau vole grâce à une forme aérodynamique, un plumage, et des muscles. Par mimétisme, les premières tentatives de l'homme sont trop calquées sur la nature et il en découle des résultats peu probants. L'articulation d'ailes mécaniques sur l'ornithoptère n'est pas des plus fiables et ne lui permet pas de pouvoir modifier sa propre direction. Sur un produit mature comme l'A380 d'Airbus, tout les éléments critiques ont été modifiés : propulsion, atterrissage et direction. Il ne restera au final qu'une vague allure du modèle biologique. De manière similaire, il peut être judicieux de s'éloigner du fonctionnement du cerveau afin de profiter pleinement de l'implémentation silicium d'une architecture neuromorphique. On sera alors en mesure d'être optimal en termes de robustesse, d'énergie, de programmabilité et de surface.

L'étude du cerveau a commencé bien plus tard mais, du fait de sa complexité, son observation est des plus critiques. La plupart de ses constituants sont identifiés sans que l'on puisse, à l'heure actuelle, en établir une cartographie précise. La compréhension de la totalité des mécanismes biologiques est également difficile. Ceci dit, leur reproduction à l'identique dans une architecture neuromorphique n'est pas nécessaire à des fins

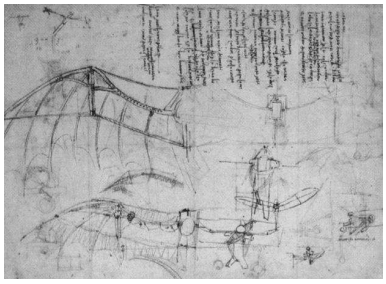
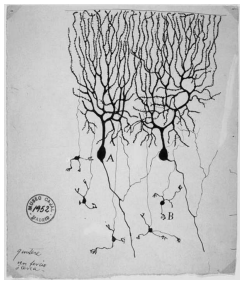
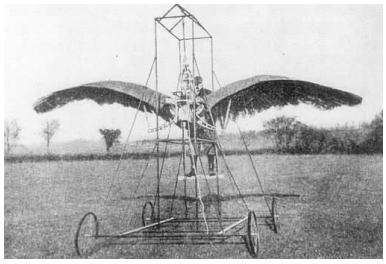
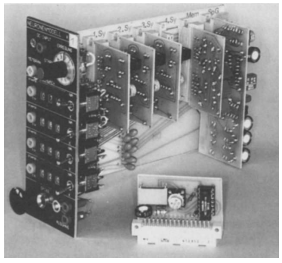
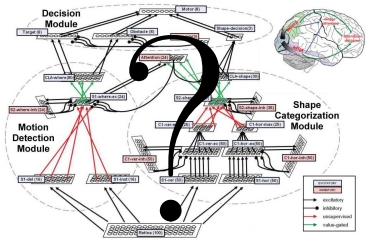
<table>
<tr><td>  </td><td>  </td><td rowspan="2">De la biologie</td></tr>
<tr><td>Goéland</td><td>Un illustre cerveau en vélo</td></tr>
<tr><td>  </td><td>  </td><td rowspan="2">à la conceptualisation</td></tr>
<tr><td>Design of a flying machine,
Leonard de Vinci, 1488</td><td>Cellules de Purkinje,
Santiago Ramón y Cajal, 1899</td></tr>
<tr><td>  </td><td>  </td><td rowspan="2">et aux premiers essais</td></tr>
<tr><td>Ornithoptère d'Edward Frost, 1902</td><td>Un neurone pour la recherche et l'enseignement (42)</td></tr>
<tr><td>  </td><td> 
</td><td rowspan="2">pour l'industrialisation...</td></tr>
<tr><td>Airbus A380 en 2010</td><td>Système neuromorphique (60)</td></tr>
</table>

Figure I.14: En s'inspirant de la biologie, l'homme a pu voler. Le modèle le plus fidèle n'est pas toujours le plus efficace pour le but recherché. Qu'en est il du neuromorphique ?

applicatives, et il est possible de construire un système répondant à une certaine spécification. Lors de la phase de conception, il est impératif de s'interroger sur l'objectif à atteindre. De cette façon, on évitera une implémentation trop précise, couteuse ou encore inadaptée pour obtenir une architecture optimale en fonction de l'objectif visé.

C. Objectifs de cette thèse

Cette thèse s'est déroulée selon deux grands axes de recherche. Le premier suit le cadre établi par le projet Arch^{2}Neu dont on détaillera ci-dessous les objectifs. Il a ensuite été question d'évaluer l'impact des nœuds technologiques avancés dans une approche *more Moore* mais également l'influence de technologies émergentes pour les architectures neuromorphiques.

1. Le projet Arch^{2}Neu :

Face aux accélérateurs matériels de types FPGA (reprogrammable) ou systèmes sur puce (hétérogène), un circuit intégré neuromorphique permettrait de prendre en charge deux problèmes récurrents en micro-électronique. L'objectif du projet Arch^{2}Neu était de montrer l'intérêt d'une architecture neuromorphique employée comme accélérateur matériel programmable.

En réponse au premier point, les réseaux de neurones utilisent un codage de l'information par fréquence d'impulsions, et sont basés sur une architecture hautement décentralisée. La distribution et l'entrelacement de la mémoire au sein des cœurs de calcul, auxquels s'ajoutent une communication robuste, devraient pallier les futurs aléas de fabrication de la microélectronique.

Notre architecture neuromorphique doit également répondre aux contraintes de consommation. Les applications réalisables étant fonction du nombre de neurones, il en résulte des contraintes de densité d'intégration d'une part et de faible consommation énergétique d'autre part.

Afin d'utiliser une architecture neuromorphique à des fins applicatives, il est donc important de la considérer comme étant constituée de 3 parties :

- le routage des potentiels d'actions
- les neurones
- les synapses

Au cours de cette thèse, on s'est tout particulièrement intéressé à l'implémentation des neurones et de leurs synapses.

A la manière de l'implémentation d'une application sur un FPGA, il est nécessaire de prévoir une chaine de programmation et de compilation pour permettre l'utilisation d'un grand nombre de neurones pour la réalisation de tâches complexes, et pour établir une matrice de connectivité optimisée. Ce flot, indiqué sur la figure I.15, permettra à un utilisateur de programmer un réseau de neurones à impulsion par le biais d'une librairie d'opérateurs. Après décomposition en fonctions élémentaires, une opération complexe (comme une FFT par exemple) pourra être implémentée.

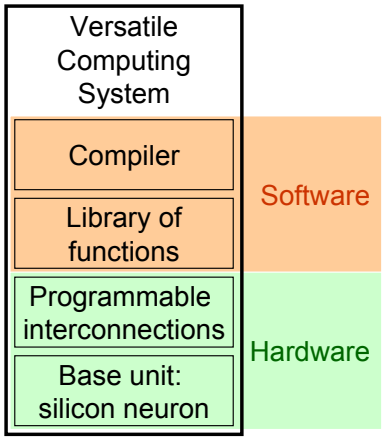


Figure I.15: Flot de programmation d'une architecture neuro-inspirée.

2. Contributions au projet Arch^{2}Neu.

On verra dans le prochain chapitre comment les spécifications de l'architecture ont permis de préciser les caractéristiques du neurone et l'élaboration de son cahier des charges. Sa réalisation a également été fonction de la technologie choisie qui aura donné lieu à certains impératifs. Deux démonstrateurs silicium ont ainsi été fabriqués dont l'un a pu faire l'objet de caractérisations électriques.

3. Etudes des technologies avancées.

Les progrès réalisés dans le cadre d'une approche *more Moore* ont permis de réduire la taille des transistors. Une des approches privilégiées actuellement modifie l'architec-

ture pour la rendre plus parallèle et hétérogène (réseau sur puce, système sur puce).

D'autre technologies apparaissent pour répondre à la diversification des besoins. En effet, la performance n'est actuellement plus la seule priorité : l'énergie et le coût sont devenus des contraintes majeurs. Ainsi le développement des circuits plastiques, par opposition au silicium, cible un marché qui requiert un faible coût de production au détriment de la performance et de la surface.

A contrario d'autres pistes se basent sur le flot de fabrication silicium existant. L'exploitation de la troisième dimension à l'aide de *Through-Silicon-Vias* (TSV's) permet, par exemple, de diminuer la puissance dissipée lors de la communication entre deux puces.

On observera l'influence des avancées technologiques sur une architecture neuromorphique, en explorant l'impact d'un passage vers des technologies plus avancées. On étudiera également l'implication de composants plus originaux, basés sur des technologies 3D ou memristives.

Conclusion

Cette partie avait pour but d'apporter les bases de biologie et d'électronique nécessaires à la lecture de ce manuscrit. Après avoir décrit le contexte favorable au développement d'un accélérateur neuromorphique, nous allons maintenant nous intéresser à l'implémentation du neurone dans le cadre du projet Arch^{2}Neu.

II

Intégration d'un neurone robuste pour des applications computationnelles

Peut-être fabriquerons-nous un jour ce qui nous comprendra.

Jean Rostand, *Pensées d'un biologiste*, 1939.

Sommaire

A. Quelques notions de conception en microélectronique 34
1. Flot de développement 34
2. L'analogique et le numérique 38
3. Quelques exemples d'implémentations analogiques de neurones proposées dans la littérature 40
B. Conception d'un neurone LIF robuste 44
1. Spécification des caractéristiques du neurone 44
2. Choix et réalisation des structures élémentaires 47
3. Fonctionnement du neurone 55
4. Résultats de simulation 60
5. Réalisation du layout du neurone 66
C. Circuits réalisés 68
1. Implémentation dans le circuit *Reptile* 68
2. Caractérisation et test 72
3. Évolutions dans le circuit *Spider* 81

Le concept d'accélérateur neuromorphique soutenu dans le projet Arch^{2}Neu se base sur des neurones à impulsions pour être robuste et présenter une faible consommation (28). Cet accélérateur peut effectuer un large spectre de tâches, notamment dans le domaine du traitement du signal, en décomposant celles-ci en opérations élémentaires réalisables par des neurones.

Dans ce chapitre, nous verrons tout d'abord l'environnement de développement utilisé en conception microélectronique. Nous détaillerons ensuite l'élaboration et le dimensionnement des neurones dans le premier circuit réalisé, nommé *Reptile*. Les résultats de sa caractérisation auront permis de concevoir un deuxième circuit, *Spider*, l'aboutissement du projet Arch^{2}Neu.

A. Quelques notions de conception en microélectronique

1. Flot de développement

Cette partie détaille l'environnement de conception (Design Kit ou DK) utilisé pour la réalisation des circuits du projet Arch^{2}neu (*Reptile* et *Spider*). Le DK ST65 v.5.6.4 développé par *ST microelectronics* est utilisé. On verra par la suite les différents logiciels employés lors des étapes de développement. La procédure de conception est montrée dans le tableau II.1, détaillant les principales étapes.

Phase	Objectif	Outil(s)
Élaboration	Création du schéma	Cadence Virtuoso Schematic Editor
Simulation pré-layout	Validation comportementale + étude de la consommation et impact de la variabilité	Cadence ADE Eldo
Layout	Elaboration	Cadence Virtuoso Layout XL
	Vérification	Mentor Calibre
	Extraction des parasites	Synopsis Star-rcxt
Simulation post-layout	Simulation sur schéma extrait comportement + consommation	Cadence ADE Eldo

Table II.1: Flot de conception utilisé.

a) Généralités

Cette technologie permet l'emploi de transistors pouvant fonctionner à différentes tensions d'alimentation allant de 1 à 2.5 V. La partie numérique étant alimentée à

1.2 V, on utilisera par conséquent cette même tension pour les neurones. La liaison entre les neurones analogiques et la partie numérique sera ainsi facilitée.

La partie numérique est conçue à l'aide de cellules (ou portes) standardisées réalisant des opérations booléennes de type AND, OR, etc. Pour les connecter et plus généralement relier des transistors entre eux, on peut utiliser sept niveaux de métaux différents. Le premier niveau de métal (M1) est utilisé au sein des cellules standards du design kit. Les niveaux 2 à 5 sont utilisés pour le routage des signaux entre les cellules. Les niveaux 6 et 7, organisés en grille, servent à l'alimentation des cellules numériques du circuit pour les tensions de référence Vdd et Gnd.

Le design kit propose plusieurs options d'implémentations de capacités illustrées sur la figure II.1. Ce choix est déterminant pour le bon fonctionnement du neurone, et sera détaillé en II-B.2.a). Les capacités peuvent être de type MOS lorsque l'on utilise la grille d'un transistor dont on court-circuite le drain et la source. A contrario, les capacités MIM sont uniquement métalliques et se déclinent sous différentes architectures. La première utilise les métaux de routage et l'oxyde les séparant. Elles sont alors planes ou réalisées à l'aide de doigts interdisposés. La deuxième, optionnelle dans le design kit, est une capacité placée entre les métaux 5 et 6.

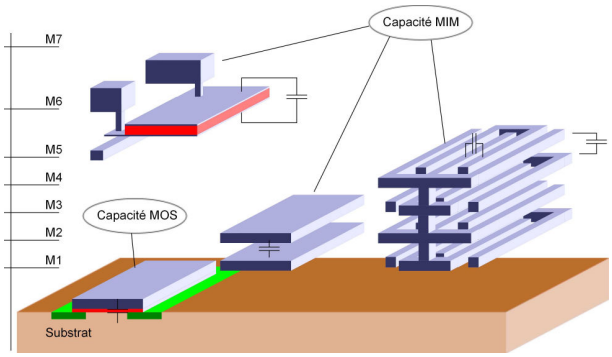


Figure II.1: Ce schéma illustre les différentes capacités disponibles. Les capacités MOS occupent une surface de silicium. A contrario, les capacités MIM utilisent les métaux de routage.

La longueur minimale de grille des transistors est de 65 nm. Lors d'une conception d'une partie analogique ou pour diminuer l'impact de la variabilité, on pourra prendre une grandeur plus élevée. Les transistors sont déclinés sous plusieurs versions : low Vt, standard Vt, High Vt, High Performance Analog. Pour les 3 premiers, un changement du

dopage décale la tension de seuil V_{th} des transistors et permet d'optimiser la puissance dissipée. Le dernier a une longueur de grille minimale de 140 nm et un courant I_{on} plus élevé.

b) Conception pre-layout

Le DK ST 65 permet d'utiliser tous les logiciels nécessaires à la conception et à la vérification. Ainsi le dimensionnement se fait graphiquement au niveau du schéma à l'aide du logiciel *Virtuoso schematic editor* et permet d'obtenir une *netlist*. La simulation a été réalisé à l'aide d'ADE (Analog Design Environment) et du simulateur par défaut *Eldo* (Mentor Graphics). Le comportement du bloc doit à ce stade respecter les spécifications dans les conditions nominales.

On réalise ensuite des analyses de type PVT (*Process-Voltage-Temperature*). Le circuit doit continuer à respecter ses spécifications avec des tensions d'alimentation autres (généralement + ou − 10%) ou des variations de température (-40 à $120°C$). L'étude de l'impact des procédés de fabrication est réalisée de 2 manières. La première est de type "pire cas" (corners) : les transistors peuvent avoir des caractéristiques nominales (**NOM**), rapides (**F**ast) ou lentes (**S**low). Cette méthode, particulièrement utile pour la conception de circuits numériques, combine ces caractéristiques pour chaque type de transistors d'une porte. On obtient alors les corners NOM, FF, SS, FS, SF où les lettres représentent respectivement un transistor de type P et de type N. A noter qu'il existe des corners surévaluant l'impact des aléas de fabrication et permet la conception de circuits analogiques sensibles. La lettre A est alors ajouté aux noms de corners précédents et devient alors FFA, SSA, etc. L'autre méthode utilise des tirages aléatoires (Monte Carlo) des paramètres physiques des transistors (épaisseur de l'oxyde de grille,...). Le logiciel définit une valeur selon la probalité définie par une distribution gaussienne, fonction de la moyenne et de l'écart type du paramètre. Les études de fonctionnalités robustes aux variations de tensions et de températures sont fondamentales lors de la réalisations de circuits fonctionnant sur batteries et/ou dans des conditions extrêmes. Pour les démonstrateurs fabriqués, alimentés par des générateurs fixes et ne souffrant pas de problèmes d'échauffement, ces études ont été réalisées seulement dans le cas nominal.

Le layout, dessin physique des masques permettant la fabrication des différents niveaux du circuit, se fait dans *Virtuoso Layout Editor*. Tout ces niveaux sont dessinés

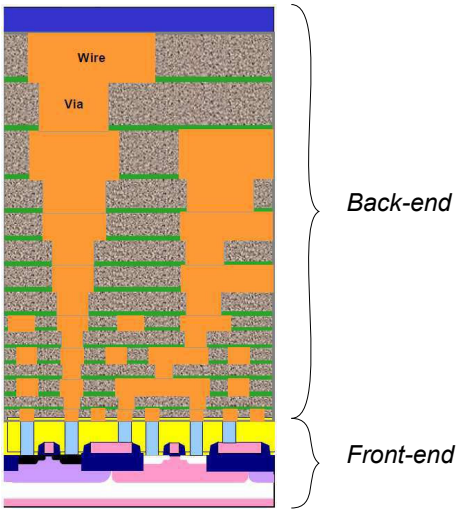


Figure II.2: Les transistors, appartenant au *front-end*, sont connectés au *back-end* par des fils métalliques. D'après (36)

en partant en général du bas, *Front-End*, vers le haut, *Back-End*. Une vue en coupe de la figure II.2 fait référence à cet empilement. Le *Front-End* contient les transistors et plus généralement tout ce qui se trouve sur le substrat. On dessinera à ce niveau les zones de dopages, les grilles, les caissons, etc. Le *Back-End* concerne la partie de routage des signaux à l'aide des métaux décrits précédemment. Nous allons maintenant décrire le flot permettant le contrôle du layout créé.

c) Vérification - Layout & Post-layout

En plus des vérifications comportementales à l'aide des simulations, il existe d'autres tests réalisés à la fois lors du layout et après celui-ci. Les deux premiers sont réalisés par le logiciel Calibre distribué par Mentor Graphics. Le test DRC (Design Rule Check) vérifie le respect des contraintes géométriques liées à la technologie de fabrication. A titre d'exemple, on peut citer les règles d'espacement, de densité ou encore de surface. Les règles de densités sont respectées par ajout de cellules ou de tuiles de matériaux (polysilicium, métaux,etc).

L'étape LVS (Layout Versus Schematic), consiste en l'extraction d'une *netlist* correspondant au layout puis à sa comparaison avec celle générée par le schéma du circuit. A ce stade, on s'assure de la bonne connectivité des composants entre eux, ainsi que de leurs propriétés (W, L, ...).

L'extraction de parasites consiste en la transcription du layout au schéma en tenant compte de la proximité physique des composants. Elle se fait à l'aide de Star-RCXT de Synopsys. L'exemple de la figure II.3 montre le couplage de fils de métal ainsi que les capacités parasites générées. Une simulation du schéma contenant ces parasites permettra ensuite de vérifier que ces derniers n'influent pas sur le comportement du bloc.

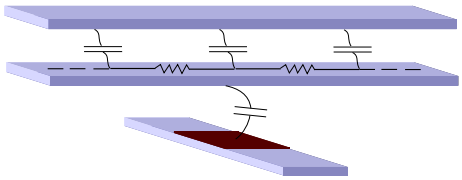


Figure II.3: Phénomènes de couplage pouvant existant lors de la réalisation physique d'un circuit. On montre ici un cas classique : la formation d'une capacité parasite lorsque deux fils se superposent ou se croisent.

Une fois le comportement d'un bloc validé par des analyses PVT et post-Layout, il peut être intégré dans un circuit. Lors de leur placement, il peut être judicieux de regrouper les blocs sensibles voire même de les isoler à l'aide de caissons ou d'anneaux de garde qui auront pour effet d'absorber des charges parasites du substrat. Ceci est particulièrement employé pour les blocs analogiques qui peuvent être sensibles aux bruits d'horloges numériques.

2. L'analogique et le numérique

Nous rappelons les principales différences entre l'électronique numérique et analogique dans le tableau II.2. Ces deux domaines sont complémentaires et indissociables puisque notre environnement réel est spatialement et temporellement analogique. Les processus de discrétisation et d'échantillonnage des signaux en vue de leur utilisation dans des systèmes numériques sont particulièrement adaptés pour du calcul précis et déterministe. Cependant l'opération de conversion a un coût et n'est pas nécessairement requise pour toutes les utilisations.

Selon le nœud technologique utilisé, les figures II.4 montrent les intérêts des deux domaines en indiquant les points d'équivalence de surface et de consommation en fonction du rapport signal à bruit (sur la figure à 60 dB). Si ce rapport est inférieur,

Analogique	Numérique
Valeur continue, généralement comprise entre V_{th} et $V_{dd} - V_{th}$	Valeur discrète 0 ou 1, G_{nd} V_{dd}
Utilisation directe des propriétés physiques des composants électroniques	Utilisation de l'algèbre booléenne (porte OU, ET, ...)
Un nœud peut contenir plusieurs bits d'information à un instant donné	Chaque fil transporte un bit d'information à un instant donné
De la variabilité entre les composants résulte une fenêtre d'erreur	La variabilité a peu d'importance, cependant une erreur sur un bit de poids fort est dramatique
Le bruit provient des fluctuations thermiques au sein des transistors	Le bruit est introduit lors de l'arrondissement des valeurs
Le bruit est amplifié d'étage en étage	Le signal est régénéré à chaque porte
Un système avec de nombreux étages en cascade devient complexe à réaliser	Le chainage des différentes portes est aisé

Table II.2: Quelques propriétés de l'analogique et du numérique. (71)

l'implémentation analogique a l'avantage en termes d'énergie et de surface.Au delà, on constate d'une part que l'implémentation numérique prévaut et d'autre part qu'il existe une limite pour l'analogique provenant du bruit thermique ou de scintillation.

Le point d'équivalence indiqué à 60 dB, correspondant à 10 bits, semble décroitre selon les nœuds technologiques. Contrairement au numérique, la passage à l'échelle d'un circuit analogique n'est pas trivial et ne permet pas de diminuer la consommation et la surface de manière aussi directe. En conception analogique, et pour des questions de variabilité, il est préconisé d'utiliser une longueur de grille de l'ordre de deux fois supérieure à la longueur minimale. Ceci explique l'emploi recommandé d'une longueur de grille minimale de 140 nm pour les transistors analogiques fournis dans le DK. Cependant, il a été observé à partir du nœud 90 nm que cette recommandation n'est plus aussi efficiente puisque la dispersion du courant généré par un miroir de courant a tendance à augmenter (4). Pour pallier ce phénomène, on augmente de fait, soit la consommation soit la surface. On peut également citer l'ITRS (36) dont les prévisions sur la dispersion de la tension de seuil ($A_{V_{th}}$) et sur la mobilité (A_β) indiquent la même tendance. On verra dans la partie III-A. quelles seraient les caractéristiques possibles d'un neurone dans des nœuds avancés.

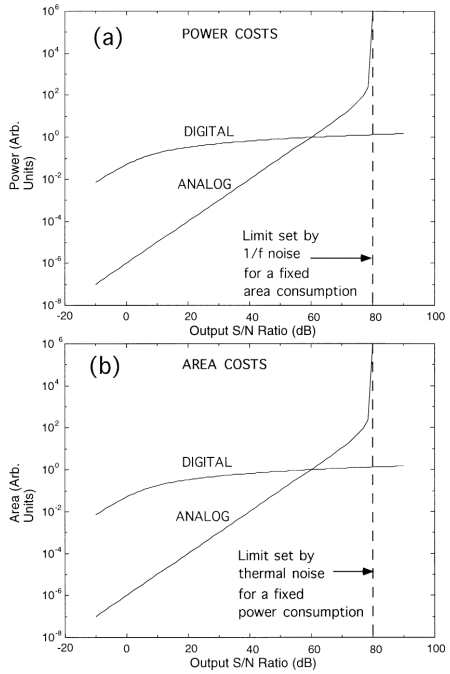


Figure II.4: Analogique et numérique. Figure (a) : Considération énergétique. La limite analogique est fixée par un bruit en 1/f. Figure (b) : Considération surfacique. La limite est alors contrainte par le bruit thermique (71)

Les conséquences du passage à l'échelle pour l'analogique sont également moins bénéfiques en termes de consommation. La diminution de la tension d'alimentation réduit les possibilités de monter les étages d'opérations en cascode. Par conséquent, ceux-ci sont mis en cascade et débitent davantage de courant.

Les applications envisagées requièrent une faible précision. Par conséquent, le rapport signal sur bruit du neurone électronique est relativement bas et enclin à une implémentation analogique.

3. Quelques exemples d'implémentations analogiques de neurones proposées dans la littérature

Pour préparer la conception de notre neurone, on détaille les implémentations proposées par différents groupes de recherche. Celles-ci seront mises en relation et étayeront les choix effectués pour notre futur implémentation de neurone. On observe la présence de blocs récurrents dans plusieurs implémentations. Ceux-ci sont mis en évidence par

des couleurs identiques sur les 3 premiers schémas alors que le dernier contient directement le nom de la fonction implémentée. La capacité de membrane C_m, en orange, constitue le cœur du neurone et permet le stockage des charges.

a) Neurone 1 - *Van Schaik et al.* - (85)

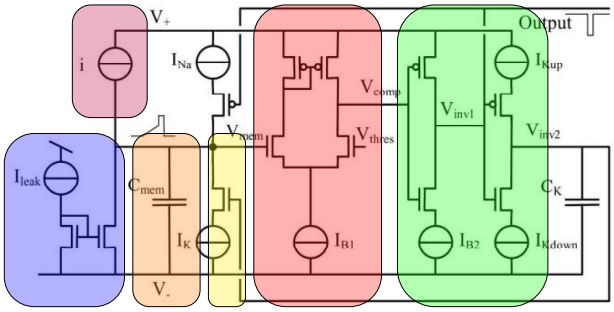


Figure II.5: Schéma d'un neurone d'après *Van Schaik* (85).

Cette implémentation simplifie le modèle d'Hodgkin-Huxley aux deux principaux canaux ioniques : potassique (K^+) et sodique (Na^+). Il a l'avantage d'être particulièrement simple puisque les différents sous-blocs fonctionnels sont facilement identifiables et connus. L'intégration du courant, en violet, dans la capacité C_{mem} est comparé à $V_{threshold}$ par le biais de l'amplificateur à transconductance en rouge. Afin de raccourcir le temps de basculement signifiant l'émission d'un spike, et donc de réduire la consommation, la sortie d'un premier inverseur établit une boucle qui va injecter le courant I_{Na}. La sortie du second inverseur permet la remise à zéro du neurone selon une dynamique fonction de la capacité C_K et des sources de courants I_{Kup} et I_{Kdown}. Le bloc de fuite permet, lors d'un stimulus trop faible, le retour à l'équilibre du neurone par le biais du miroir de courant en bleu sur le schéma.

Dans cet exemple, seuls des poids positifs peuvent être injectés et le détail de leurs modulations n'est pas décrit. On peut tout à fait imaginer une modulation en amplitude, en temps ou les deux.

b) Neurone 2 - *Vogelstein et al.* - (86)

Cette implémentation simplifie le modèle d'Hodgkin-Huxley en mutualisant un transfert de charge entre trois capacités, $C_{0,1,2}$ et la capacité de membrane C_m. Les

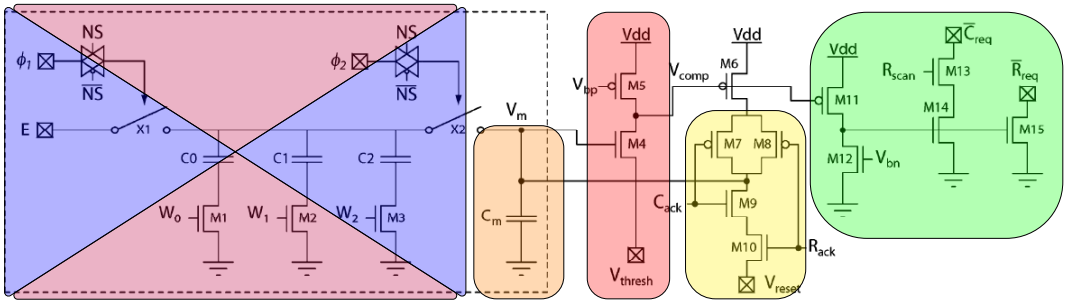


Figure II.6: Schéma d'un neurone d'après *Vogelstein* (86).

blocs bleu ou violet permettent donc, suivant l'ordre d'activation de ϕ_1 et ϕ_2, l'injection ou la soustraction de charges dans C_m si le neurone est sélectionné par le biais du signal NS. Ils jouent donc le rôle des canaux ioniques des neurones en émulant leurs conductances par un transfert de charges. Au dessus d'un certain seuil fonction de V_{thresh} et V_{bp}, le neurone demande l'autorisation d'émettre une impulsion en émettant sa position déterminé par une ligne et une colonne, respectivement R_{req} et C_{req}. Une fois la "poignée de main" établie, le potentiel V_m est remis à zéro et le routage d'impulsions est réalisé par un protocole AER.

Dans ce cas, l'évolution du potentiel de membrane est causée par un convertisseur numérique/analogique (CNA). Les signaux $W_{0,1,2}$ dépendent des valeurs des synapses encodées sur 3 bits dans une mémoire externe. La bande passante du CNA étant fixe, il en découlera une limitation du nombre de neurones.

c) Neurone 3 - *Wijekoon et al.* - (87)

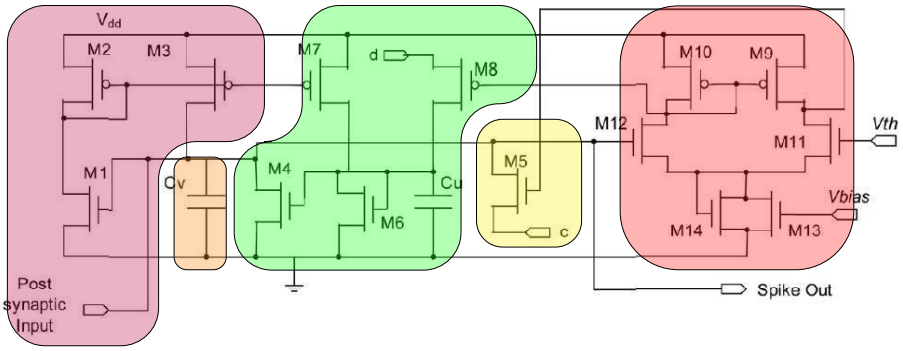


Figure II.7: Schéma d'un neurone d'après *Wijekoon* (87).

Ce schéma représente une implémentation du modèle d'Izhikevich. La capacité C_v est équivalente à la capacité de membrane dont le potentiel suit l'équation I.12. Le bloc vert, contenant la capacité C_u, est un bloc d'adaptation du neurone à mettre en rapport avec l'équation I.13. Les paramètres numériques ainsi que les variables a et b sont alors définis par les valeurs des composants. Les variables c et d utilisées lors de la remise à zéro sont toutefois explicites puisqu'elles permettent au neurone de présenter différents régimes d'émissions d'impulsions. Pour faire évoluer le potentiel de membrane et moduler la fréquence de ses impulsions, le courant post-synaptique est un créneau dont l'amplitude est variable.

d) Neurone 4 - *Livi et al.* - (49)

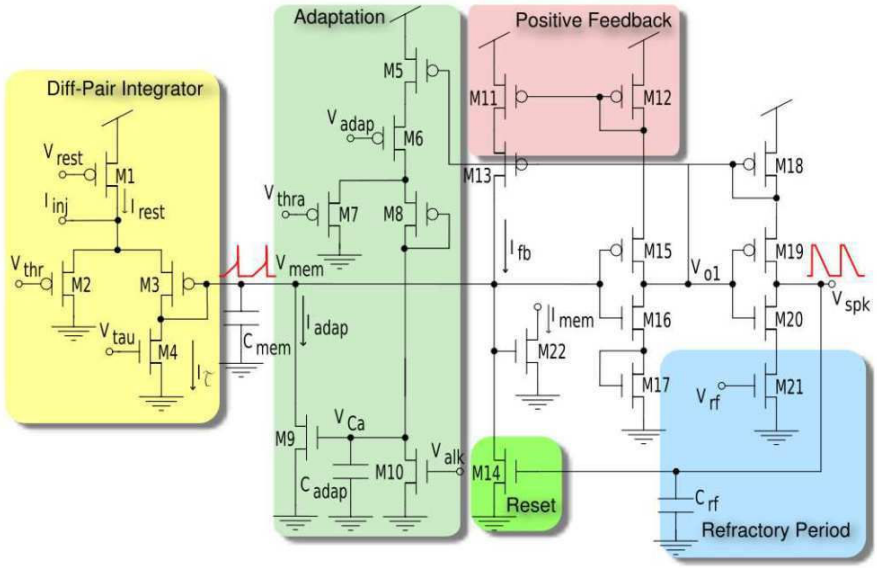


Figure II.8: Schéma d'un neurone d'après *Livi* (49).

Cette implémentation reprend en partie le schéma du neurone 1 puisque le bouclage s'effectue également à l'aide de deux inverseurs. La différence majeure est le passage en courant du neurone, en d'autre terme l'évolution jusque ici de V_{mem} est transposé à I_{mem}. Le courant injecté dans le capacité passe préalablement par le filtre surligné en jaune. Il permet le réglage de la fuite, de la tension de repos et de la tension de seuil. Cette même topologie est utilisée dans l'adaptation du neurone afin de mieux simuler la conductance des canaux calciques.

La modulation temporelle ou de l'amplitude du courant d'entrée n'est également pas précisée.

e) Conclusion

D'un point de vue fonctionnel et extrêmement simplifié, on peut expliciter les différents constituants d'un neurone. La modulation de l'entrée entraîne une évolution du potentiel de membrane. Si celui-ci dépasse un certain seuil, le neurone génère une impulsion qui entraine la remise à zéro du potentiel V_m.

L'étage de comparaison peut se faire de plusieurs manière. Ceci est réalisable directement avec un comparateur, la tension V_{th} est alors explicitement connectée à une structure de type amplificateur et comparée à V_m. Les autres méthodes se réfèrent aux figures II.6 et II.8 dans lesquels la tension de seuil n'a pas de relation directe avec la tension de membrane.

Le retour au potentiel de repos de la capacité se fait également de deux façons. Il peut être soit interne au neurone, auquel cas un mécanisme de bouclage active généralement un transistor responsable de la remise à zéro, soit externe. Un protocole, pouvant être par exemple de type *poignée de main*, est alors utilisé pour réinitialiser le potentiel de membrane du neurone. En plus de la fuite intrinsèque de la capacité, certains neurones proposent une fuite additionnelle, réglable à l'aide d'une tension ou d'un courant.

B. Conception d'un neurone LIF robuste

Nous présentons ici la conception du neurone à partir de ses spécifications. On a privilégié une approche fonctionnelle bloc par bloc. Chacun fera l'objet d'un paragraphe dont l'objectif est de justifier l'emploi d'une certaine topologie. Ces briques de base permettront ensuite l'élaboration du neurone dont nous décrirons le fonctionnement globale ainsi que les simulations auxquelles il a été soumis lors de sa réalisation.

1. Spécification des caractéristiques du neurone

De par le fonctionnement hybride des neurones, on peut espérer des gains en surface et en consommation lors des passages vers les nœuds futurs. En effet, comme la partie

routage des impulsions peut être réalisée par le biais d'impulsions numériques, elle profitera pleinement du passage à l'échelle.

Si le routage a tout intérêt à se faire de manière numérique (signaux binaire par nature), les neurones et les synapses peuvent être implémentés des deux façons, analogique ou numérique. Leur nombre doit être le plus élevé possible tout en maintenant une puissance dissipée maitrisée. C'est pour cela qu'il a été fondamental de bien spécifier le cahier des charges du neurone à fabriquer.

a) Modèle de neurone à implémenter

Nous avons choisi d'utiliser le modèle de neurone Leaky Integrate-and-Fire dont le comportement est décrit au paragraphe I-A.2.d). D'un point de vue applicatif, le modèle LIF rend possible un grand nombre d'opérations (28), à savoir l'addition ou la soustraction (par le biais de poids positifs ou négatifs), ou encore la multiplication grâce à la fuite (79). Il a donc été choisi pour ses capacités calculatoires et sa simplicité permettant une implémentation efficace. En vue d'être intégrés à grande échelle, les neurones LIF devront avoir une petite taille, une faible consommation et enfin être robustes à la variabilité du procédé de fabrication.

D'autres contraintes sur la conception proviennent de l'architecture et des applications envisagées. Il a ainsi été déterminé que le poids synaptique des neurones doit être encodé sur 7 bits. Cela implique également que le neurone doit avoir un SNR de 30 dB. Certaines applications nécessitent une fréquence de fonctionnement du neurone élevée afin de diminuer le temps de calcul et fixent par conséquent une fréquence maximale de 1 MHz. La fréquence minimale correspondra à des limites matérielles détaillées dans la partie II-B.2.a).

b) La structure du neurone analogique

De par sa nature, il apparaît naturel de concevoir un neurone analogique. En effet, il exploite pleinement les équations physiques de l'électronique à savoir l'intégration du courant aux bornes d'une capacité et la sommation des courants.

Comme illustrée sur la figure II.9 Le bruit généré par un neurone n'est pas propagé d'étage en étage mais supprimé lors de l'émission de l'impulsion logique. Ce fonctionnement hybride analogique/numérique permet de régénérer le signal et allège les contraintes de conception du neurone en réduisant son rapport signal à bruit. D'après

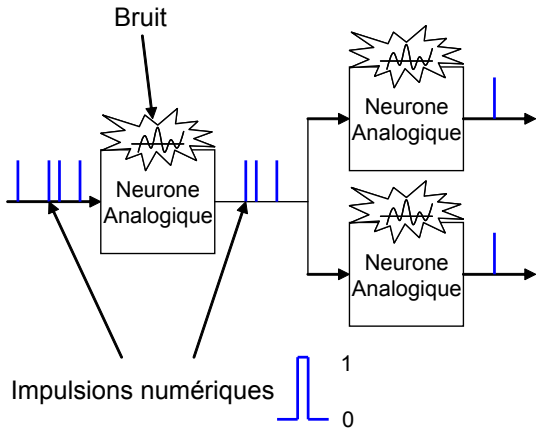


Figure II.9: Fonctionnement mixte des neurones. Le bruit introduit dans les neurones analogiques est supprimé lors de la génération d'un potentiel d'action sous la forme d'un signal numérique.

le paragraphe II-A.2., toutes les conditions sont réunies pour qu'une implémentation analogique d'un neurone soit la plus adaptée. Sa faible résolution permettra d'espérer des gains en termes de surface et de consommation.

L'observation d'un neurone biologique et du modèle LIF permet d'identifier les blocs à concevoir pour une implémentation électronique. Ceux-ci sont montrés sur la figure II.10. Par analogie avec la biologie et de gauche à droite, on a représenté les synapses, l'accumulation d'ions au niveau du soma et enfin la propagation du potentiel d'action. Les blocs électroniques à concevoir sont donc les suivants : un convertisseur numérique/analogique, une capacité, un comparateur, un point mémoire, une fuite contrôlée et une remise à zéro.

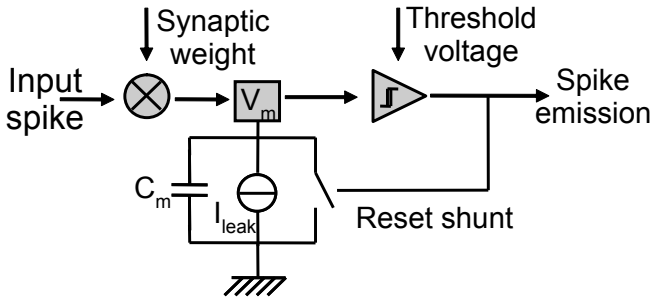


Figure II.10: Schéma bloc d'un neurone analogique LIF, un mécanisme de modulation du poids (synapse biologique) et un mécanisme de propagation du potentiel d'action.

Après avoir détaillé le cahier des charges du neurone, il faut à présent définir une

méthode de travail. Les différents blocs de la figure II.10 sont également présents sur le schéma II.11 que nous avons suivi lors de la conception. Les flèches de gauche constituent les données principales à prendre en compte lors de la réalisation du bloc. Sa création peut générer des phénomènes indésirables, visibles sur les flèches de droite, entrainant la modification d'une étape précédemment réalisée.

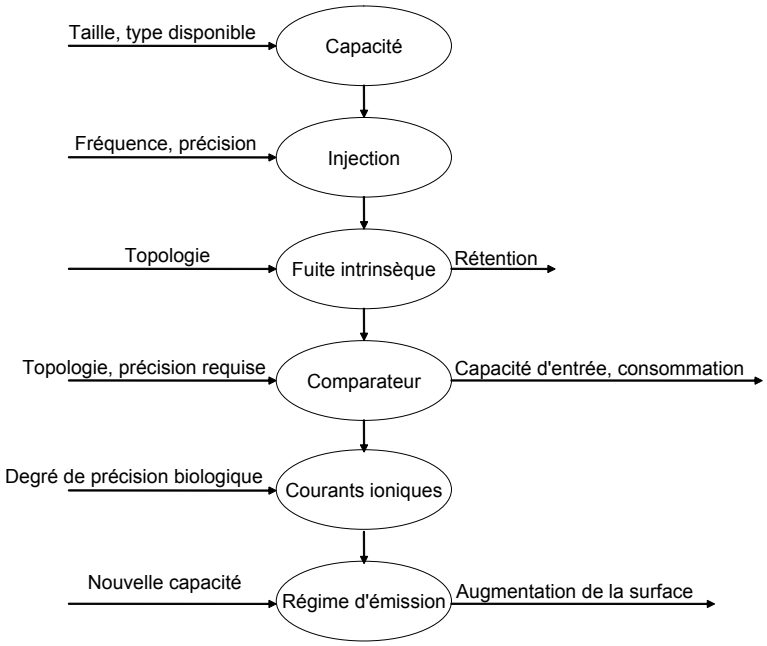


Figure II.11: Méthodologie adoptée pour la conception du neurone. Dans le contexte du projet Arch^{2}Neu, l'avant dernière étape se résume à une remise au potentiel de repos et à une fuite simple (constante et programmable). Quant à la dernière étape, elle ne nous est pas nécessaire pour notre implémentation.

Une première étude de la taille de capacité permettra, selon la topologie d'injection choisie, de réajuster sa valeur pour optimiser sa rétention. On réalisera un bloc de comparaison et un autre de fuite, représentés par les courants ioniques sur le schéma. Le régime d'émission concerne principalement les implémentations de neurones biologiques qui peuvent générer des motifs de potentiels d'action. Ce dernier bloc, nécessitant une capacité d'adaptation, n'est pas utilisé dans notre implémentation.

2. Choix et réalisation des structures élémentaires

Il est question ici de détailler les différentes topologies et dimensionnements des blocs qui constituent le neurone. Suivant le flot détaillé précédemment, on commence

naturellement par choisir la capacité.

a) Choix de la capacité de membrane

Une liste des capacités disponibles dans le design kit ainsi que leurs caractéristiques principales sont listées dans le tableau II.3. Il est possible de déterminer le type approprié de capacité en regardant ses fuites intrinsèques ainsi que la fréquence de fonctionnement désirée du neurone.

Type de capacité	I_{fuite}	$C_{surfacique}$	Observations
Poly N+ Pwell	1 pA/um^2	13.2 $fF/\mu m^2$	@ $0.9V$
Poly P+ Nwell	$0,2$ pA/um^2	12.3 $fF/\mu m^2$	@ $0.9V$
M1/M2	-	330 $aF/\mu m^2$	-
Striped Stack metal M1/M5	-	830 $aF/\mu m^2$	Surface > 120 μm^2
MIM	20 aA/um^2	5 $fF/\mu m^2$	$1,4$ nA/cm^2@ $1V$

Table II.3: Liste des capacités disponibles et quelques unes de leurs caractéristiques à 27°C. I_{fuite} n'est pas indiqué pour 2 capacités du fait de leur faible capacité surfacique.

Pour réaliser un neurone fonctionnant à une fréquence basse f_{low} égale à 1 kHz, on doit déterminer préalablement la fuite intrinsèque maximale de la capacité notée $I_{fuite_{max}}$. On se base sur l'équation suivante :

$$I_{fuite} = C_m \frac{dV}{dt} \tag{II.1}$$

On peut ainsi en déduire $I_{fuite_{max}}$ en considérant $f_{low} = 1$ kHz, l'amplitude de la variation de la tension de membrane V_M ($V_{threshold} - V_{rest} < V_{dd}$), et de sa résolution en bit ($n = 7$). On a alors $\delta V = \frac{V_{threshold} - V_{reset}}{2^n}$ qui représente l'incrémentation d'un poids minimal en mV. Si l'on veut conserver δV pendant 1 ms, on obtient alors :

$$\begin{aligned} I_{fuite_{max}} &= C_m \frac{V_{threshold} - V_{rest}}{2^n} \; f_{low} && \text{(II.2)} \\ &= C_m * \delta V * f_{low} && \text{(II.3)} \\ &= C_m * \frac{0.6}{128} * \frac{1}{1.10^{-3}} && \text{(II.4)} \\ &\simeq C_m * 4.10^{-3} && \text{(II.5)} \end{aligned}$$

On peut alors calculer le courant maximal de fuite autorisé en considérant la capacité et le courant de fuite surfaciques. Par élimination, on s'aperçoit que seule la capacité MIM respecte cette contrainte puisque :

$$I_{fuite_{maxMIM}} = 5.10^{-15} * 4.10^{-3} = 2.10^{-14}\ A/\mu m^2 \tag{II.6}$$

Et l'on vérifie bien $I_{fuite_{maxMIM}} \gg I_{fuite_{MIM}}$. On ne peut cependant pas déterminer à ce stade la taille de la capacité. Pour que le neurone puisse fonctionner à 1 kHz, il faudrait connaitre dès à présent les autres courants de fuites liés aux blocs adjacents à la capacité (69). Nous allons donc maintenant passer à l'étude des blocs connectés à la capacité.

b) Topologie d'injection

Les schémas décrits dans la partie précédente montrent différentes approches permettant d'injecter un poids synaptique dans un neurone. Plus généralement, la modulation d'un poids synaptique peut se faire de deux façons. La première est la modification du flux de charges injectés. Ceci peut se faire grâce à l'aide de capacités comme sur la figure II.6, directement via des résistances/transistors (fig. II.8), ou plus généralement avec un CNA. La seconde est de modifier le temps pendant lequel un flux constant est injecté.

L'architecture proposée par Vogelstein (86) n'est pas une solution envisageable. La RAM externe contient les poids synaptiques et un CNA est commun à tout les neurones. Le CNA deviendrait le point sensible lors de la fabrication et on perdrait ainsi l'objectif d'architecture robuste massivement parallèle.

Une implémentation de *mini-DACs* (48) permettrait de répondre à l'objectif de robustesse en distribuant la conversion numérique/analogique. La précision obtenue ne serait pas suffisante si l'on restreint la consommation en vue d'une implémentation à grande échelle.

La seconde solution, c'est à dire la modulation temporelle d'un flux constant, dépend de la variabilité des temps de montée et descente des portes de contrôle numériques illustrés sur la figure II.12. Cette variabilité a été estimée sur un échantillon de 500 MUX 2 :1. La sélection du multiplexeur peut permettre au choix l'injection de courant par le biais d'un interrupteur ou son blocage par une valeur par défaut. Lorsque la durée moyenne d'une impulsion est de 938 ps, l'écart type est de $2,7\ ps$. Cette robustesse est

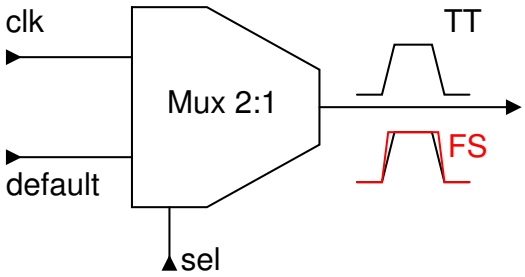


Figure II.12: Effets de la variabilité sur un MUX 2 :1. La largeur de l'impulsion délivrée varie selon la fabrication. Lorsque le signal d'horloge et sélectionné par le multiplexeur, le créneau de sortie peut avoir deux formes suivant les caractéristiques des transistors (ici TT et FS).

équivalente à un SNR de 51 dB qui est bien supérieur aux spécifications visées pour le neurone (30 dB). La précision nécessaire de 7 bits est effectivement réalisable ; la modulation synaptique générique de la figure II.10 se précise et devient une modulation par impulsions comme le montre la figure II.13.

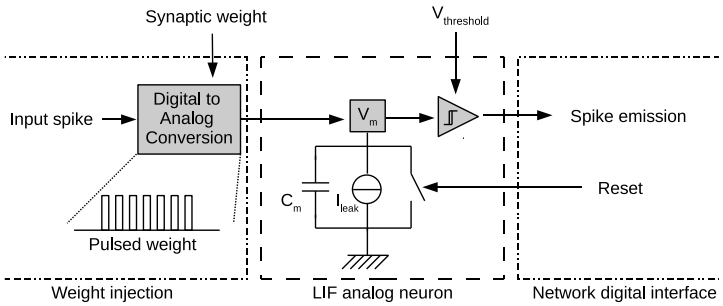
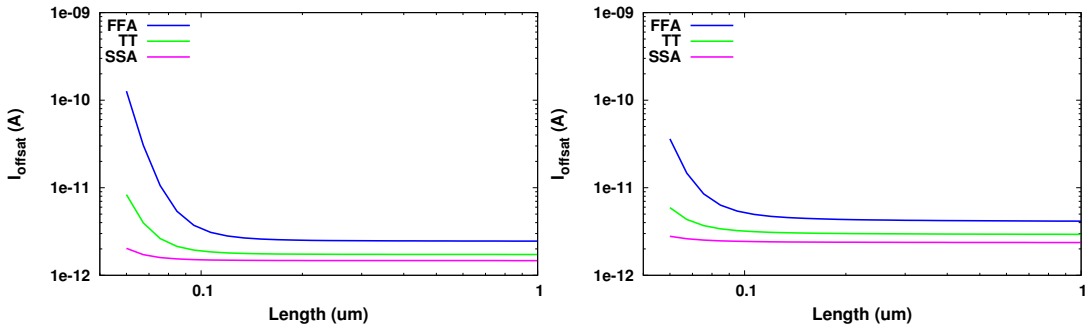


Figure II.13: Schéma bloc du neurone analogique LIF, le mécanisme de modulation du poids est réalisé par une conversion en nombre d'impulsions : 128 impulsions correspondent à un poids dont la valeur absolue est égale à 1.

Par conséquent, on se propose de réaliser la modulation du poids synaptique à l'aide d'un interrupteur S_{pulse} commandé par l'horloge de la partie numérique ayant une fréquence de 500 MHz. Le mécanisme de modulation du poids est réalisé par une conversion en nombre d'impulsions : 128 impulsions correspondent à un poids dont la valeur absolue est égale à 1, valeur normalisée selon $V_{threshold}$. Pour le réaliser, une étude du courant de fuite en saturation I_{offsat} ($V_{ds} = 1,2$ et $V_{gs} = 0$) est présentée sur la figure II.14. Ceci permet d'estimer la dimension nécessaire aux transistors pour

permettre une isolation suffisante de la capacité C_m.



(a) Nmos transistors **(b)** Pmos transistors

Figure II.14: I_{offsat} de transistors MOS de type HVTLP en fonction de leur longueur de grille pour différents corners : TT, FFA, SSA. $V_{gs} = 0\ V$, $V_{ds} = 1,2V$, $W = 1\ \mu m$ et $T = 27°C$

En dessous de $0,2\ \mu m$, la variabilité locale est prépondérante comme en témoigne l'écartement entre les courbes des différents corners FFA et SSA. Lorsque la longueur de grille est supérieure à $0,2\ \mu m$, l'influence des variations globales devient majoritaire dans la variation du courant de fuite. Ceci se traduit sur la figure par l'écart constant entre les corners. Il n'est donc pas utile de prendre une longueur du canal plus élevée. C'est donc celle-ci que l'on a utilisé pour l'implémentation des transistors de l'interrupteur S_{Pulse}.

Le signal d'horloge passera au travers d'un MUX dont il faudra configurer l'état par le bit de sélection ainsi que la valeur par défaut. Ceci indique le besoin d'un bloc de contrôle numérique qui se sera nécessaire au fonctionnement du neurone.

c) Comparateur

Les différentes topologies utilisées dans les neurones sont généralement des OTAs (85) ou des inverseurs (86). Cependant, un meilleur appariement n'est possible que lorsque deux transistors de même type sont utilisés dans l'étage d'entrée. Dans une approche de robustesse à la variabilité, nous avons par conséquent privilégié les topologies basées sur une paire différentielle comme les OTAs.

En considérant les contraintes de consommation, un problème apparaît lorsque la valeur à comparer V_m se trouve juste en dessous de la valeur seuil $V_{threshold}$. La sortie du comparateur se trouve autour d'une valeur proche de $V_{dd}/2$. Ceci provoque la création d'un chemin de conduction, illustré sur la figure II.15, entre V_{dd} et la masse à l'interface des parties analogique et numérique du circuit.

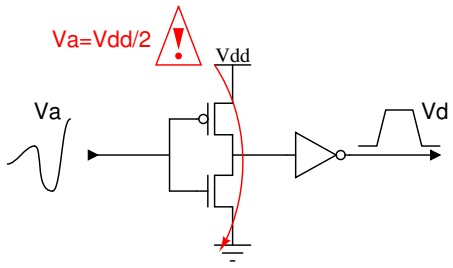


Figure II.15: Interfaçage entre analogique et numérique. Un chemin de conduction apparait si la tension analogique reste à un potentiel permettant la conduction simultanée des transistors de la porte numérique, ici un inverseur.

Afin de diminuer la consommation, certaines implémentations utilisent une boucle rétroactive connectée à la sortie de l'amplificateur permettant d'accélérer le basculement du neurone (voir sur les fig. II.5, II.7 et II.8).

Le neurone de la figure II.6 a une structure un peu moins traditionnelle. Sa consommation est maitrisé par les potentiels V_{bp}, V_{bn} et enfin V_{thresh}. Elle utilise un transistor M5 en *pull up* couplé à M4. M6 devient passant lorsque V_m augmente, et V_{comp} prend la valeur de V_{thresh}. Si la vitesse de charge de la capacité C_m est trop lente, le transistor M12 de *pull down* permettra également d'optimiser la consommation. Cette topologie ne possède toutefois pas la paire différentielle requise du point de vue de la robustesse.

Nous avons opté pour l'utilisation d'un comparateur à bascule (25). Son schéma est présenté sur la figure II.16. Celui-ci a l'avantage d'être alimenté uniquement lors de son activation. La topologie induira la rapidité de décision et la faible consommation. Les transistors d'entrée de la paire différentielle sont de type PMOS pour leur meilleur appariement et leur plus faible courant de grille. Ils permettront donc d'accroitre la précision et d'augmenter la rétention des charges de la capacité.

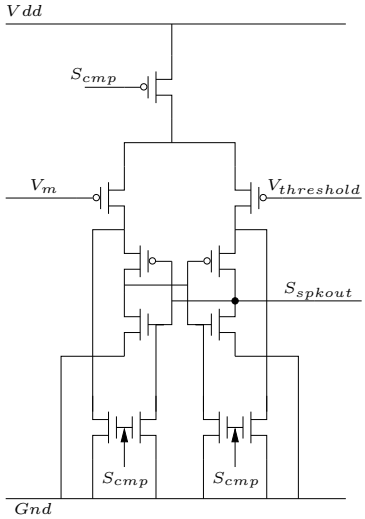


Figure II.16: Comparateur à bascule finalement retenu dans notre implémentation.

d) Dimensionnement de la capacité

A ce stade, on a déterminé les deux topologies qui vont isoler la capacité du reste du neurone. On peut identifier, sur la figure II.17, les courants de fuite dans l'interrupteur utilisé pour la modulation des poids synaptiques d'une part et à travers la grille du comparateur d'autre part. On peut ainsi estimer la valeur de la capacité requise pour une rétention d'un poids minimal δV pendant 1 ms.

$$\Sigma I_{fuite} = I_{fuite_{MIM}} + I_{fuiteS_{pulse}} + I_{fuite_{PMOS}} \quad \text{(II.7)}$$

$$\simeq I_{fuiteS_{pulse}} + I_{fuite_{PMOS}} \quad si \quad C_m < 1000\ um^2 \quad \text{(II.8)}$$

$$= 1.5\ pA \quad \text{(II.9)}$$

$$C_m = \frac{\Sigma I_{fuite}}{\delta V f_{low}} \simeq 350\ fF \quad \text{(II.10)}$$

La valeur de 350 fF est obtenue en considérant uniquement les courants de fuite. Par conséquent, elle est majorée pour pallier une fabrication dont la capacité surfacique serait plus faible. C_m a donc été fixée à 500 fF et correspond à une surface de 100 um^2. La capacité étant intégrée au niveau du *back-end*, la surface sous-jacente est employée au placement des transistors du neurone, le coût en surface est virtuellement nul.

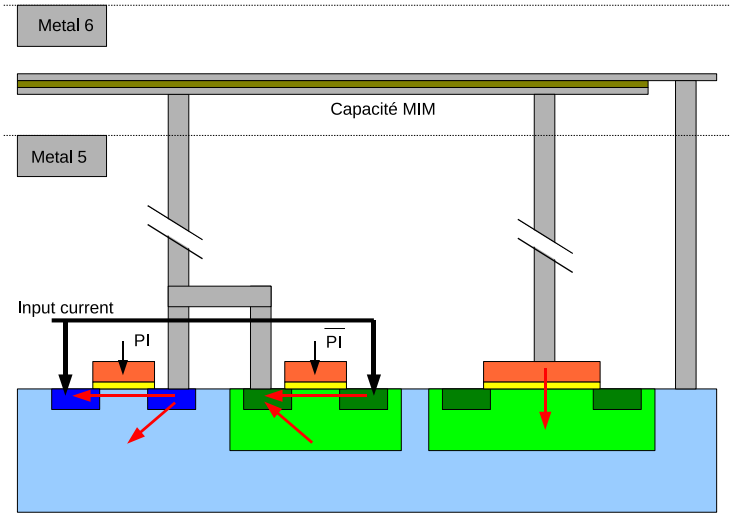


Figure II.17: Schéma des différents courants de fuites aux bornes de la capacité : à gauche l'interrupteur commandé par le signal PI (pulsed injections), à droite le PMOS à l'entrée du comparateur. On peut y identifier les 3 types de fuite : a) courant de la jonction de diode polarisé en inverse, b) courant de diffusion dans le canal, c) courant de grille.

e) Bloc de fuite programmable et remise à zéro

Nous présentons ces deux blocs car ils sont liés tant sur le schéma que lors de leurs utilisations. Le fonctionnement du neurone en mode "*fuite*" permet d'étendre ses capacités calculatoires. L'ajout d'une fuite permet notamment de détecter la coïncidence entre deux trains impulsions. Elle doit pouvoir prendre différentes valeurs en fonction de l'emplacement du neurone dans le flot de calcul.

La stratégie adoptée est l'implémentation d'une fuite programmable à l'aide de trois bits notés $\tau\langle 0:2\rangle$. L'activation de plusieurs bits est possible et permet une meilleure

	τ_0	τ_1	τ_2
Fuite (μs)	100	20	10
Fréquence (kHz)	10	50	100
W Nmos (μm)	1	5	10

Table II.4: Constantes de temps de fuite implémentées et dimensions des transistors correspondants.

granularité de la fuite. La tension $V_{leakbias}$ ajoute un degré de liberté pour le choix de la valeur de la fuite. Pour les valeurs indiquées dans le tableau II.4, $V_{leakbias} = 800\ mV$.

La remise à V_{reset} du potentiel V_m du neurone se fera à l'aide d'un interrupteur placé en parallèle aux trois transistors de fuite. Elle nécessitera la génération d'une impulsion générée par le neurone, puis sa réception et sa validation par un bloc numérique, détaillé dans la partie II-B.3.b). Pour la fuite comme pour le retour au potentiel de repos, le comportement des interrupteurs S_{leak} et S_{pulse} devra également être configuré.

3. Fonctionnement du neurone

On rappelle dans les tableaux II.5 puis II.6 les signaux numériques d'entrées/sortie et les valeurs analogiques nominales de polarisation nécessaires aux neurones.

Signal	Fonction
S_{pulse}	Modulation du poids synaptique
S_{cmp}	Comparaison
$S_{\Phi_{dec}}$	Activation du neurone, poids négatif
$S_{\Phi_{inc}}$	Activation du neurone, poids positif
$S_{Tau}\langle 0:2\rangle$	Paramètre de fuite
S_{leak}	Fuite, stabilisation et remise à zero
S_{reset}	Remise à V_{reset}
S_{spkout}	Potentiel d'action généré

Table II.5: Signaux numériques du neurone : 9 entrées, 1 sortie

Notation	Fonction	Valeur typique
$V_{threshold}$	Potentiel de seuil	900 mV
$V_{leakbias}$	Potentiel de fuite	800 mV
V_{reset}	Potentiel de repos	300 mV
I^+	Courant pour incrémentation	1.2 uA
I^-	Courant pour decrémentation	1.2 uA

Table II.6: Valeurs analogiques nominales de polarisation

a) Description

Nous allons maintenant détailler le fonctionnement du neurone à l'arrivée d'une impulsion. Son schéma, présenté sur la figure II.18, est constitué des blocs précédemment conçus.

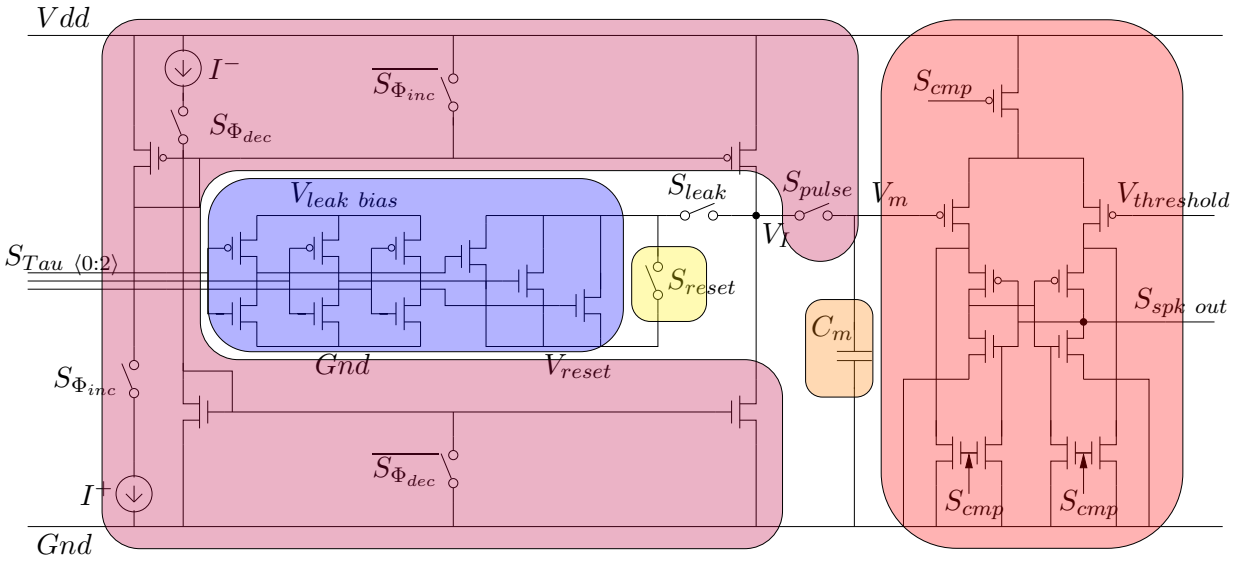


Figure II.18: Schéma du neurone analogique réalisé. On identifie l'injection (violet), la fuite (bleu), la remise à zéro (jaune), la capacité (orange), le comparateur (rouge)

Pour l'exemple qui suit, le neurone est en mode IF (sans fuite) : les signaux $S_{Tau}\langle 0:2\rangle$ sont à 1. Lors de la stimulation d'un neurone, le miroir correspondant à la polarité du poids est activé : les interrupteurs $S_{\Phi_{inc}}$ et son complément $\overline{S_{\Phi_{inc}}}$ (respectivement $S_{\Phi_{dec}}$ et $\overline{S_{\Phi_{dec}}}$) sont utilisés pour un poids positif (respectivement négatif). Quand le miroir est stable, l'interrupteur S_{leak} est ouvert et l'injection de charges dans la capacité via S_{pulse} peut débuter. Le nombre d'impulsions de cet interrupteur est fonction du poids synaptique. Une fois l'injection terminée, le potentiel V_{reset} est appliqué au nœud V_I afin de diminuer le potentiel V_{MI} aux bornes de S_{pulse}. Pendant ce temps, le comparateur est activé par S_{cmp}, un événement logique est généré si $V_M > V_{threshold}$. Il est envoyé vers d'autres neurones via un mécanisme de routage qui remet le potentiel V_M à la valeur V_{reset} en fermant S_{reset} et S_{leak}. La comparaison est activée lors de l'injection de poids positifs puisque le dépassement de $V_{threshold}$ ne peut avoir lieu lorsque le potentiel V_M est diminué.

b) Conversion numérique analogique et commande numérique

Si la simulation du neurone peut être réalisée à l'aide de sources idéales de tension, l'intégration silicium requiert un bloc numérique permettant de contrôler le neurone. Il a été succinctement évoqué sur la figure II.13 sous le terme de *Digital to Analog Conversion* (DAC) qui se distingue d'un CNA par sa spécificité envers le neurone. En effet, ce bloc DAC génère, à l'aide d'une machine d'état (FSM), les signaux de contrôles numériques du neurone précédemment décrits. Ils permettent ainsi la configuration et le fonctionnement du neurone, comme détaillé sur le schéma de la figure II.19.

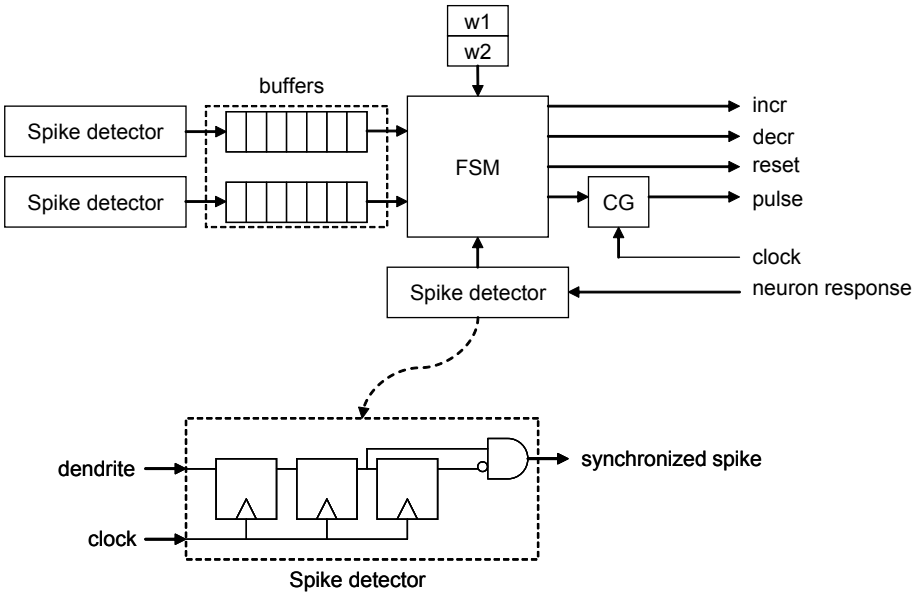


Figure II.19: Schéma du DAC, bloc de contrôle du neurone.

Lorsqu'une impulsion stimule un neurone, elle est stockée dans une FIFO (*buffers*) puis transmise à la FSM. La FSM permet la génération des signaux selon la valeur du poids w_i et le mode d'utilisation du neurone, configuré avec ou sans fuite. Un bloc de *Clock Gating* permettra la génération des impulsions nécessaires à la modulation du poids. Un potentiel d'action, émis par un neurone, est synchronisé à l'aide de trois bacules avant d'entrer dans la FSM. Elle générera ensuite le signal de remise à zéro du neurone. Le DAC occupe une surface de 1300 μm^2 et consomme une énergie de l'ordre 0.3 mW. A noter qu'il peut être mutualisé entre différents neurones.

Les caractéristiques temporelles des signaux de contrôles sont déterminées lors de la conception du neurone et sont visibles dans le tableau II.7 ainsi que sur la figure

II.20. Leurs spécifications sont nécessaires lors du codage de la FSM. Le signal S_{pulse} est activé à $500MHz$ lors de l'injection. Le temps de remise à zéro, τ_{reset} est fonction de la résistance en série des interrupteurs S_{pulse}, S_{leak} et S_{reset}. La durée d'activation τ_{cmp} du signal S_{cmp} est fonction du temps de basculement du comparateur. Le temps de stabilisation τ_{stab} pour les miroirs activés par les signaux S_{Φ} est fonction de leur taille. Il est visible sur cette même figure, en bas, par le passage de S_{leak} et S_{reset} à 0, 22 ns après le passage à 1 de $S_{\Phi_{inc}}$.

Description	Signal concerné	Durée (ns)
$T_{circuit}$	S_{pulse}	1
τ_{cmp}	S_{cmp}	14
τ_{stab}	S_{Φ}	22
τ_{reset}	S_{reset}	70

Table II.7: Tableau des principaux temps nécessaires à la réalisation du DAC

c) Simulation mixte

L'intérêt de la simulation mixte est de fournir à la fois polyvalence et souplesse lors des tests. Elle permet également de valider le comportement de la partie numérique lorsqu'elle est connectée à la partie analogique. La topologie de la simulation mixte est montrée sur la figure II.21.

Le bloc "*Testbench*" est décrit à un niveau comportemental et procure toute la souplesse nécessaire aux simulations. On pourra aisément modifier les valeurs et les signes des poids synaptiques à utiliser, l'instant et l'envoi des impulsions, etc. Dans un premier temps, le DAC est décrit à un niveau comportemental. On vérifie à ce stade que le comportement du neurone est en adéquation avec les spécifications. L'outil de simulation place des convertisseurs idéaux qui permettent la correspondance entre les niveaux logiques et les tensions Vdd ou Gnd. Le DAC est ensuite remplacé par une vue décrite à l'aide des caractéristiques des portes logiques fournies dans le design kit. Pour les signaux critiques, on a simulé également les sorties à un niveau transistor. Les éventuels problèmes temporels sont détectés lors de cette dernière simulation.

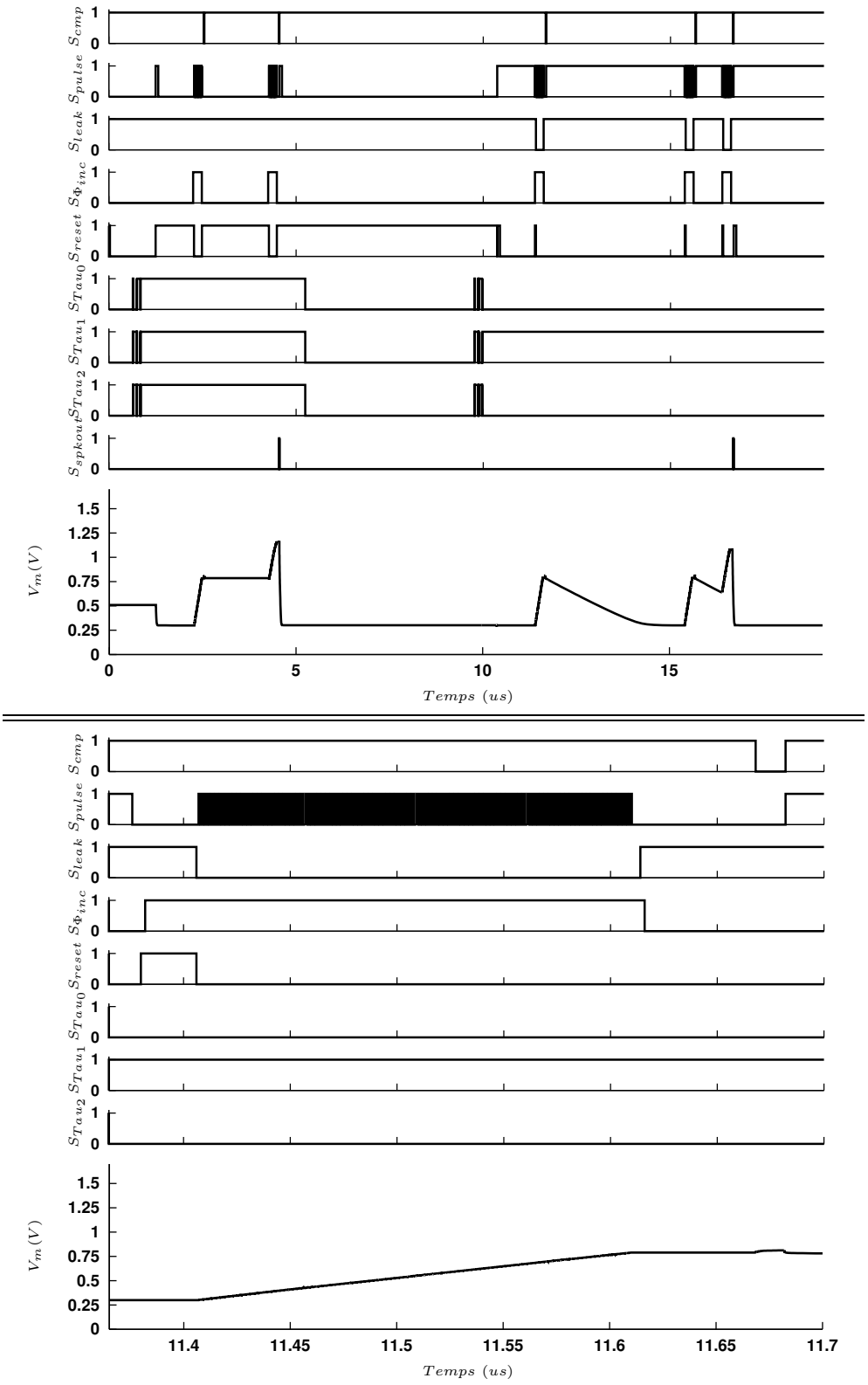


Figure II.20: Chronogramme des signaux de contrôle du DAC lors de l'évolution du potentiel de membrane V_m. En haut : Le neurone est configuré en mode IF puis après reconfiguration en mode LIF ; les poids sont égaux à +0.8. En bas : zoom lors de l'injection d'une impulsion en mode LIF.

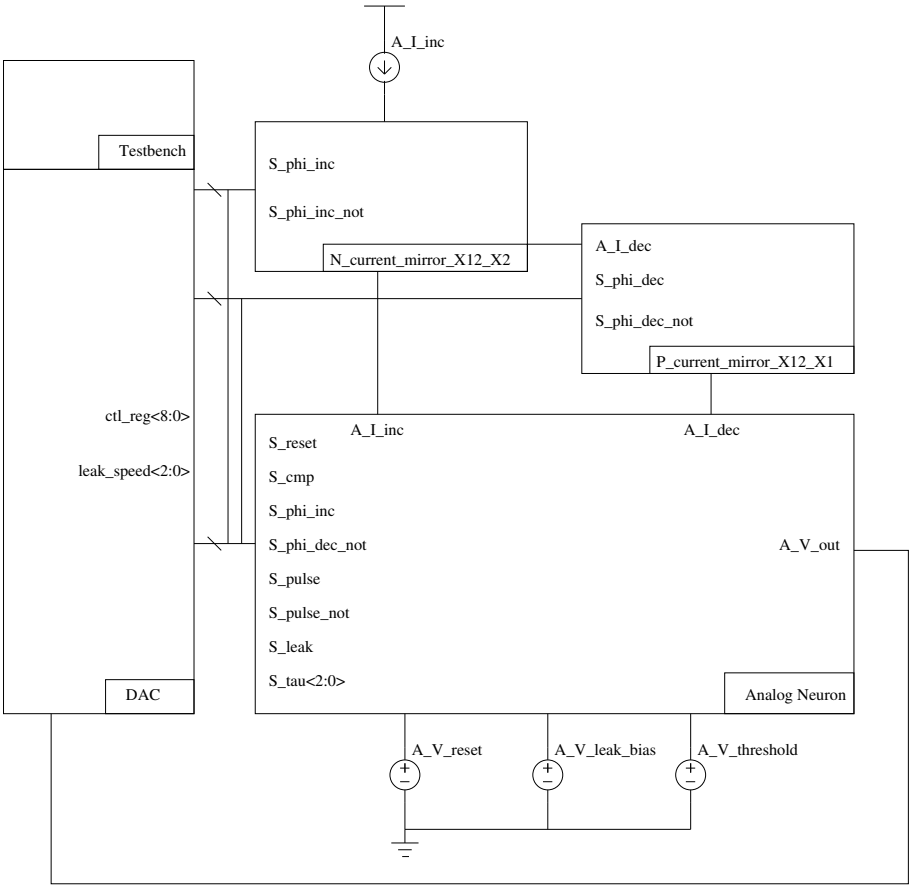


Figure II.21: Banc de test mixte pour neurone analogique.

4. Résultats de simulation

On vérifie tout d'abord le bon fonctionnement du neurone. Nous avons procédé à plusieurs analyses de variations globales et locales afin d'étudier la robustesse du neurone face à la variabilité du procédé de fabrication. En effet, tous les neurones d'une même puce subiront les mêmes variations globales (variations de lot à lot, de plaque à plaque et de puce à puce). Des variations locales vont également influencer l'appariement entre transistors. Les simulations de type Monte-Carlo simulent la variabilité en générant un tirage aléatoire des paramètres d'un composant. On présente ici l'impact de ces deux types de variations.

a) Simulation de fonctionnement

Les premières études du comportement du neurone ont été réalisées uniquement avec Eldo, les signaux de contrôle du neurone étaient alors simulés à l'aide de source de tension idéale. Les résultats de cette approche sont présentés sur la figure II.22. Dans cette simulation, le neurone reçoit successivement les poids +0.8, -0.5, +0.8. Ceci correspond soit à la génération de 104 créneaux du signal S_{pulse} et l'activation du signal $S_{\phi_{inc}}$, soit de 64 créneaux avec $S_{\phi_{dec}}$. Autour de 400 ns, on observe une augmentation de V_m provenant de sa comparaison avec V_{th}. Celle-ci est causée par la capacité parasite entre la source et la grille du transistor d'entrée du comparateur lors de son activation. A la fin de la comparaison, V_m retourne a un potentiel avec un écart plus faible que la valeur d'un incrément et n'aura, par conséquent, qu'un très faible impact sur le résultat du calcul. La valeur de ces incréments est visible sur l'encart de la figure lors de la deuxième injection de poids positifs et est de l'ordre de 4 mV.

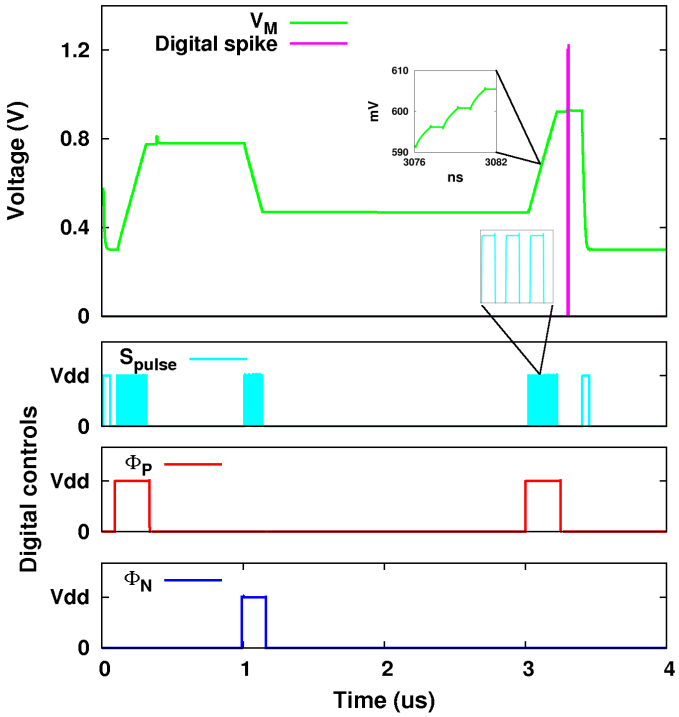


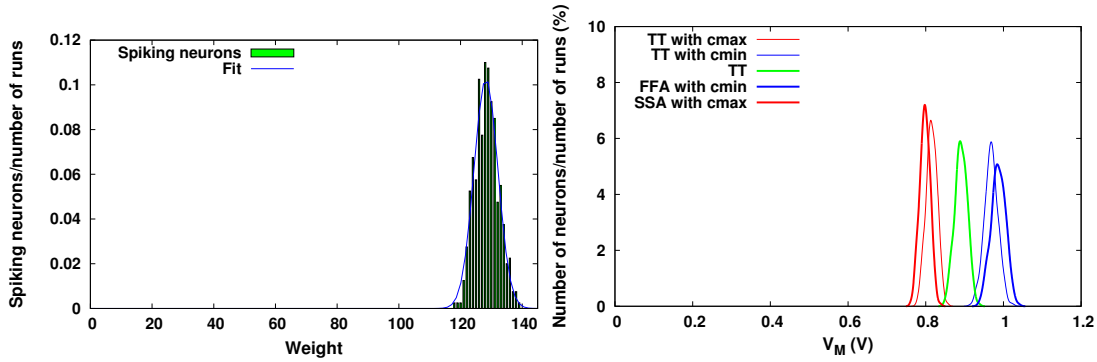
Figure II.22: Évolution du potentiel de membrane V_m et sortie numérique du neurone. Le neurone est configuré en mode IF, ses principaux signaux de contrôle numérique permettent l'injection de 3 poids consécutifs : respectivement +0.8, −0.5, et +0.8.

La simulation par des sources de tensions analogiques bénéficie d'une facilité de mise en place initiale puisqu'aucun environnement spécifique n'est requis. Dans un deuxième temps, et pour pouvoir tester davantage de configurations, nous avons effectué des simulations mixtes numériques/analogiques. On a ainsi pu valider le bon comportement du neurone avec son DAC comme en témoigne la figure II.20. Ceci a été réalisé dans de nombreuses configurations : avec/sans fuite, poids positifs/négatifs, séquences variées de stimulations. De 0 à 10 us sur cette figure, le neurone est d'abord configuré en mode sans fuite. L'arrivée de deux impulsions successives permet au neurone d'émettre une impulsion montrée sur la ligne S_{spkout}. Ce signal est reçu par le DAC qui permet le retour de V_m au potentiel V_{reset} la remise à zéro du neurone et envoie en parallèle l'impulsion dans le réseau. Une reconfiguration du DAC a lieu à 10 us par le biais du bloc "*Testbench*" qui lui est attaché (cf. structure de test sur la fig.II.21). Il contrôle alors le neurone de sorte qu'il fonctionne à présent avec une fuite, comme en témoigne les signaux modifiés S_{pulse}, S_{pulse}, S_{leak}, S_{reset} et S_{Tau_1}. Il est maintenant nécessaire que deux impulsions soient suffisamment proches dans le temps pour que le neurone déclenche un potentiel d'action, visible vers 17 us.

b) Variations locales dans le corner nominal (TT)

L'histogramme de la figure II.23a montre l'influence des variations locales entre transistors lorsque les variations globales sont nominales. A chaque simulation, un poids est injecté dans un échantillon de 400 neurones. Le nombre de neurones dont la sortie a été active est reporté dans l'histogramme. Sans variabilité, tous les neurones devraient émettre un potentiel d'action lorsque le poids vaut 1. Ce poids étant codé sur 7 bits, il requiert 128 impulsions de S_{pulse} pour permettre un déclenchement du neurone. En pratique on observe une distribution, centrée autour de 128, du nombre d'impulsions nécessaires à la génération d'un potentiel d'action.

La distribution peut être modélisée par une loi normale dont l'espérance $\mu = 128.2$ et l'écart type $\sigma = 3.9$. Ceci correspond à un coefficient de variation de 3% équivalent à un rapport signal à bruit de 30.3 dB. Les spécifications fixées au début de ce chapitre sont effectivement respectées.



(a) Impact au sein du corner TT **(b)** Impact des variations globales

Figure II.23: Impact du procédé de fabrication sur le comportement des neurones. Figure (a) : Histogramme vert : nombre de neurones ayant généré un potentiel d'action à un poids donné. Courbe bleue : fit par une loi normale ($\mu = 128.2$ et $\sigma = 3.9$). Figure (b) : Distribution du potentiel de membrane avec différents corners. On note l'effet prépondérant des variations de la capacité de membrane C_m.

c) Comparaison entre différents corners : SSA-TT-FFA

Dans la figure II.23b, la forme de chaque courbe représente les variations du potentiel de membrane V_m lorsqu'un neurone reçoit une impulsion pondérée d'un poids égal à 1. Chaque courbe représente un type de variation globale de fabrication : cas nominal (TT), transistors lents avec grande capacité (SSA-Cmax), transistors rapides avec petite capacité (FFA-Cmin). En observant les courbes TT-Cmax et TT-Cmin, on remarque l'importance prépondérante de la variation globale de la capacité sur le potentiel V_m.

Les effets des variations globales peuvent être caractérisés post-fabrication et compensés par l'adaptation de la tension de seuil $V_{threshold}$, l'amplitude du courant injecté, la fréquence de fonctionnement de S_{pulse}, ou bien encore par le nombre d'impulsions correspondant à un poids égal à 1.

d) Estimations de consommation

Dans le cadre d'une étude *bottom-up*, il est important de prévoir la consommation du neurone exprimée en watt. Elle permettra à plus haut niveau d'estimer l'impact de la consommation des neurones analogiques au niveau du circuit.

Pour émuler l'intégration du neurone dans un environnement numérique, la sortie

du comparateur du neurone est connectée à un inverseur. La charge de sortie C_L a une valeur de l'ordre de 1 fF. L'étude de la consommation inclut les courants pour l'injection des charges dans la capacité, la comparaison, ceux de fuite et enfin ceux utilisés pour l'activation des interrupteurs de contrôle. On considère donc $S_{\Phi_{inc}}$ et $S_{\Phi_{dec}}$ pour l'activation du neurone selon le signe du poids synaptique mais également et surtout S_{pulse} pour la modulation du poids.

Comme on s'intéresse à la consommation intrinsèque au neurone, on ne considère que les courants qui leur sont nécessaires. On définit 2 régimes d'émission des potentiels d'action d'un neurone :

- t_{active} : période active pendant laquelle une impulsion est intégrée dans la capacité, V_m est comparé avec $V_{threshold}$ pour générer l'impulsion numérique et enfin le neurone est remis à son potentiel de repos V_{reset}. L'énergie nécessaire à la génération d'un potentiel d'action E_{spike} est obtenue en intégrant la puissance instantanée du neurone pendant t_{active}. Cette puissance est quant à elle obtenue par multiplication du courant avec la tension d'alimentation.

- $t_{standby}$: période pendant laquelle le neurone attend l'arrivée du prochain potentiel d'action. Il possède donc une puissance statique P_{leak}.

En considérant une fréquence d'émission des impulsions égale à f, on obtient l'équation suivante :

$$\frac{1}{f} = t_{standby} + t_{active} = t_{period} \tag{II.11}$$

Pendant ces 2 périodes, l'énergie d'un potentiel d'action E_{spike} est estimée à 2 pJ (expression utilisée par (49), (87), (54)) et la puissance statique causée par les courants de fuites P_{leak} vaut 1.2 nW. Ces deux unités sont utilisées puisque t_{active} et les courants de fuites sont supposés constants. On peut cependant définir une énergie ou une puissance si on fixe la fréquence d'émission du neurone. Ceci s'exprime par une puissance moyenne P_{spike} pendant t_{active} et une énergie dissipée pendant $t_{standby}$.

L'équation suivante exprime la puissance dissipée par le neurone P_f selon sa fréquence d'émission f :

$$\begin{aligned} P_f &= P_{spike} + P_{leak} \\ P_f &= E_{spike} * f + P_{leak} \\ P_f &= 2.10^{-12} * f + 1.2.10^{-9} \end{aligned} \tag{II.12}$$

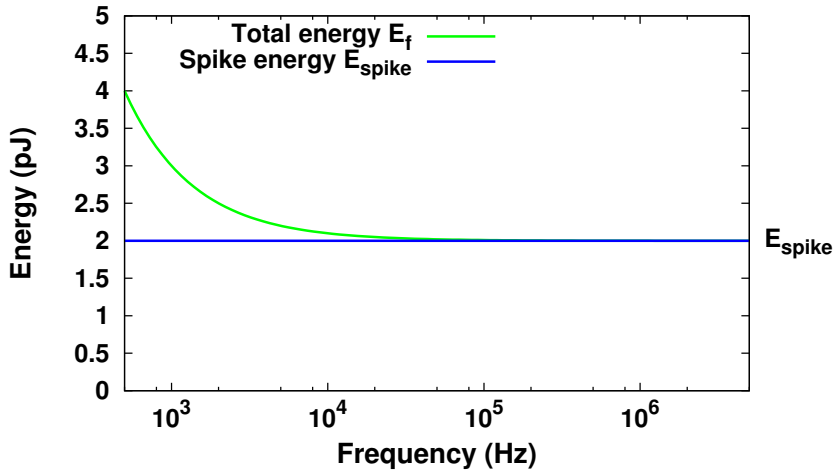


Figure II.24: E_f correspond à l'énergie dissipée pour la génération de potentiels d'action à une fréquence f. Elle inclue l'énergie nécessaire à l'émission du potentiel d'action (puissance active fonction de la fréquence) et la puissance statique. A partir de $f = 50\ kHz$, la dissipation d'énergie causée par les courants de fuite est négligeable pour considérer $E_f = E_{spike}$

Des fréquence typiques pour des études haut niveau sont : $f_{mean} = 100\ kHz$, $f_{min} = 1\ kHz$ et $f_{max} = 1\ MHz$. Les puissances suivantes correspondent au cas nominal :

$$\begin{aligned} P_{f_{mean}} &= 200\ nW \\ P_{f_{min}} &= 3\ nW \\ P_{f_{max}} &= 2\ \mu W \end{aligned} \tag{II.13}$$

E_f représente l'énergie dissipée pendant t_{period}. Elle est tracée sur la figure II.24 en fonction de la fréquence d'émission f. La courbe est obtenue à l'aide de l'équation II.12 lorsque chacun des membres est divisé par f. A basse fréquence ($t_{standby} \gg t_{active}$), l'énergie pour générer un potentiel d'action est principalement dissipée pendant $t_{standby}$. L'énergie dissipée est donc fortement dépendante de la fréquence. En moyenne-haute fréquence, l'énergie dissipée pendant t_{active} est supérieure à celle dissipée entre la génération des potentiels d'action. Par conséquent, E_f est égale à E_{spike}.

5. Réalisation du layout du neurone

Après avoir validé le comportement du neurone à l'aide des simulations, les différents masques nécessaires à la fabrication sont dessinés dont une vue d'ensemble est présentée en figure II.25. La forme en carré a été pensée pour l'implémentation de bloc de neurones et permettre ainsi le regroupement de parties analogiques au sein du circuit.

Les transistors sont positionnés sous la capacité pour minimiser la surface consommée, qui sera égale à 120 um^2. Toutes les structures de type miroir ou paire différentielle respectent une symétrie centrale afin de minimiser les impacts de fabrication. En haut à droite se trouve la paire différentielle du comparateur à laquelle sont réliée C_m et V_{th}. Les miroirs de courants pour l'injection et la décrémentation se trouvent respectivement en haut à gauche et à droite. Les 16 transistors responsables de la fuite programmable, détaillés dans le tableau II.4, sont au milieu en bas. L'interrupteur S_{pulse} est placé en bas à gauche de la cellule et est entouré par un anneau de garde. Utilisé comme isolant, il a pour objectif d'absorber les charges générées lors du basculement de l'interrupteur et d'atténuer la transmission de fréquences parasites au neurone.

Dans un souci d'intégration afin de former une tuile, toutes les entrées/sorties numériques sont en bas, alors que les entrées analogiques se situent sur les côtés à droite et en haut. L'extraction des parasites est réalisée à partir de ce layout et permet la génération d'un nouveau schéma qui tient en compte la proximité et les interactions des composants. De nouvelles simulations ont été effectuées qui ne montraient pas d'évolutions majeures. Douze neurones ont alors été assemblés comme indiqués sur la figure II.26. Ils forment ainsi une tuile qui pourra, à son tour, être intégrée dans une bloc numérique.

Conclusion

Dans cette partie, nous avons réalisé le neurone en nous appuyant sur les spécifications de l'architecture et en respectant les contraintes de fabrication. Des choix de topologies ont été faits dans l'optique d'obtenir un neurone robuste aux procédés de fabrication et capable de faire du calcul. Les simulations ont montré un comportement en adéquation avec ses spécifications. Sa variabilité a été estimée afin de connaitre les limites de l'implémentation. Un calcul préliminaire de l'énergie du neurone par impulsion a permis de prédire sa consommation. Après sa conception, nous l'avons intégré dans les circuits que nous présenterons dans les paragraphes suivants.

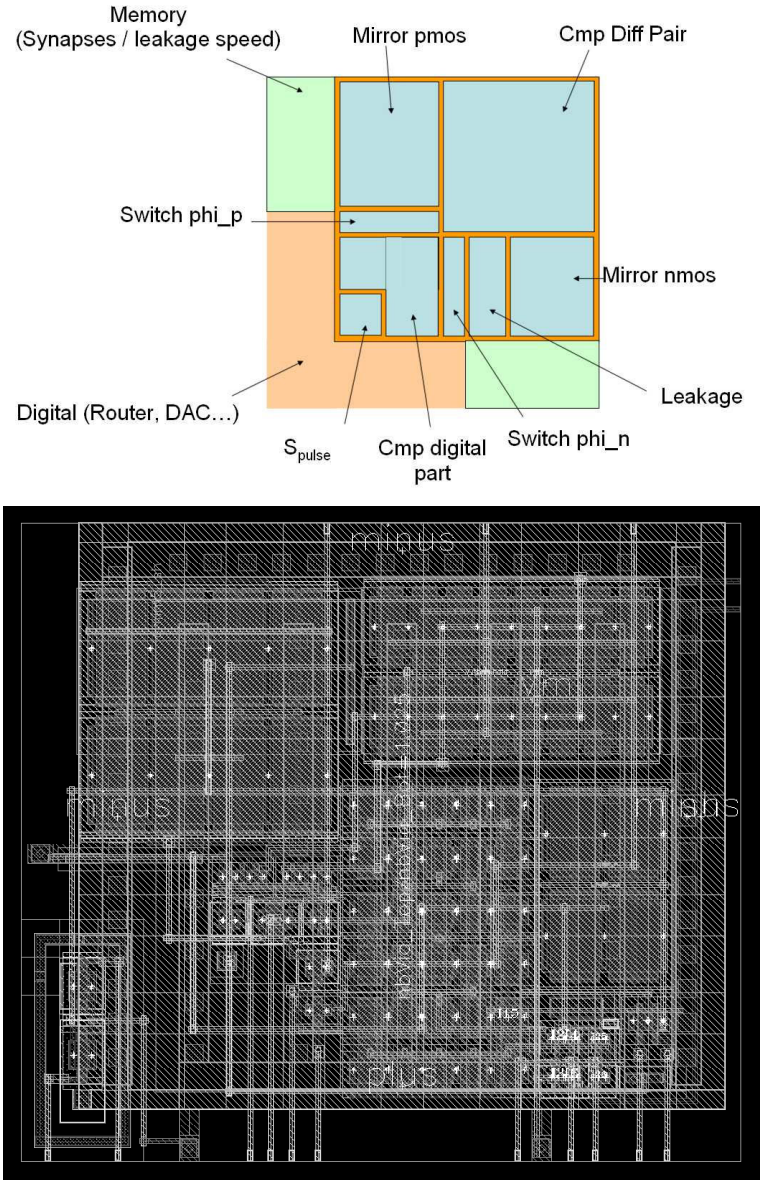


Figure II.25: En haut, schéma de préparation du layout des transistors du neurone. En bas, layout du neurone réalisé.

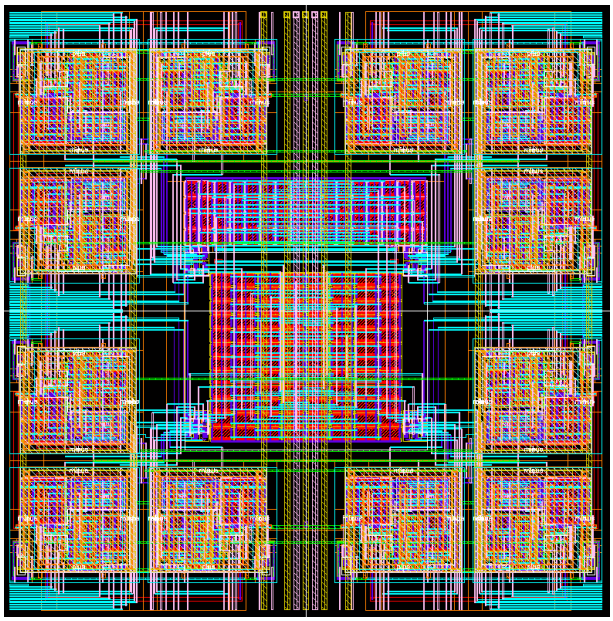


Figure II.26: Layout d'une tuile comportant 12 neurones.

C. Circuits réalisés

Deux circuits basés sur la technologie ST 65 ont été réalisés dans le cadre du projet *Arch*^{2} *Neu*. L'interfaçage entre l'analogique et le numérique étant critique, l'objectif principal du circuit *Reptile* était de valider le fonctionnement du neurone avec le DAC. Le second circuit, *Spider*, répond au problème du passage à l'échelle d'une architecture neuromorphique et étendra les possibilités de calcul et donc les applications potentielles.

1. Implémentation dans le circuit *Reptile*

A partir du neurone conçu, *Reptile* a été réalisé dans l'optique de :

- valider l'intégration du neurone analogique et son interfaçage avec l'environnement numérique

- montrer la faisabilité d'opérations élémentaires en connectant des neurones entre eux

- mesurer la consommation d'un circuit basé sur des neurones à impulsions

- caractériser la variabilité des neurones

L'architecture implémentée est présentée sur la figure II.27. Elle permet de stimuler des neurones mais également de tester indépendamment le comparateur et d'observer précisément un neurone. La programmation du circuit se fait par le biais d'un protocole RS232 et des registres chainés du circuit. La boucle à verrouillage de fréquence (FLL) (1) est en charge de générer l'horloge du circuit à 500 MHz et sera utilisée par les DAC pour contrôler les interrupteurs S_{pulse}. La matrice d'interconnexion contient le poids synaptique entre des neurones. Leur stimulation peut se faire soit de l'extérieur soit à l'aide de générateurs de Poisson dont la fréquence moyenne est réglable. Chacun des neurones possède son propre DAC. Leurs potentiels d'action peuvent être observés, retardés et/ou propagés vers d'autres neurones.

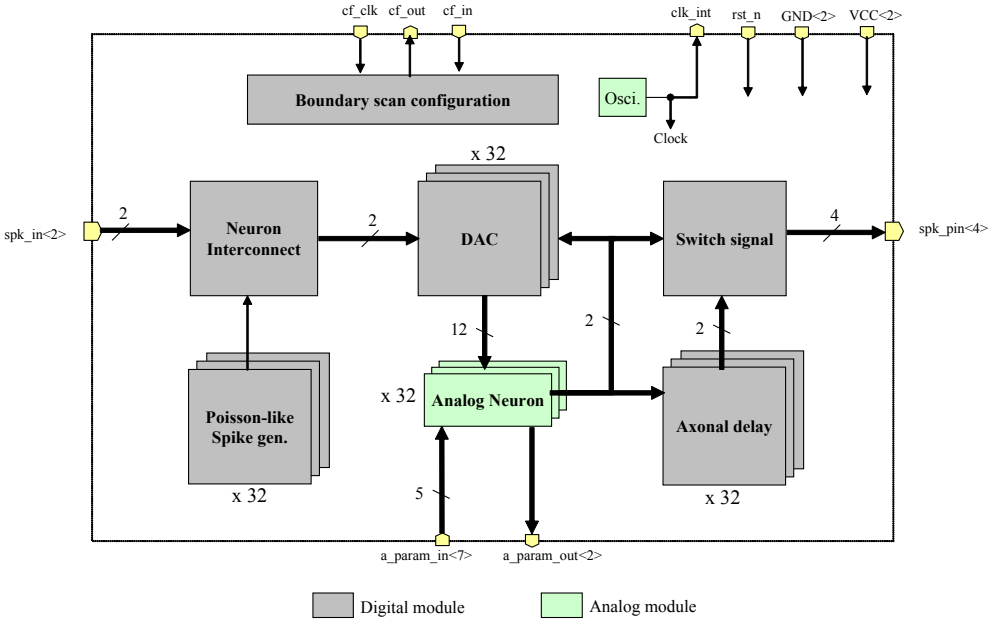


Figure II.27: Architecture du premier circuit de test *Reptile*

a) Tuiles de neurones analogiques

Une tuile est formée de 12 neurones analogiques en périphérie, au milieu desquels se trouvent des miroirs de courant en charge d'alimenter les neurones. Ceci est nécessaire puisque chacun d'entre eux est relié à un DAC qui leur permettent d'être simultanément actifs.

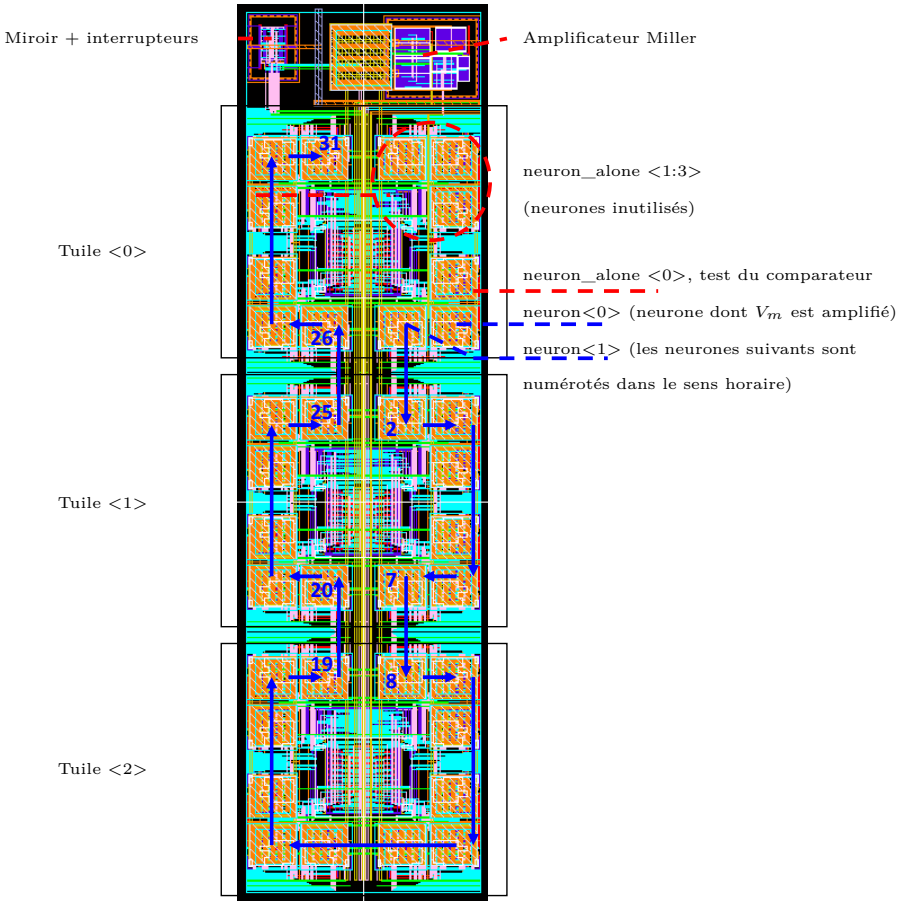


Figure II.28: Organisation physique de la partie analogique du circuit *Reptile* : alimentation des tuiles, amplificateur Miller, 3 tuiles de 12 neurones.

Les 36 neurones implémentés sont répartis en trois tuiles de douze neurones placées l'une en dessous de l'autre comme indiqué sur la figure II.28. Parmi eux, 32 sont interconnectés dans le réseau par la partie numérique du circuit. Le comparateur d'un neurone déconnecté est instrumenté afin de tester uniquement ce sous-bloc. Trois des quatre restants, sont entièrement inactifs : leurs entrées sont maintenues à Vdd ou Gnd. On pourra noter que toutes les entrées/sorties numériques se font sur les côtés.

Les tuiles ont plusieurs potentiels de polarisation. La tension de seuil, le potentiel de fuite, le potentiel de reset sont indépendants les uns par rapport aux autres, et communs à tous les neurones. Le courant de soustraction est répliqué à partir du courant d'injection par un miroir de courant. Sa valeur est commune à toutes les tuiles mais

chacune peut être activée séparément à l'aide du miroir dans le coin supérieur gauche.

b) Amplification du potentiel de membrane du neurone<0>

Un amplificateur de type Miller a été conçu pour l'analyse de signaux analogiques. Monté en suiveur dans le circuit, il permet l'observation d'un nœud en diminuant l'impact éventuel d'un instrument de mesure. Il isole le potentiel observé et peut fonctionner théoriquement jusqu'à 500 MHz avec une charge en sortie de 5 pF. Son schéma est présenté sur la figure II.29. L'amplificateur est connecté au potentiel V_m du neurone<0> afin de suivre son évolution.

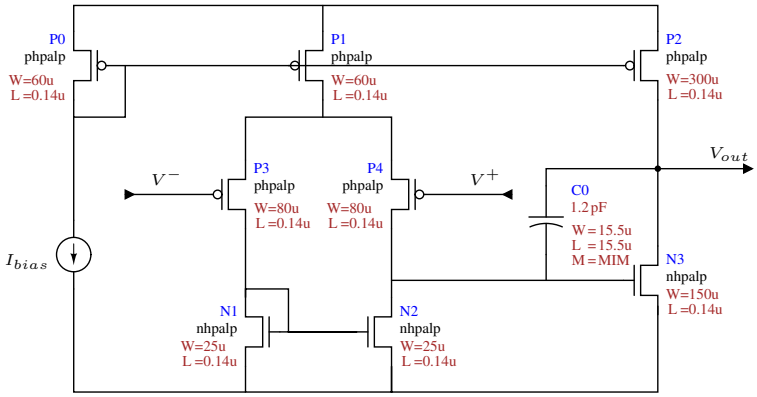
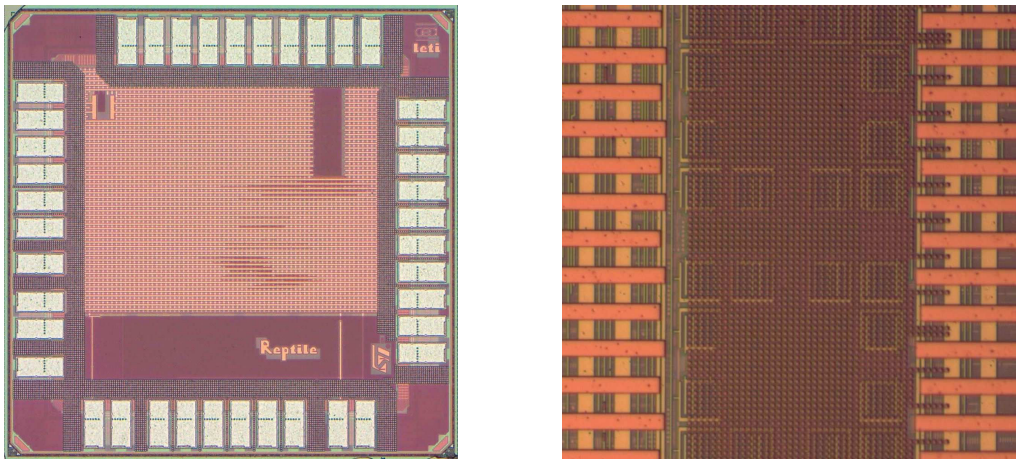


Figure II.29: Schéma de l'amplificateur pour l'observation du potentiel de membrane. Il est connecté en suiveur dans les circuits *Reptile* et *Spider*. I_{bias} est de l'ordre du mA (0.9 à 1.1 mA) et polarise les transistors de l'amplificateur à l'aide d'un générateur externe.

c) Vue générale

Une photographie au microscope d'un exemplaire du circuit nu, par opposition à ceux en boitier, est visible sur la figure II.30a. On y identifie les entrées/sorties, en périphérie du circuit, qui sont reliés aux pattes du boitiers. On observe les grilles d'alimentation de la partie numérique formées par les métaux 6 et 7. La partie inférieure, où se trouve le logo *Reptile*, est majoritairement inutilisée. Les deux autres zones noires constituent les parties analogiques : en haut à gauche se trouve la FLL, en haut à droite est intégrée la partie analogique neuronale présentée sur la figure II.28. La surface totale du circuit mesure 1 mm^2.



(a) **(b)**

Figure II.30: Photographies du circuit *Reptile* : **(a)** Vue globale, cœur du circuit et couronne de plots (taille 1mm × 1mm). **(b)** Zoom sur les neurones. On observe les tuiles générées pour la densité au niveau du métal 7 ainsi que le métal 6 d'accès aux capacités MIM.

La photographie sur la figure II.30b est un agrandissement de la zone analogique du circuit qui contient les neurones. On y distingue les carrés de métaux nécessaires à l'homogénéité de la surface du circuit lors des étapes de fabrication. Par transparence, on peut distinguer les métaux inférieurs dont ceux d'accès à la capacité MIM des neurones.

2. Caractérisation et test

a) Environnement de test

Une carte de test, présentée sur la figure II.31 et en annexe V-B., a été réalisée pour permettre la caractérisation du circuit. Elle permet l'observation directe des signaux analogiques à l'aide d'un oscilloscope. Pour les signaux à haute fréquence, une sonde Agilent 1156A a été utilisée pour visualiser le potentiel V_m et la sortie de la FLL.

Sur la photographie, on identifie les alimentations, en rouge, des parties analogique et numérique du circuit ainsi que de la couronne de plots en 2.5 V. Le cœur du circuit fonctionne en effet à 1.2 V et les plots servent d'intermédiaire pour protéger le circuit des décharges électro-statiques.

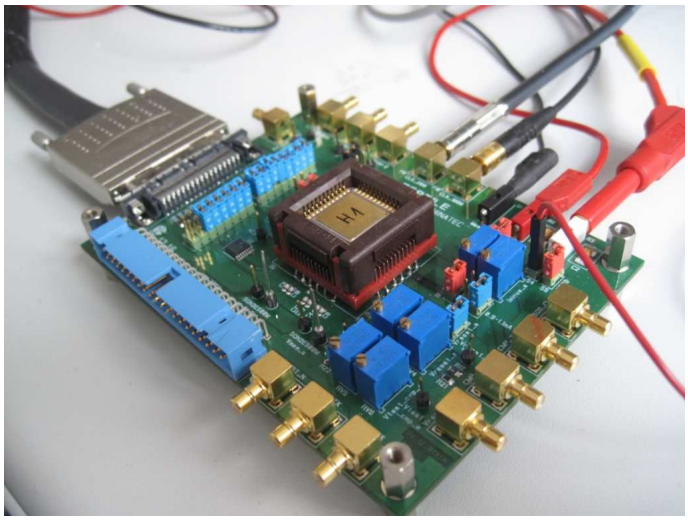


Figure II.31: Carte de caractérisation et circuit Reptile

Deux horloges sont connectées à la carte sur son bord supérieur (câbles noirs). Elles permettent la synchronisation de la FLL et le séquençage de l'envoi de donnée pour la configuration de la puce. Six potentiomètres (en bleu) sont chargés de générer et de régler les potentiels et courants de polarisation.

Les entrées numériques peuvent être générées soit à l'aide d'un FPGA soit à l'aide d'un PC. En plus de ces deux options, les sorties numériques sont également observable à l'oscilloscope.

Nous n'avons pas opté pour l'utilisation d'un FPGA comme en témoigne le connecteur bleu, à gauche, laissé vacant. La stimulation du circuit se fait à l'aide d'un châssis NI PXIe-1062Q de National Instrument, visible à gauche sur la figure II.32, dans lequel est insérée une carte de contrôle NI PXIe-6545. L'ensemble permet l'envoi et l'acquisition de données par le biais d'une interface logicielle Labview créée spécifiquement pour le test du circuit. Tous les signaux utilisent cette interface, excepté les signaux analogiques, les alimentations et l'horloge à 20 MHz permettant l'envoi des données.

b) Comportement général

La configuration du circuit permet de générer l'horloge à 500 MHz nécessaire au fonctionnement du neurone. Une fois celle-ci établie, on s'est attaché aux réglages des différents potentiomètres pour obtenir les valeurs envisagées lors de la simulation. Dès

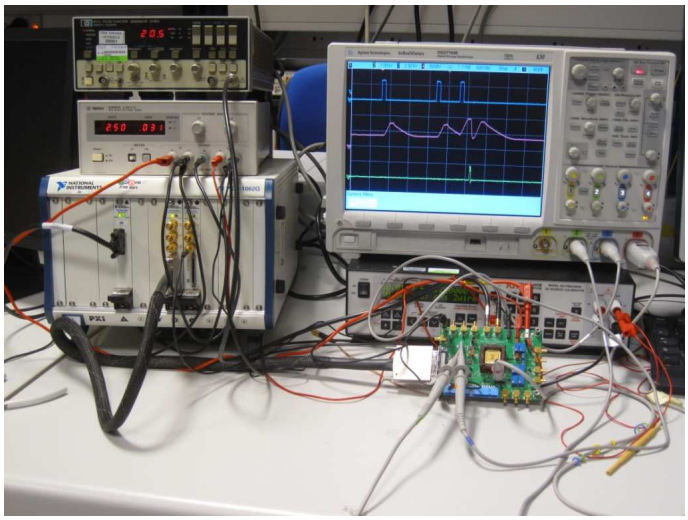


Figure II.32: Environnement de caractérisation : oscilloscope, carte de test, circuit, générateur de fréquence, sources d'alimentation.

lors, on a pu vérifier le bon comportement du comparateur et poursuivre par l'observation du neurone instrumenté.

La figure II.33 présente l'évolution du potentiel de membrane du "neuron<0>" configuré avec fuite lorsqu'il est soumis à trois impulsions successives de poids synaptiques + 0.8. Lorsque 2 impulsions sont suffisamment proches, leur corrélation est détectée, un potentiel d'action est généré, V_m retourne à V_{reset}. On vérifie un fonctionnement tout à fait conforme à celui prévu lors de la simulation.

Ces premiers tests témoignent de la difficulté d'observer des signaux à haute fréquence dans la mesure où ceux-ci sont matériellement limités. Le signal d'horloge à 500 MHz généré par la FLL est modulé par les deux autres fréquences d'horloge à 100 et 20 MHz. Son signal, utilisant pourtant un plot analogique pour augmenter la fréquence de coupure, est fortement atténué. De même, le bruit généré par la partie numérique et les circuits RC parasites ne permettent pas d'identifier précisément les incréments du potentiel de membrane du neurone.

c) Consommation

Le protocole établi pour l'étude de la consommation du circuit est le suivant. Deux sources de tension sont connectées pour alimenter séparément les parties analogique et

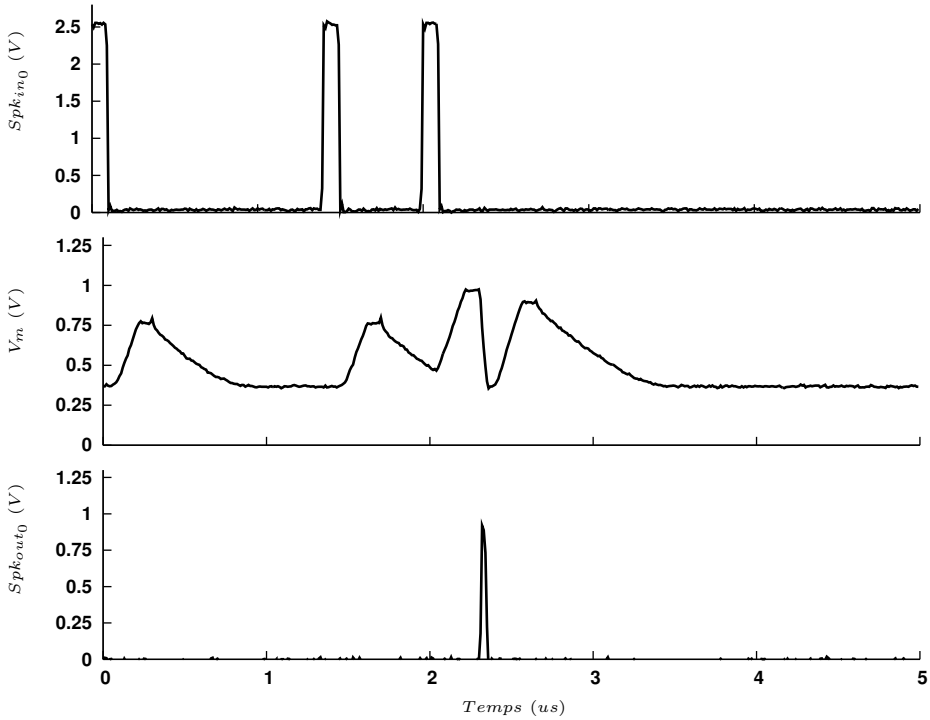
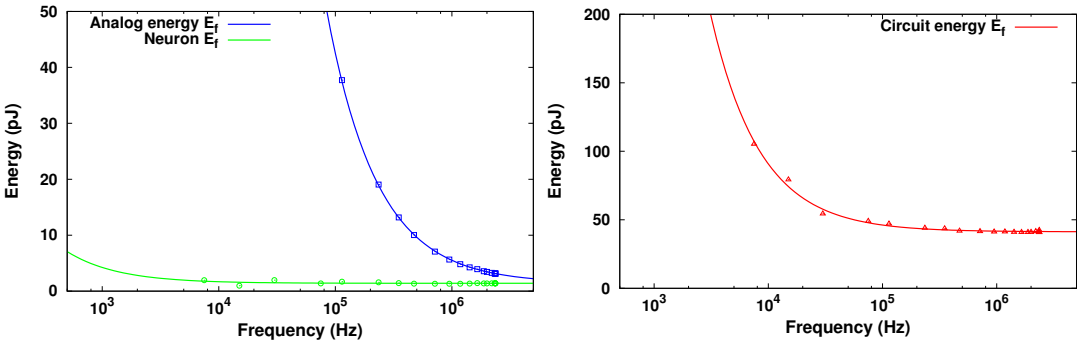


Figure II.33: Comportement d'un neurone en présence de 3 stimuli successifs sur Spk_{in_0} : lorsqu'ils sont suffisamment corrélés, le neurone émet un potentiel d'action.

numérique. La configuration du circuit active 30 boucles dans le but d'étudier la relation entre fréquence des impulsions et consommation du circuit. Une boucle comporte un générateur de Poisson, une partie de la matrice d'interconnexion, un DAC et un neurone.

Les générateurs de Poisson stimulent les neurones à différentes fréquences. On mesure le courant débité dans chaque partie du circuit. La puissance moyenne est calculée en le multipliant par la tensison d'alimentation. Selon la fréquence de stimulation, on déduit la valeur de l'énergie consommée en fonction de la fréquence tracée sur la figure II.34. Les courbes d'ajustement ont le même type de fonction que dans l'équation II.12.

Les résultats de la figure II.34a sont en accord avec les simulations (cf fig.II.24). L'énergie E_{spike} est estimée à 1.4 pJ (2 pJ par simulation). Elle correspond à la valeur de E_f pour un neurone stimulé à haute fréquence. L'énergie E_f de toute la partie analogique s'éloigne très rapidement de la valeur E_{spike}. La consommation statique des



(a) **(b)**

Figure II.34: Caractérisation de l'énergie d'un spike dans le circuit *Reptile* : **(a)** En vert, énergie d'un neurone permettant de générer une impulsion. En bleu, énergie consommée par la partie analogique (tuiles de neurones, miroir d'alimentation des tuiles). **(b)** Énergie totale d'une boucle numérique+analogique pour générer une impulsion.

Référence	Techno (um)	Modèle	E_{spike}	E_{tot}
(87)	0.35	Izhikevich	8.5-9.0 pJ	-
(33)	1.5	IF	3-15 nJ	-
(49)	0.35	$\sim$ aEIF	7-267 pJ	-
(54)	0.045 SOI	Digital IF	-	45-80 pJ
Circuit *Reptile*	0.065	IF	1.4-10 pJ	40 pJ

Table II.8: Caractéristiques énergétiques de neurones implémentés dans la littérature. En étant au niveau de l'état de l'art, le circuit *Reptile* ouvre des perspectives nouvelles pour des architectures neuromorphiques basse consommation.

tuiles activées pour les neurones est une première explication. Les courants de fuite, plus élevés qu'en simulation puisque $Vdd_{analog} > Vdd_{nominal}$, en sont une autre.

Au niveau du circuit, c'est-à-dire en considérant la boucle complète analogique et numérique, la figure II.34b indique une énergie E_{spike} de l'ordre de 41 pJ. Ceci est à mettre en comparaison avec les travaux récents publiés par IBM (54). L'énergie nécessaire par impulsion est de 45 pJ en technologie 45 nm. Des résultats de recherches d'équipes plus anciens sont également présentées dans le tableau II.8.

d) Fuite intrinsèque et programmable

Les caractérisations de la fuite du "neuron<0>" sont montrées sur la figure II.35. Les mesures montrent que la fréquence minimum de fonctionnement de 1 kHz est respectée. En se référant à la courbe sans fuite programmée (en bas à droite) et à la mesure de la constante de temps exprimée dans le tableau II.9, on observe que le potentiel de membrane diminue de l'ordre de 3.6 mV en $1ms$. Cette valeur est approximative puisqu'elle est acquise à l'oscilloscope par le biais de l'OTA. Elle est une image de ce qui se passe réellement dans les autres neurones. En effet, les fuites à travers la grille du transistor d'entrée de l'étage différentiel de l'OTA contribue à la diminution du potentiel de membrane. Sous forme d'équation, la rétention vérifie bien $\frac{1}{\delta V * f_{low}} = 0.25\ s \simeq \tau$.

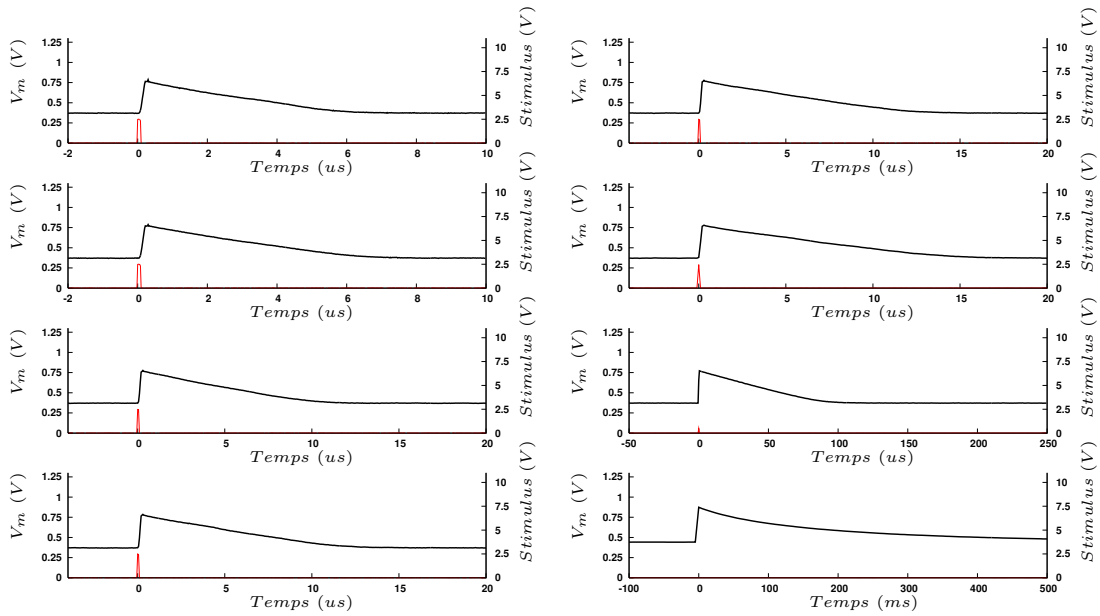


Figure II.35: Fuite programmable du neurone après augmentation du potentiel causé par une impulsion et un poids égal à +0.8. Elle est présentée de la plus à la moins importante de haut en bas et de gauche à droite. Causé par l'échantillonnage de l'oscilloscope, le potentiel d'action entrant disparait pour des temps d'acquisition trop longs.

Les autres constantes de temps ont été mesurées et sont également présentées dans le tableau II.9. Selon leur configuration, celles-ci pourront soit être utilisées à des fins de purs intégrateurs soit détecter des coïncidences entre deux impulsions. Si on se réfère au tableau II.4, on note un décalage par rapport aux résultats attendus. Il n'est toutefois

pas possible à ce stade de déterminer si la cause provient de la fabrication ou de la présence de l'OTA. La dynamique du potentiomètre actuellement soudé sur la carte ne permet pas d'augmenter la fuite suffisamment.

Valeur de config	S_{τ_2}	S_{τ_1}	S_{τ_0}	Fuite mesurée
0	0	0	0	5.68 us
1	0	0	1	6.24 us
2	0	1	0	10.11 us
3	0	1	1	11.5 us
4	1	0	0	12.12 us
5	1	0	1	14.61 us
6	1	1	0	89.99 us
7	1	1	1	$[220; 256]$ ms

Table II.9: Caractéristiques mesurées de la fuite du "neuron<0>" selon ses bits de configuration

e) Variabilité des neurones

En relation avec les simulations réalisées dans les partie II-B.4.b) et II-B.4.c), nous avons caractérisé la variabilité des neurones. Ceci a été fait pour 310 neurones sur dix circuits différents, dans lesquels nous n'avons pas inclus le résultat du neurone instrumenté.

Le poids de la synapse d'entrée du neurone est fixé initialement autour de +0.7 puis progressivement augmenté. Tant que le poids de déclenchement n'est pas atteint, la stimulation du neurone testé n'a pas d'impact. Lorsque le neurone émet des impulsions à sa fréquence maximale de fonctionnement, cela signifie que le poids critique pour le seuil de déclenchement a été atteint.

La figure II.36 montre la distribution des poids nécessaires à la génération d'un potentiel d'action des différents neurones. La courbe gaussienne de lissage est de moyenne $\mu = 100.17$ et d'écart type $\sigma = 6.25$. En effet, une moyenne inférieure à 128 (correspondant à 7 bits) a été privilégiée, dans un souci de simplification du protocole de test et de validation comportementale du circuit avant caractérisation du "neuron<0>". On peut en effet stimuler un neurone avec une seule synapse, dont le poids est également codé sur 7 bits. Les écarts entre la variabilité estimée par simulation ($\sigma = 3.9$) et

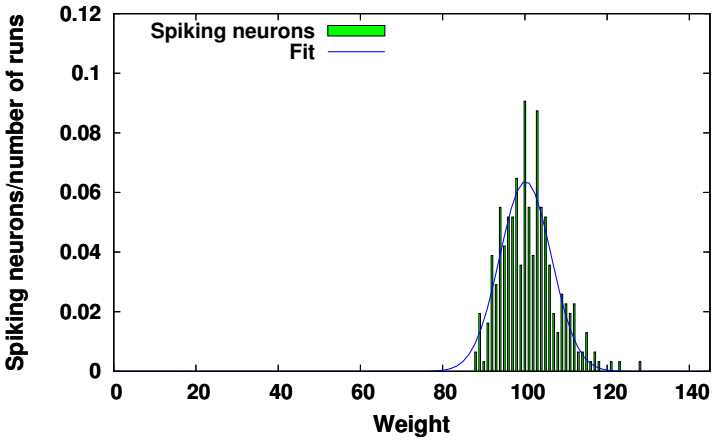


Figure II.36: Caractérisation de la variabilité des neurones. Nombre de neurone ayant généré un potentiel d'action à poids donné. L'échantillon est de 310 neurones, la moyenne $\mu = 100.17$ et $\sigma = 6.25$.

les valeurs mesurées ($\sigma = 6.25$) peuvent tout d'abord s'expliquer par le maintien de l'ensemble des paramètres de test. En effet, les tensions et les courants de polarisation n'ont pas été modifiés entre les caractérisations de chaque circuit. Cette approche se justifie par le nombre de neurones par circuit (32) qui n'aurait effectivement pas permis d'étudier la variabilité sur un échantillon suffisamment représentatif.

Une seconde explication provient des spécifications et de la conception du neurone. La simulation montrait la variabilité intrinsèque du neurone équivalente à 30.3 dB dans un environnement idéal. La caractérisation du circuit mesure, quant à elle, la variabilité totale qui comprend également les blocs suivants :

- Le contrôle numérique de l'interrupteur
- Le miroir de la tuile pour alimenter les neurones
- Le miroir pour alimenter les tuiles

Il est donc tout à fait normal de mesurer une variabilité plus élevée.

f) Vers une procédure de réglage automatique des paramètres

Après les caractérisations précédentes, on réalise que les procédés de fabrication ont eu un impact important sur le circuit. On peut légitimement penser à une méthode

pour déplacer la gaussienne, ou même réduire l'écart type de cette distribution. Tout comme un même modèle de processeur est vendu selon la fréquence maximale qu'il supporte, on propose ici quelques étapes pour optimiser l'utilisation du circuit :

- Réglages matériels initiaux -

Ils sont nécessaires pour obtenir une puce fonctionnelle en vue de sa caractérisation. On utilisera pour ceci un neurone dont le potentiel de membrane est observable, dans notre cas le "neuron<0>". A ce stade, il est important de vérifier le bon comportement du neurone et de son interface avec le numérique, à savoir : une comparaison qui génère une impulsion captée par le numérique, une remise à V_{reset}, la gestion des poids positifs et négatifs. Ceci se fera en plaçant préalablement les différents potentiels et courants de polarisation à leur valeur typique, si besoin en les modifiant à la main successivement.

- Caractérisation -

Elle permet de connaitre la configuration générale des neurones du circuit. La caractérisation d'un échantillon représentatif du nombre total des neurones intégrés permet de déterminer la moyenne et l'écart type de la distribution de leur poids de déclenchement. Ceci devra, contrairement à ce qui a été fait dans la partie précédente, être réalisée de manière totalement automatisée.

- Modification des paramètres pour une utilisation optimale -

Les valeurs V_{th}, V_{reset}, I_{plus} et la fréquence peuvent être modifiées pour optimiser l'emploi de l'ensemble des neurones afin de se placer au point nominal de fonctionnement du circuit. En les modifiant, on placera la courbe gaussienne de manière à la centrer en 128. Si la partie numérique est suffisamment contrainte lors de la conception, on peut augmenter $Iplus$ et la fréquence, ceci diminuera alors l'écart type de la gaussienne.

A présent, il suffira alors d'itérer ces deux dernières étapes pour placer le circuit à son point optimal de fonctionnement en exploitant au mieux une majorité de neurones. On notera cependant que cette optimisation est réalisable seulement si les caractérisations des neurones sont automatisées. Bien que le positionnement de la gaussienne sur 128 est réalisable par un simple ajustement de V_{th}, le protocole de test actuel ne permet pas une nouvelle caractérisation rapide de l'écart type.

3. Évolutions dans le circuit *Spider*

Les évolutions des neurones analogiques au sein du circuit *Spider* répondent aux observations des caractérisations ainsi qu'à la nouvelle topologie du circuit. Étant toujours en fabrication, il n'a pas encore pu faire l'objet de caractérisations électriques.

a) Vue générale

Spider, dont le layout est présenté sur la figure II.37, a pour objectif de répondre au problème de passage à l'échelle d'une architecture neuromorphique. Il est constitué de 25 modules mixtes analogiques/numériques contenant chacune une tuile de 12 neurones analogiques soit un total de 300 neurones. L'avancée majeure réside dans l'utilisation d'un motif élémentaire réplicable autant de fois que nécessaire.

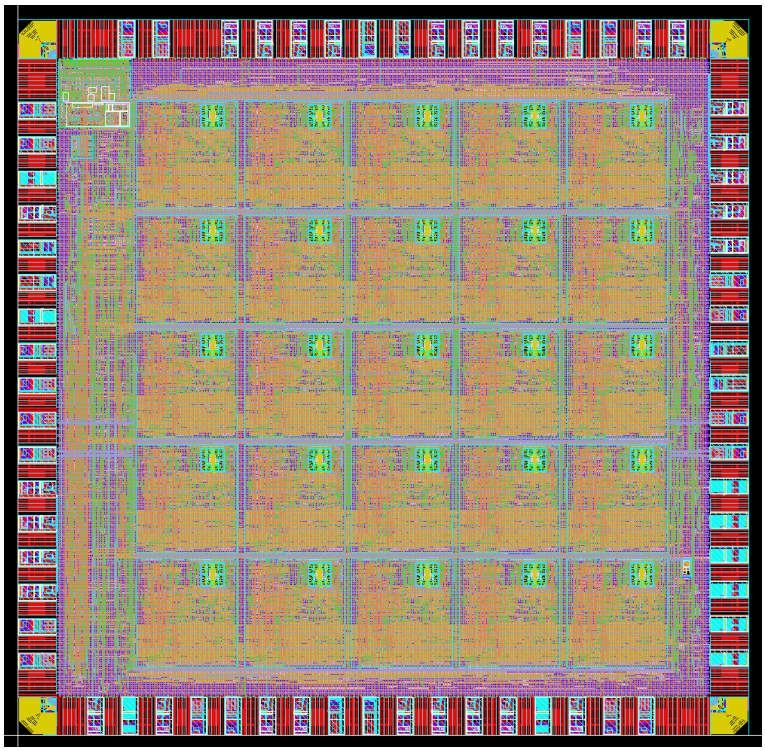


Figure II.37: Layout du deuxième circuit réalisé *Spider*. On distingue : la matrice de 5*5 modules mixtes analogiques/numériques, la couronne de plots, une cellule de compensation requise pour adapter leur fonctionnement post-fabrication. D'autres blocs comme les générateur de Poisson et les blocs de délai sont situés en périphérie.

Il n'existe plus qu'un DAC par module. Par conséquent, les potentiels d'action entrants dans un module sont stockés dans une FIFO entrainant la stimulation séquentielle des neurones. La surface de la tuile aurait pu être réduite par le retrait des miroirs de courants centraux. Par gain de temps, ceci n'a pas été fait et le layout de la tuile de 12 neurones de *Reptile* a été réutilisé.

b) Modifications par rapport au circuit *Reptile*

Suite aux caractérisations du circuit Reptile, certains points ont été modifiés. L'interfaçage de la partie numérique avec les neurones a en effet montré quelques dysfonctionnements. La comparaison n'ayant pas toujours le temps d'être effectuée, le neurone ne générait aucune impulsion et son retour au potentiel de repos n'était pas commandé par la partie numérique. Lors du test du circuit *Reptile*, il a été nécessaire d'augmenter la tension d'alimentation des neurones. Ainsi le comparateur a pu effectuer une comparaison dans un intervalle de temps plus court. La tension à appliquer sur les transistors analogiques ont dû par conséquent être de l'ordre de 1.34 V. Pour résoudre ce problème dans l'implémentation du circuit *Spider*, le temps d'activation du comparateur par le signal S_{cmp} a quant à lui été augmenté de 10 à 20 ns.

Dans le circuit *Reptile* et comme en témoigne la figure II.38, le courant est distribué tuile par tuile et permet de ce fait leur inactivation. Ce courant de polarisation est remplacé par une tension appliquée sur l'ensemble des transistors alimentant le miroir des tuiles dans *Spider*.

Du point de vue conception, l'intégration de la tuile analogique est améliorée. L'intégration dans *Reptile* fut longue et potentiellement source d'erreurs. Chaque bloc fut intégré manuellement dans un autre puis connecté, au niveau final, à la couronne de plot. Dans *Spider*, les neurones sont encapsulés dans un module numérique. Il bénéficie d'un niveau final totalement numérique dont l'assemblage est automatisé.

Conclusion

Cette partie a montré la mise en place du flot de conception d'un neurone analogique. Celui-ci a été utilisé pour la fabrication d'un circuit dans une technologie maitrisée ($ST65$). Les caractérisations du circuit *Reptile* ont confirmé le bon fonctionnement du neurone analogique et conforté son intégration dans un second circuit *Spider*. Son

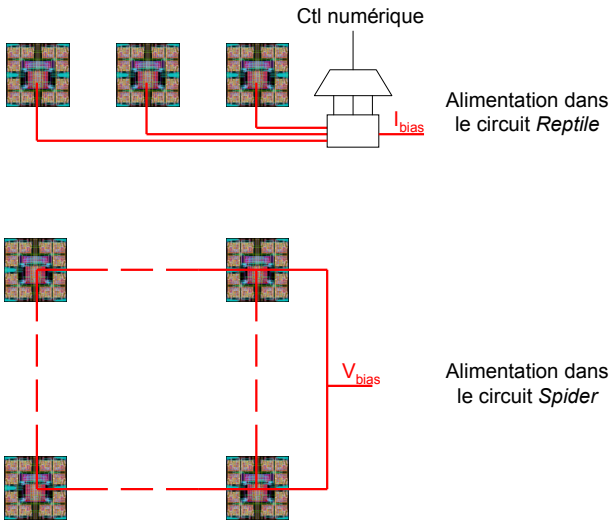


Figure II.38: Evolution de l'alimentation des tuiles du circuit *Reptile* au circuit *Spider*

comportement est conforme aux spécifications et aux simulations. Confortant les valeurs obtenues par simulations, l'énergie intrinsèque du neurone mesurée est sensiblement inférieure à 2 pJ, ce qui le place au niveau de l'état de l'art mondial. La variabilité a été précisément étudiée sur un ensemble représentatif de neurones. L'écart par rapport à la simulation s'explique par la caractérisation de plusieurs circuits, mais aussi dans une moindre mesure par la variabilité d'autres blocs intégrés.

II. INTÉGRATION D'UN NEURONE ROBUSTE POUR DES APPLICATIONS COMPUTATIONNELLES

III

Études sur les technologies avancées

La vraie nouveauté naît toujours dans le retour aux sources.

Edgar Morin, *Amour, poésie, sagesse*, 1997.

Sommaire

A. Quel avenir pour un neurone analogique ? 86
1. Comparaison au nœud technologique 65 nm 86
2. Extrapolation vers des nœuds avancés 88
B. Notions de mémoire résistive 89
1. Un phénomène ancien, une théorie récente 89
2. Un intérêt grandissant dans la communauté neuromorphique 90
3. De nombreux phénomènes physiques 92
4. Précisions sur les mémoires résistives utilisées 93
C. Cœur du neurone : élément capacitif 95
1. État de l'art des capacités 96
2. Les capafils 96
3. Utilisation d'un *Through Silicon Via* 98
4. Utilisation d'une mémoire à changement de phase 102
D. Implémentation des connexions synaptiques 106
1. Utilisation des mémoires à changement de phase 106
2. Utilisation des mémoires à filament conducteur 107

De par leur nouveauté, les technologies émergentes sont souvent mal maitrisées. Parfois, il semblerait que leur développement ne permette plus un contrôle total. En raison de leur très faibles dimensions, elles sont effectivement plus sensibles lors de leur fabrication et donc sujettes à des phénomènes aléatoires.

De par son fonctionnement massivement parallèle, une architecture neuromorphique reste fonctionnelle même si les disparités technologiques sont fortes. L'utilisation de technologies avancées est cependant rendue difficile puisqu'il existe peu d'outils pour les étudier. Certains dispositifs émergents ont par conséquent nécessité le développement de modèles ou de modules pour les rendre compatibles avec un flot industriel existant.

A. Quel avenir pour un neurone analogique ?

Il s'agit ici d'étudier l'impact qu'auront les procédés technologiques sur la conception d'un neurone. L'implémentation des synapses sera étudiée dans la partie suivante et nous étudierons, par conséquent, l'implémentation du cœur et du comparateur du neurone.

1. Comparaison au nœud technologique 65 nm

En parallèle du développement du neurone analogique décrit dans le chapitre précédent, le développement d'une version numérique du neurone LIF a été nécessaire. Il a permis initialement de réaliser des simulations au sein d'un flot uniquement numérique dans l'objectif de développer l'architecture du circuit. Il nous est alors apparu judicieux d'étayer les propos énoncés dans la partie II-A.2. en analysant ces deux implémentations de neurones. Dans le souci d'effectuer une comparaison équitable des deux versions, ce neurone numérique doit respecter un comportement similaire au neurone analogique. Il a ensuite été optimisé en vue d'une implémentation compacte et à faible consommation. Son architecture est décrite sur la figure III.1.

En lien avec le neurone analogique, on retrouve ici les mêmes blocs fonctionnels rassemblés dans trois parties : mécanisme de modulation de poids, cœur du neurone et comparateur. Le mécanisme de modulation de poids par impulsions est semblable à celui utilisé dans le chapitre précédent. L'entrée d'un bit de polarité dans le cœur du neurone, auparavant incluse dans la partie synapse, a été modifiée dans le but d'optimiser

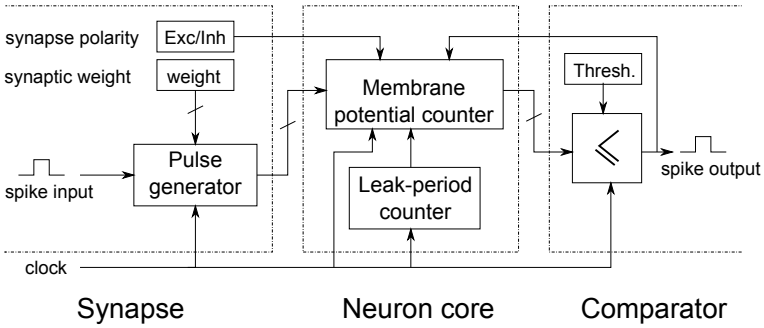


Figure III.1: Implémentation numérique du neurone LIF.

l'implémentation numérique. La capacité de membrane est remplacée par un compteur de sept bits, incrémenté lorsque le poids est positif. Il peut être décrémenté lorsque le poids est négatif ou par l'intermédiaire du compteur de fuite. Sa configuration lui permet d'accélérer ou de ralentir la décrémentation du compteur du potentiel de membrane. Lorsque ce compteur atteint la valeur 127, le comparateur génère un potentiel d'action qui sera également responsable du retour à zéro du potentiel de membrane.

Ce neurone a ensuite été synthétisé à l'aide des portes logiques du flot de conception numérique ST CMOS 65 nm. L'horloge à 256 MHz permet d'obtenir une fréquence maximale de génération d'impulsions du neurone égale à 1,9 MHz afin d'être semblable à l'implémentation analogique.

Le tableau III.1 récapitule les principaux résultats d'implémentations entre un neurone analogique et un neurone numérique en termes de surface et de consommation. On constate que malgré la taille importante occupée par la capacité du neurone analogique, cette implémentation reste cinq fois plus compacte que sa version numérique. Elle est également vingt fois plus efficace en termes d'énergie, ce qui s'explique par la faible contrainte en termes de rapport signal à bruit, la sortie du neurone analogique se régénérant dans une impulsion numérique.

Implémentation	Analogique	Numérique
Fréquence max. des impulsions (Mspike/s)	1.9	1.9
Surface du cœur du neurone (μm^2)	120	538
Énergie du cœur du neurone (pJ/spike)	2	41
Temps de développement	☹	☺

Table III.1: Implémentations d'un neurone LIF : analogique VS numérique

2. Extrapolation vers des nœuds avancés

Étudions maintenant l'impact de la réduction de la longueur du canal de grille en vue d'implémentations analogique et numérique d'un neurone LIF. Si un neurone numérique pourra pleinement tirer avantage du passage à l'échelle, ce n'est pas le cas d'un neurone analogique, ceci en partie à cause de la capacité. Cependant, prenant pour acquis le facteur 5 en termes de surface au nœud 65 nm en faveur de l'analogique, on peut estimer qu'une telle implémentation conservera son avance jusqu'au nœud 22 nm.

Ceci est illustré sur la figure III.2, présentant quelques implémentations de neurones LIFs analogiques (12, 14, 27, 35) et numériques (22). Le passage à l'échelle de l'analogique est estimé selon le rapport de prévisions établies par l'ITRS qui indique certains paramètres comme Vdd ou $\sigma_{V_{th}}$ et ceci jusqu'au nœud 20 nm pour des procédés similaires (substrat massif, CMOS basse consommation). Concernant le numérique, on a estimé la surface selon la réduction de la longueur de grille. La résolution de sept bits et la variabilité des neurones, maintenues constantes, mènent à un ralentissement de la diminution de la surface occupée par un neurone analogique.

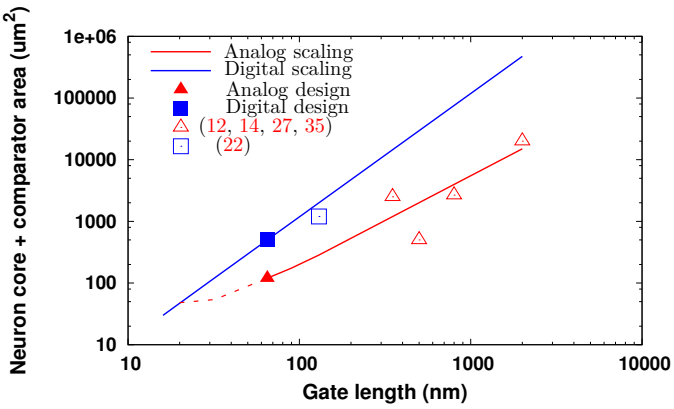


Figure III.2: Passage à l'échelle de neurones LIF : évolution des implémentations analogiques et numériques selon le nœud technologique. Un croisement relativement proche existe et semble se situer autour du nœud 22 nm.

Le neurone analogique ne contient pas d'étage en cascode qu'il serait nécessaire de mettre en cascade. Ceci entrainerait une consommation additionnelle lors de la diminution de V_{dd} dans un nœud plus agressif. En considérant la consommation d'un

neurone analogique fixe, il devrait conserver l'avantage sur son comparse numérique et ceci même si l'on considère une réduction optimiste d'un facteur 2 à chaque nœud technologique. Pour conclure, même si la capacité d'un neurone analogique semble être le principal obstacle à la réduction de sa surface, il continuera de présenter de sérieux atouts et ceci, jusqu'à des nœuds très avancés. Outre les memristors favorisant une approche analogique, l'intégration de capacités ultra-denses au niveau du back-end pourra également être envisagée. Elles seront étudiées et feront l'objet des prochaines parties.

Ainsi, nous commencerons par un petit tour d'horizon des mémoires résistives afin de rappeler brièvement leur comportement et leur intérêt. Des nouvelles technologies, capacitives ou memristives, feront l'objet d'une première étude sur l'implémentation du cœur du neurone. Dans un second et dernier temps, nous proposerons quelques circuits employant ces composants résistifs pour l'implémentation de synapses.

B. Notions de mémoire résistive

Après la présentation d'un premier dispositif étudié ci-après, nous verrons les raisons de l'engouement provoqué par le memristor lorsqu'il est employé dans des architectures neuromorphiques. Nous distinguerons et présenterons plus précisément deux d'entre eux que nous avons choisi d'étudier.

1. Un phénomène ancien, une théorie récente

Le memristor - contraction de *memory* et *resistor* - a été décrit théoriquement en 1971 par Leon O. Chua (16). Pour cela, il s'est appuyé sur les quatre grandeurs physiques de l'électromagnétisme à savoir la tension, le flux, le courant et la charge.

D'une part, la dérivée temporelle permet de relier courant et charge, tension et flux. D'autre part, la résistance, l'inductance et la capacité reliaient respectivement tension et courant, flux et courant, et enfin charge et tension. En raisonnant par symétrie, il a proposé que le memristor de memristance M relie la charge q au flux φ selon l'équation :

$$d\varphi = M dq \tag{III.1}$$

Quand M est constant, l'équation du memristor est équivalente à la loi d'Ohm. Au contraire, si M est fonction de q, de φ ou du temps, il se comporte comme une

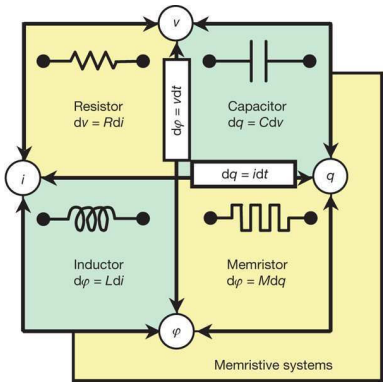


Figure III.3: Relations entre grandeurs électriques et dispositifs élémentaires : résistance, inductance, condensateur et mémristor. D'après (80).

résistance variable dont la valeur dépend du courant qui l'a traversée.

Ce phénomène a été rapporté dans la littérature plusieurs fois et ceci depuis plusieurs décennies (62). La réduction des temps d'observation ainsi que la diminution et la maitrise de faibles dimensions ont permis le rapprochement entre l'effet memristif et le comportement de certains dispositifs à mémoire résistive. C'est seulement en 2008 que des chercheurs de Hewlett Packard (80) ont relié le comportement d'un composant physique avec le modèle de Chua émis 37 ans plus tôt.

2. Un intérêt grandissant dans la communauté neuromorphique

L'arrivée de ces dispositifs dans le domaine de la micro et nanoélectronique en a fait un axe principal de recherche pour l'intégration de mémoires. L'intérêt majeur provient de leur forte densité et de leur potentiel à être miniaturisés. La rémanence de l'état du memristor en l'absence d'alimentation lui confère un attrait tout particulier dans des applications *instant off/instant on*. Enfin, l'intégration de ces dispositifs se faisant au niveau du *back-end*, les contraintes chimiques de fabrication sont moins restrictives qu'une technologie *front-end*. L'intégration de plusieurs couches permettrait également de répondre aux contraintes des mémoires actuelles, pouvant occuper une partie non négligeable d'un circuit.

L'attrait des industriels pour les memristors provient d'un besoin de réduction des structures de mémoires. Celles-ci comportent plusieurs transistors (6 pour une SRAM) ou encore une capacité (DRAM) qui ne permettaient plus la réduction efficace de leur surface. Les mémoires résistives sont, quant à elles, capables de stocker des données

binaires en utilisant seulement un transistor d'accès et présentent également un fort potentiel de stockage de valeurs multi-bits, multi-valuées ou analogiques. En effet, l'intégration d'une résistance de forte valeur est possible pour un coût en surface dérisoire. Sa valeur est modifiable et on peut ainsi concevoir, par exemple, des filtres denses et programmables.

Parallèlement au développement des composants à caractères memristifs, on assiste à une explosion de leurs applications potentielles, en particulier dans l'ingénierie neuromorphique. Un scénario classique de simulation consiste à implémenter des synapses dotées de règles d'apprentissages (STDP) (5, 44, 88).

Dans le contexte du neuromorphique, Leon Chua a exposé dès 1976 (17) l'intérêt d'un memristor pour remplacer les canaux ioniques d'un neurone basé sur le modèle Hodgkin-Huxley. En effet, les canaux ont été décrits comme des résistances dépendantes du temps et potentiellement remplaçables par des systèmes memristifs du premier ou du deuxième ordre. Un neurone basé sur ce modèle verrait son nombre de transistors décroitre et par conséquent sa surface se réduire.

Comme on l'a déjà évoqué, l'implémentation des synapses dans une architecture neuromorphique pose également un réel problème. Si on cherche à reproduire sur silicium le fan-out d'un neurone pour 10 000 synapses, leur surface devient prépondérante. L'implémentation de règles d'apprentissage est également complexe puisqu'elle nécessite un bloc dédié. La mise à jour des poids synaptiques deviendrait alors extrêmement couteuse en termes de surface ou de temps dans le cas où le bloc serait mutualisé. Par conséquent, ceci contraint certains groupes de recherche à privilégier l'apprentissage hors-puce.

En temps que mémoire analogique, un memristor peut directement faire la conversion entre le potentiel d'action numérique et le neurone. Le memristor connecté au potentiel de membrane à l'aide d'un transistor agit comme un filtre RC programmable que l'on assimilerait aux canaux ioniques. Il a également été observé que l'évolution de la résistance d'un memristor présente un comportement similaire à la règle d'apprentissage STDP (76), que nous avons vu sur la figure I.3. En effet, si les impulsions des neurones pré et postsynaptiques coïncident, une différence de potentiel aux bornes du memristor diminuera sa résistance et donc renforcera son poids.

Certains de ces dispositifs sont actuellement sujets à une très forte variabilité en termes de courant ou de temps de programmation. L'étude réalisée dans (63) a cependant montré que malgré une variabilité de 50 %, un réseau de neurones reste fonctionnel. Les memristors sont par conséquent de sérieux candidats pour des applications neuromorphiques et en particulier pour des implémentations synaptiques.

3. De nombreux phénomènes physiques

La littérature recense de nombreux phénomènes physiques qui présentent un comportement memristif. Comme on le voit sur la figure III.4, Waser et al. ont proposé de les classer en différentes catégories selon les mécanismes physiques qui régissent leur fonctionnement.

- *Phase Change Memory* (PCM) : le matériau chalcogénure qui constitue cette mémoire se fige en phase cristalline ou amorphe selon la durée pendant laquelle il est soumis à une température. Elle possède par conséquent une résistance modifiable, fonction du ratio des phases aux résistivités distinctes.
- *Thermal Chemical Memory* : le basculement résistif est assuré par la dissolution d'un filament conducteur dictée par l'effet Joule ou la formation de celui-ci par un champ électrique appliqué.
- *Valency Change Memory* : l'effet d'un champ électrique permet la modification de valence du matériau induisant un changement de ses propriétés résistives.
- *Electrochemical Metallization Memory* : une réaction d'oxydoréduction permet la dissolution d'une électrode au sein d'un matériau perméable. Sous l'effet du champs électrique, la migration des ions permet la formation et la dissolution d'un filament conducteur.
- *Electrostatic/Electronic Effects Memory* : le changement de la résistance est obtenu par modification de la barrière électronique à l'interface du dispositif. Cette modulation est causée par plusieurs phénomènes différents comme l'utilisation d'un dipôle ferroélectrique ou le piégeage de porteurs aux interfaces.

Initialement optimisée pour du stockage de données optiques numériques, la technologie PCM est actuellement la plus maitrisée. Plus généralement, ces mémoires ont chacunes des atouts et des inconvénients qui les destinent à être utilisées dans des cadres d'utilisations spécifiques. Il existe évidemment des considérations en termes de

consommation et de surface mais on détaillera ici les avantages qui leurs permettraient de remplir le rôle de synapses.

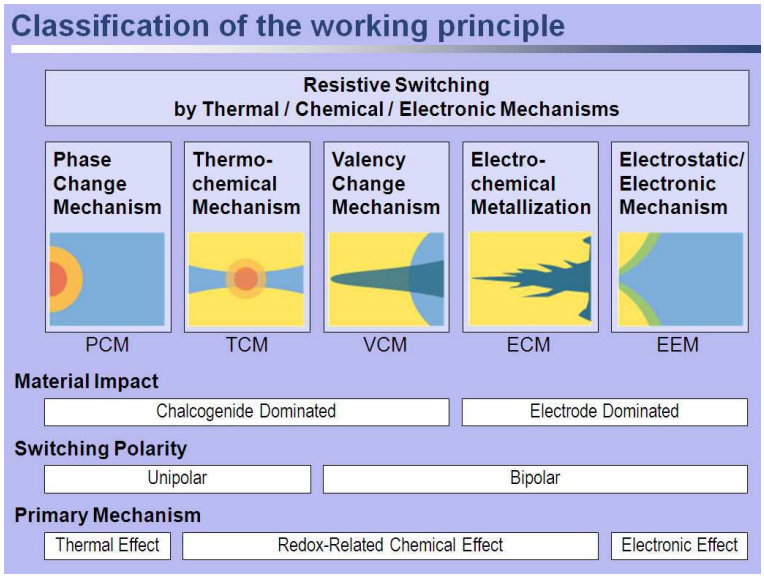


Figure III.4: Une classification possible des dispositifs de mémoires résistives selon les phénomènes physiques mis en jeu. Waser et al.

Une mémoire bipolaire sera plus facile d'emploi en vue d'une implémentation synaptique puisqu'elle pourra à la fois prendre les fonctions de LTP et de LTD. Le rapport Ron/Roff est également un critère de choix puisqu'il permet d'étendre la dynamique de la résistivité d'une synapse. Selon le rôle adopté, elle pourra alors avoir un impact très faible ou relativement important, augmentant fortement le potentiel de membrane. Une large dynamique est nécessaire mais pas suffisante. Les synapses biologiques sont, en effet, capables de prendre tout un panel de valeurs intermédiaires. Cet éventail est particulièrement requis pour des applications d'apprentissage, dans lesquelles les mises à jours des poids synaptiques se font par leurs modifications successives.

4. Précisions sur les mémoires résistives utilisées

Nous avons étudié l'implémentation de deux mémoires résistives couplées avec des neurones électroniques. Les PCMs seront intégrées au sein d'un neurone ou comme synapses multivaluées de neurones dont on détaillera la topologie. Les CBRAMs appartiennent à la classe des ECM et seront employées dans un rôle de synapses binaires.

a) Détails de fonctionnement d'une PCM

Le schéma de la figure III.5 montre les composants physique d'une PCM. Elle se compose comme toute mémoire d'une *bit line* et d'une *word line* nécessaires à son écriture et à sa lecture. Un matériau chalcogénure (GST) est placé à l'intersection de ses deux lignes. Il possède deux valeurs de résistances distinctes, noté R_a et R_c, selon sa phase amorphe ou cristalline. De la fonte de la couche de GST au sein de cette structure, résulte un changement de phase fonction de la vitesse de son refroidissement. Une paroi isolante entoure un fil qui permettra de concentrer l'échauffement du matériau chalcogénure.

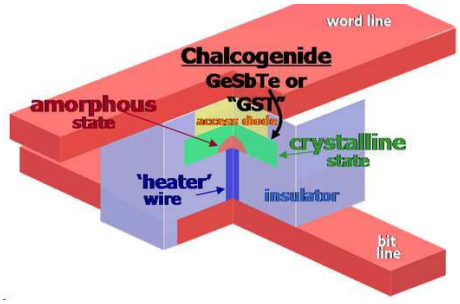


Figure III.5: Représentation des différents composants d'une PCM. Figures d'après (31).

La simulation de l'évolution de la température au sein d'un dispositif est montrée sur la figure III.6. La température de l'ordre de 700°C entraine la fonte du matériau chalcogénure. Selon sa vitesse de refroidissement, sa structure devient cristalline, amorphe ou un mélange des deux. Il est alors possible d'avoir tout un panel de valeurs de résistances intermédiaires entre R_a et R_c selon le ratio des phases du matériau.

b) Détails de fonctionnement d'une CBRAM

Une séquence de fonctionnement d'une CBRAM est présentée sur la figure III.7. Elle présente les différentes phases d'écriture, réalisée par la modification du filament par migration d'ions sous l'effet du champ V_{prog}. Lorsque celui-ci entre en contact avec l'électrode opposée, la résistance du dipôle est fortement diminuée. La lecture se fait alors à faible tension afin d'éviter l'écriture de la cellule. Selon la valeur de la résistance, la mémoire contient alors la valeur 1 ou 0. L'effacement s'effectue par l'inversion de la

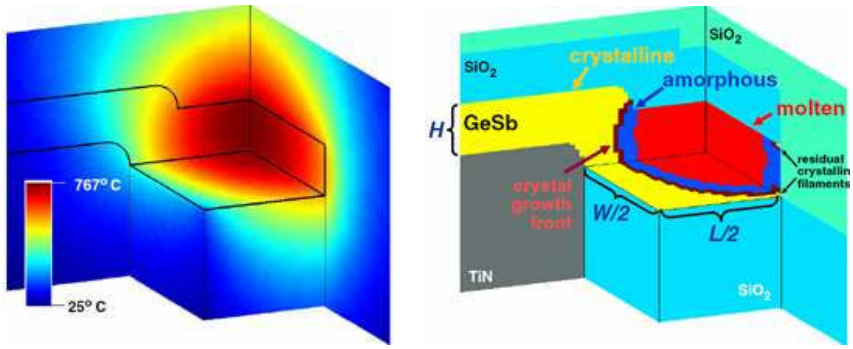


Figure III.6: Fonctionnement thermique d'une PCM. Figures d'après (31).

polarité du champ appliqué sur les électrodes. Il entraine un mouvement inverse des ions qui parvient à rompre le filament.

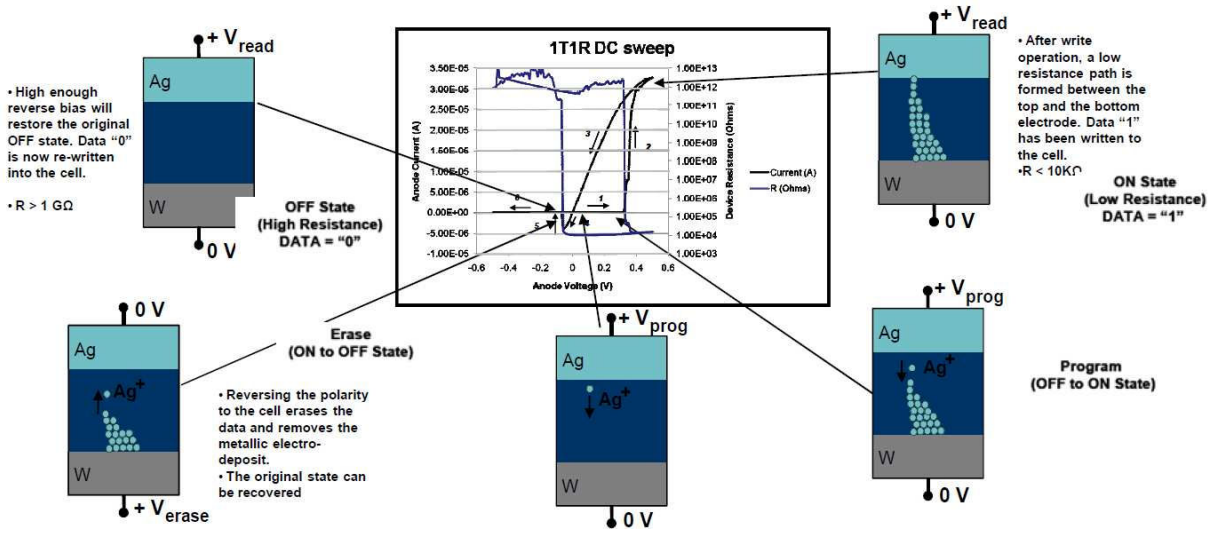


Figure III.7: Séquences de fonctionnement d'une CBRAM lors d'un cycle : programmation, lecture, effacement. D'après (7).

C. Cœur du neurone : élément capacitif

Un neurone contient différents blocs qui, selon le modèle et l'architecture utilisés, occupent une part non négligeable de la surface ou de la consommation. Si on a vu dans la structure élaborée au cours du chapitre précédent que le neurone LIF était compact,

sa taille était principalement limitée par sa capacité. La raison sous-jacente est qu'elle doit pouvoir stocker les charges pendant un temps déterminé pour les applications désirées. La capacité devenant la principale contrainte de surface, la dimension des transistors située au niveau du *front-end* a été adaptée pour diminuer la variabilité des neurones.

Dans les circuits *Reptile* et *Spider*, la capacité MIM répondait aux critères de surface et de rétention. Le passage à un nœud technologique plus avancé augmentant les fuites, il est probable que la capacité MIM ne réponde plus à ces critères. Nous aborderons ici les potentialités des technologies émergentes permettant de résoudre ces problèmes de densité et/ou de rétention.

1. État de l'art des capacités

On distingue trois approches qui permettent l'augmentation de la densité des capacités. La première consiste en la réduction de l'épaisseur d'oxyde entre les électrodes. Cette approche a été longtemps suivie mais les faibles dimensions en présence augmentent considérablement les fuites intrinsèques alors causées par effet tunnel. La rétention, pouvant être primordiale suivant l'optique suivie lors de la conception du neurone, s'en trouve fortement diminuée. La deuxième approche emploie des matériaux à forte permittivité semblables à ceux déjà utilisés dans la capacité MIM du neurone analogique conçu. La dernière option consiste à augmenter la surface des électrodes en regards. Afin de limiter leur empreinte sur un circuit, ces solutions utilisent des topologies en trois dimensions. Le tableau III.2 présente l'état de l'art en termes d'intégration de capacités.

Afin de réduire la surface du neurone, seules des capacités intégrées au niveau du back-end sont présentées dans le tableau. Elles permettront, outre de libérer une surface de silicium pour les transistors, d'être compatibles avec leurs fabrications.

2. Les capafils

Au delà des capacités indiquées dans le tableau précédent, nous nous intéressons au potentiel d'utilisation des capacités à base de nanofils développées par P.H. Morel (57). L'utilisation de nanofils permettrait d'accroitre considérablement la surface de regard entre les électrodes. Après croissance des nanofils conducteurs à partir d'un substrat, un dépôt d'isolant puis un dépôt de matériau conducteur sont réalisés. La capafil, pour

Tech. de structuration	Diélectrique (ϵ_r)	Densité ($\mu F/cm^2$)	Référence
Nanotubes de Carbone	HfO_2	0.65	(15)
MIM 3D	Ta_2O_5 (23)	1	(20)
MIM 3D	Ta_2O_5 (25)	1.5	(81)
MIM Plan + TSV	$SrTiO_3$ (140)	2.5	(73)
MIM 3D	Ta_2O_5 (25)	3	(39)
MIM 3D	Ta_2O_5 (25)	4.5	(3)
MIM	ST high-k	0.5	
MOS	ST transistor stack	1.2	

Table III.2: Tableau présentant les différentes caractéristiques de capacités intégrées au niveau du *back-end*. D'après (57).

capacité à base de nanofils, ainsi formée est compatible avec une intégration back-end et possède une faible empreinte. Les ordres de grandeurs de ces travaux ont démontré une capafil de type MOS de 9,6 μF/cm^2 pour une fuite intrinsèque proche de 110 aA/μm^2.

Une configuration envisagée est présentée sur la figure III.8. L'utilisation de mémoires résistives au niveau du *back-end* supprimerait les transistors d'injection et libèrerait de l'espace de silicium. Les transistors restants proviendraient principalement de l'intégration du comparateur. La contrainte de surface pourrait alors se situer au niveau de l'implémentation de la capacité. On privilégie par conséquent l'utilisation de capacité à base de nanofils dont la tension est comparée à l'aide de transistors. L'emploi de ce type de capacité permettrait une réduction d'un facteur 20 de la surface par rapport à la capacité MIM. Ceci causerait un impact majeur lors d'une implémentation analogique du neurone en lui procurant un avantage supplémentaire en termes de surface.

Ce schéma présente la stimulation d'un neurone <0> par une impulsion provenant du neurone <1>. Le potentiel traverse une structure de type crossbar dont les résistances déterminent les interconnexions des neurones. La tension aux bornes de la capacité du neurone <0> augmente et permet la génération d'un potentiel d'action lorsqu'elle dépasse une valeur de seuil. Il pourra être préalablement mis en forme et amplifié pour faciliter sa propagation vers les prochains neurones.

Nous avons vu ici les impacts d'une technologie émergente à forte valeur ajoutée

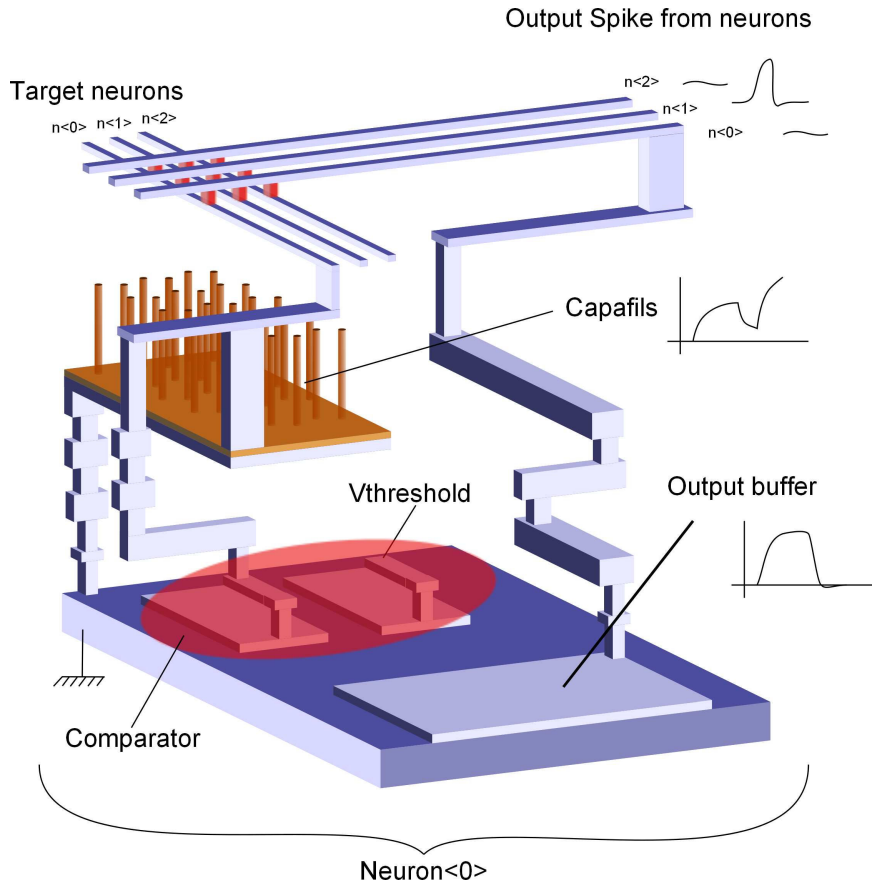


Figure III.8: Concept d'implémentation d'un futur neurone analogique. Les synapses sont réalisées par des dispositifs résistifs au niveau du back-end (en rouge). La taille de la capacité deviendra alors une contrainte forte pouvant être réduite par l'intégration de dispositifs à nanofils. Seuls les transistors de comparaison et d'amplification restent sur silicium.

d'un point de vue analogique. Des dispositifs novateurs sont également développés pour des utilisations mixtes ou numériques et feront l'objet des prochains paragraphes.

3. Utilisation d'un *Through Silicon Via*

Les *Through-Silicon-Vias* (TSVs) sont des candidats potentiels à l'empilement de puces électroniques (18, 64, 83). Ils répondent aux besoins de bande passante entre puces et permettent de réduire la consommation. Grâce à eux, il est également possible de choisir une technologie appropriée aux besoins de chaque partie d'un système. Celles-ci peuvent en effet être optimisées pour des applications mémoire, analogique, numérique

haute et basse performance. Le TSV est généralement utilisé comme un moyen de communication entre puces.

a) Positionnement de l'étude

On a vu dans le chapitre précédent que la capacité de membrane occupait une surface aussi grande que le neurone lui-même. Comme présenté sur la figure III.9, le TSV est constitué d'un métal isolé du substrat. Pour des applications courantes, cette capacité est parasite puisqu'elle augmente le temps de propagation du signal d'une puce à l'autre, en agissant comme un filtre passe-bas. La tendance actuelle est donc de réduire cette capacité pour augmenter les performances. Comme la surface occupée par un TSV est de l'ordre d'un neurone, nous nous proposons ici de l'exploiter pour en faire l'élément capacitif du neurone.

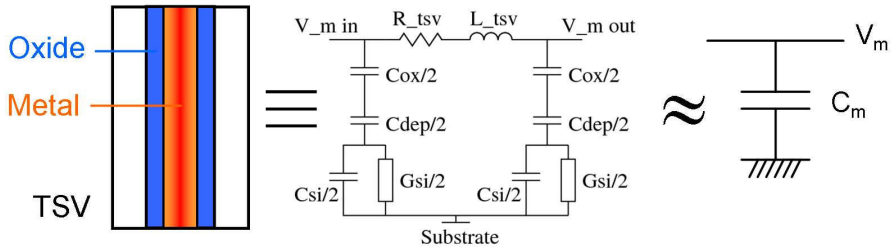


Figure III.9: Structure du TSV et modèle équivalent employé.

Paramètres du modèle	Faible densité (Low)	Moyenne densité (Medium)	Haute densité (High)
R_{tsv} $(m\Omega)$	30	20	100
L_{tsv} (pH)	50	25	5
C_{ox} (fF)	700	545	40
C_{Si} (fF)	160	55	6
G_{Si} (mS)	5	1.5	0.5
C_{dep} (fF)	2070	370	25

Table III.3: Caractéristiques mesurées d'un TSV en fonction de leur densité.

b) Modèle utilisé

Le TSV n'est pas une capacité idéale en comparaison à la capacité de type MIM utilisée lors de la conception du neurone. Il n'est pas non plus un conducteur parfait mais peut se modéliser par une structure RLC en Π (10) comme montré au centre de la figure III.9. Le modèle de TSV comporte une résistance R_{tsv} ainsi qu'une inductance L_{tsv}. Il comporte également trois capacités C_{ox}, C_{dep} et C_{si} qui isolent le métal du substrat. Les fuites parasites dans le substrat sont représentées par la conductance G_{si}. Toutes ces valeurs varient évidemment selon le procédé de fabrication du TSV mais également selon leur densité, leur hauteur, leur diamètre, leur fréquence d'opération et leur épaisseur d'oxyde. Ainsi, le tableau III.3 précise les caractéristiques électriques d'un TSV selon leur densité.

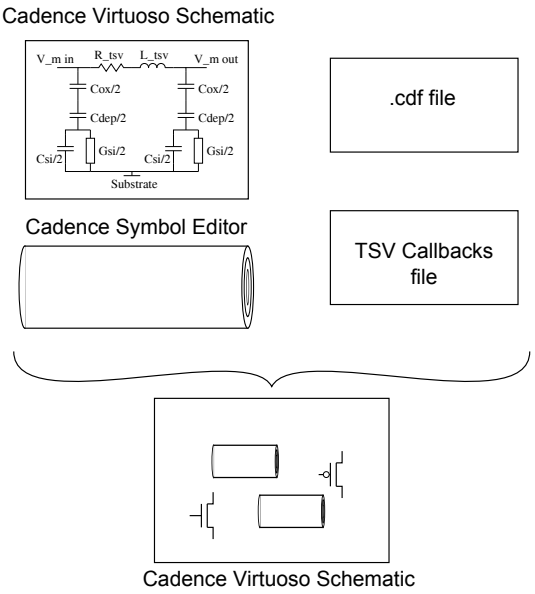


Figure III.10: Intégration du TSV dans l'environnement de simulation Cadence.

Ce modèle de TSV a été implémenté pour être utilisable dans l'environnement de simulation. Comme indiqué sur la figure III.10, la création du schéma RLC avec des paramètres génériques définisse un symbole TSV. Lors de la création d'une instance TSV les paramètres sont spécifiés à l'aide de deux fichiers annexes. Les différentes valeurs sont alors être prédéfinies, elles seront alors fonction de la densité, ou encore spécifiées manuellement.

c) Résultats de simulations

Nous avons utilisé de nouveau le banc d'essai utilisé en partie II-B. pour valider le fonctionnement du neurone. Sa capacité MIM est ici remplacée par un TSV dont les différentes valeurs sont basées sur les travaux de (10).

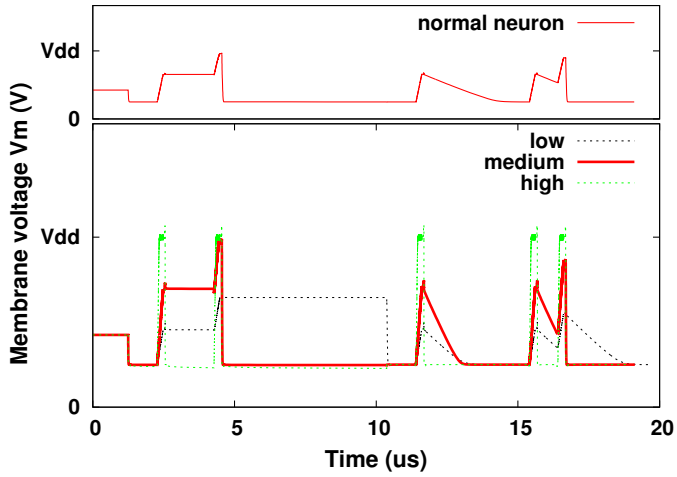


Figure III.11: Comparaison de l'évolution du potentiel de membrane d'un neurone standard et d'un neurone réalisé avec un TSV (pour différentes densités de TSV).

La valeur de la capacité totale du TSV varie selon sa topologie et en fonction de leur densité, c'est à dire le nombre de TSVs par unité de surface. Cependant, ceci a pour conséquence de modifier seulement les constantes de temps du neurone. On observe en bas de la figure III.11 que l'on peut donc identifier une courbe pour laquelle le comportement du neurone est très similaire à une version standard à base de capacité (III.11 en haut).

d) Opportunités d'une architecture 3D

Dans la figure III.12, nous présentons une vue de l'implémentation du neurone, d'abord dans un cas classique puis dans différentes propositions de topologies basées sur des TSVs. L'objectif est d'explorer les bénéfices potentiels pour une architecture neuromorphique lorsque les neurones communiquent de proches en proches.

Les gains se comptent en termes de surface de silicium, de surface de masque et de connectivité. Les résultats sont présentés dans le tableau III.4. On s'aperçoit que l'ajout

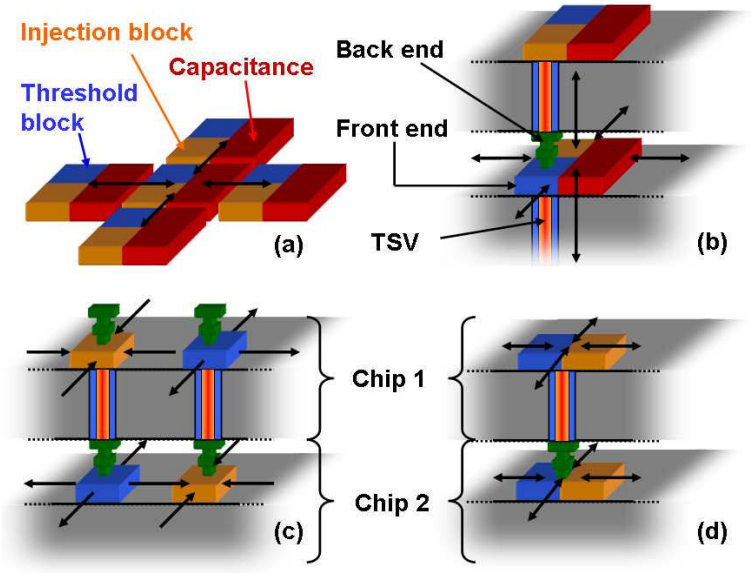


Figure III.12: Illustrations d'une architecture neuromorphique basé sur un circuit intégré (a) 2-D (b) 3D avec des neurones 2D ou 3D avec des neurones-TSV (c) et (d)

Topologie	(a)	(b)	(c)	(d)
Surface silicium	A	A	A/2	A
Surface sur masque	A	A	A/4	A/2
Connectivité	4 IOs	6 IOs	4 Is & 4 Os	8 IOs

Table III.4: Opportunités et gains potentiels de l'utilisation de TSVs pour une architecture neuromorphique

du TSV d'une manière classique permet d'augmenter l'interconnectivité des neurones. L'utilisation des TSV dans les neurones permettent d'améliorer soit la connectivité soit la surface. On peut envisager d'employer conjointement ces différentes topologies selon les besoins de l'architecture.

4. Utilisation d'une mémoire à changement de phase

Comme exposé plus tôt, les PCMs sont des memristors formés d'un matériau à changement de phase dont la résistance est modifiée suivant s'il est à structure cristalline ou amorphe. La transition d'un état à l'autre se fait par effet Joule et permet, selon les conditions d'échauffement, d'atteindre un niveau de résistance haut ou bas.

a) Positionnement de l'étude

La réduction de la taille de la grille des transistors permet l'augmentation de la capacité surfacique mais augmente parallèlement les fuites intrinsèques des composants. La rétention du potentiel de membrane pourrait devenir problématique dans un neurone ayant des constantes de temps de l'ordre de la seconde. On propose d'étudier l'emploi d'une PCM en substitution à la capacité du neurone. Elle a plusieurs atouts pour remplir ce rôle à savoir sa compacité, sa rémanence et enfin ses capacités pour du stockage multi-valué.

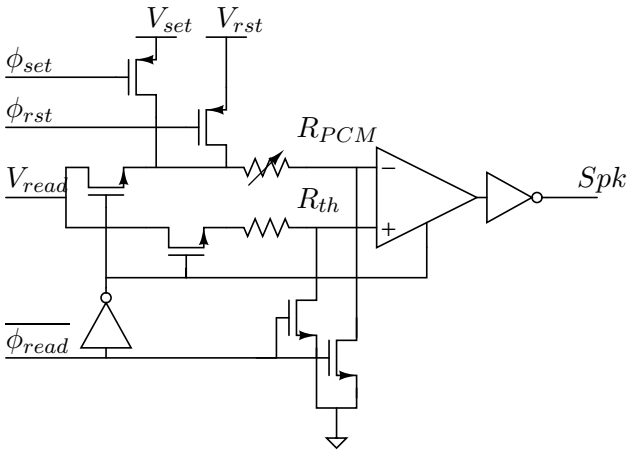


Figure III.13: Schéma du neurone IF à base de PCM : la capacité C_m est remplacée par la résistance R_{PCM}.

b) Implémentation

Dans une première approche, nous avions choisi de présenter un neurone de type LIF. En l'état actuel des caractérisations et du modèle de la PCM, nous avons préféré privilégier la piste d'un neurone IF. Le schéma du neurone est présenté sur la figure III.13 et comprend un comparateur, une PCM, une résistance de référence et des transistors de contrôle. La topologie du comparateur est la même que celle utilisée pour les neurones des circuits *Reptile* et *Spider*. Les transistors reliés à V_{set} et V_{rst} sont compatibles à des tensions de l'ordre de 2.5 V. La résistance de référence peut être soit une PCM soit une résistance polysilicium mutualisée entre plusieurs neurones.

c) Protocole de fonctionnement et résultats de simulations

Les poids synaptiques sont encodés selon un créneau de longueur τ_{set} et d'amplitude V_{set} qui diminue la résistance de la PCM. Le signal $\overline{\phi_{read}} = Vdd$ maintient le potentiel de l'entrée négative du comparateur à bascule, noté V_{PCM}, à 0 V. Le transistor, activé par le signal ϕ_{set}, permet d'appliquer la différence de potentiel de V_{set} nécessaire à la modification de la résistance de la PCM. Après écriture, une impulsion d'amplitude V_{read} et commandée par ϕ_{read}, est envoyée dans le circuit RC composé par les deux résistances R_{PCM} et R_{th}. Les capacités utilisées sont celles de la paire différentielle constituant l'étage d'entrée du comparateur. L'évolution du potentiel est fonction de la valeur de la résistance. Sous réserve que $R_{PCM} < R_{th}$, la sortie Spk passe à 1 lors de la comparaison également commandée par le signal ϕ_{read}. Dans ce cas le potentiel d'action est envoyé vers un mécanisme de routage numérique qui remet la PCM à son état de résistance haute via un protocole "poignée de main". La partie de contrôle activera pour cela le signal ϕ_{rst} qui forcera une tension V_{rst} aux bornes de la PCM pendant un temps τ_{rst}.

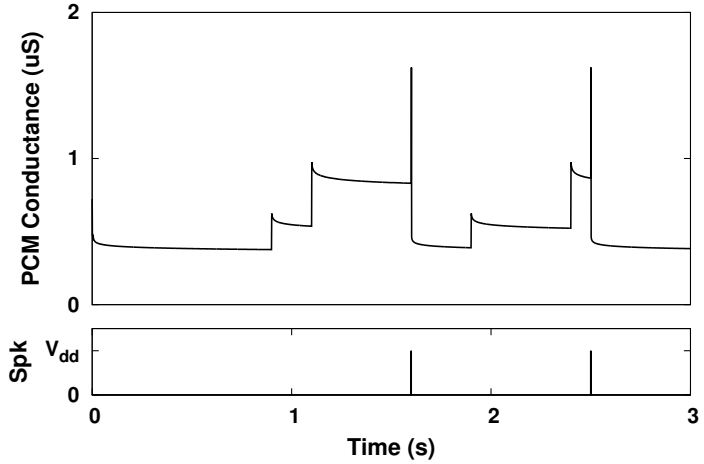


Figure III.14: Comportement du neurone IF basé sur un composant PCM.

La figure III.14 montre les résultats de simulations. On a utilisé pour ceci les valeurs indiquées dans le tableau III.5.

d) Conclusion

Cette étude a montré les possibilités offertes et résumées dans le tableau III.6 par l'utilisation d'une PCM dans un neurone. Elle permettrait la diminution significative de

Paramètres	Valeurs
R_{off}	2.2 $M\Omega$
R_{on}	10 $k\Omega$
R_{th}	1.2 $M\Omega$
τ_{set}	25 ns
τ_{rst}	200 ns
τ_{read}	250 ns
V_{set}	1.5 V
V_{rst}	3 V
V_{read}	100 mV

Table III.5: Paramètres de simulation

la surface du neurone. Elle passerait outre les problèmes de rétention dans une capacité et pourrait permettre l'implémentation de neurone IF. Les PCMs sont également de très bonnes candidates pour des implémentations synaptiques (44). Par conséquent, elles n'ajouteraient pas d'étape supplémentaire lors de la fabrication.

Cœur du neurone	Capacité	PCM
Surface	$550 + 100$	$0.06 + 100$
Rétention	1 s	> 10 ans
Résolution	7 $bits$	2 $bits$
Consommation	☺	☹
Fiabilité	☺	☹

Table III.6: Comparaison du moyen de stockage d'un neurone : une capacité face à une PCM. Deux points faibles des PCMs sont toujours en optimisation.

Il existe un phénomène de dérive de la résistance des PCMs, appelé *drift* (32), causé par la relaxation du chalcogénide. Une caractérisation plus poussée ainsi qu'un remaniement du modèle pourrait permettre l'utilisation de cet effet pour apporter la composante de fuite au neurone. Sur la figure III.14, le comportement de *drift* est visible mais une mise à jour de la valeur de la PCM supprime totalement son effet.

D. Implémentation des connexions synaptiques

Nous nous intéressons ici à l'implémentation des connexions inter-neurones qui est, comme énoncé au début de ce chapitre, un des freins au développement du bio-inspiré et encore plus du bio-mimétique. Nous avions adopté la stratégie, pour nos architectures *Reptile* et *Spider*, d'un DAC mutualisable pour plusieurs neurones. Ceci reste toutefois couteux en termes de surface, de consommation et requiert de surcroît une horloge.

L'utilisation de memristors comme synapses peut être considérée sous deux angles. Le premier tente de se rapprocher de la biologie en montrant la corrélation entre le comportement d'un memristor et les modèles de règles d'apprentissages de synapses. Le deuxième privilégie l'aspect fonctionnel de conversion entre un potentiel d'action numérique et un potentiel de membrane analogique.

1. Utilisation des mémoires à changement de phase

Contrairement à la partie précédente où la PCM était utilisée au sein du neurone, nous l'avons employé ici dans l'objectif de réaliser une connexion synaptique. Des simulations architecturales (5) ont montré la capacité de ces mémoires résistives à pouvoir apprendre de manière non supervisée. Pour permettre cet apprentissage, il est nécessaire d'augmenter le poids synaptique lorsque celles-ci permettent la génération d'un potentiel d'action.

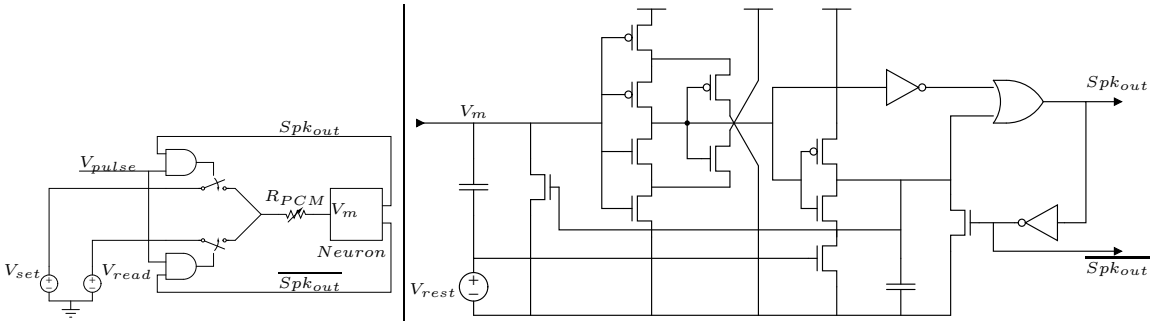


Figure III.15: Schéma de test du neurone couplé à une PCM et son propre schéma. Il permet la diminution de la résistance d'une PCM lorsqu'il y a corrélation entre les impulsions présynaptique et postynaptique.

Nous montrons sur la figure III.15 le schéma du neurone ainsi que le test permet-

tant de mettre en évidence l'évolution de résistance de la PCM. Le neurone est composé d'une capacité d'entrée chargée au potentiel de membrane V_m. Lorsqu'il dépasse un certain seuil, déterminé par le dimensionnement des transistors d'une bascule de Schmitt, il génère une impulsion sur la sortie Spk_{out}. Sa largeur, déterminée par la taille d'une deuxième capacité, permet l'écriture de la PCM s'il existe une corrélation avec une impulsion entrante Spk_{in}. L'écriture de la PCM est réalisée à l'aide d'un potentiel V_{set}. Si il n'y pas de corrélation, l'augmentation du potentiel V_m se fait par le potentiel V_{read}.

Nous présentons les résultats de la simulations sur la figure III.16. Le potentiel d'action présynatique, V_{pulse}, est périodique permettant de faciliter la simulation. La résistance de la PCM étant initialement forte, le potentiel de membrane V_m augmente lentement. Le neurone requiert initialement un grand nombre de stimulations avant de générer une impulsion. Lorsqu'il a atteint son seuil, Spk_{out} est corrélé avec V_{pulse} par une porte AND pour obtenir le signal $Update$. Il active alors l'interrupteur permettant l'établissement du potentiel V_{set} aux bornes de la PCM.

Après mise à jour, le nombre d'impulsions requis pour la génération d'un potentiel d'action est fortement réduit avant de devenir constant. Visible sur le tracé de R_{PCM}, les deux premières mises à jour permettent une réduction de la résistance avant qu'elle n'atteigne sa valeur R_{on}.

Cette implémentation propose une mise à jour des poids synaptiques lorsqu'ils sont codés par la résistance d'une PCM. On a montré ici la faisabilité de l'implémentation de la LTP, c'est à dire le renforcement des poids positifs, pour une PCM. Des questions restent toutefois en suspens, à savoir le passage à l'échelle et l'implémentation de la LTD.

2. Utilisation des mémoires à filament conducteur

Nous nous intéressons ici à la conception des synapses à l'aide de composants CBRAMS (pour Conducting Bridge RAM). Un démonstrateur en technologie 130 nm utilisant ces dispositifs a été conçu puis envoyé en fabrication. Ce sera un des premiers exemples de co-intégration memristors-cmos pour une architecture neuromorphique.

a) État des lieux du composant CBRAM

Ces mémoires appartiennent à la famille des ECM. Introduites au niveau du back-end, les CBRAMs sont plus souples en termes de choix du procédé de fabrication. Le

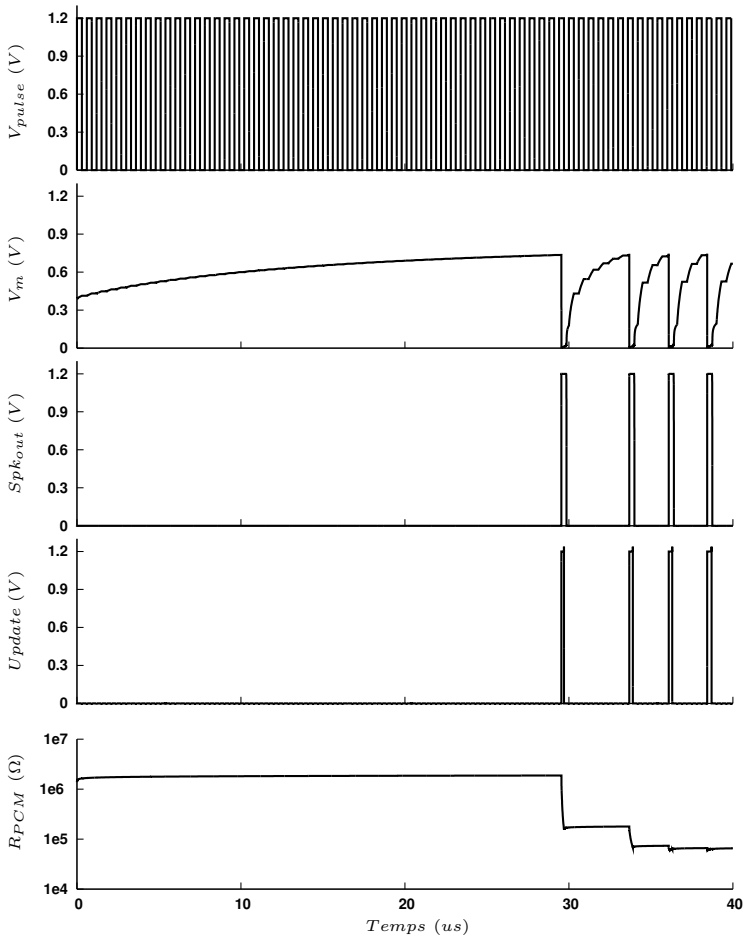


Figure III.16: Chronogramme d'une mise à jour de la valeur de la résistance de la PCM lorsqu'il y a corrélation entre une impulsion entrante et une sortante.

matériau à utiliser dans ces composants est toujours à l'étude. Lors de la conception, l'emploi des CBRAMS dans le démonstrateur s'effectue de manière à être compatible aux deux compositions chimiques potentielles. Les valeurs de R_{off} et R_{on} sont indiquées dans le tableau III.7. Il existe en effet un compromis à trouver entre le rapport R_{off}/R_{on}, les temps et les tensions de programmation et de lecture.

La première option à base de GeS_2 a l'avantage de présenter un rapport R_{off}/R_{on} élevé. Elle autorise une différence notable entre un poids synaptique fort et un faible, ce qui en fait une technologie très encline à une implémentation neuromorphique. La deuxième technologie semble plus prometteuse en termes de temps de programmation.

	GeS_2	Gd_20_3
R_{off}	20 $M\Omega$	2 $M\Omega$
R_{on}	5 $k\Omega$	50 $k\Omega$

Table III.7: Caractéristiques des CBRAMS prises en compte pour la conception

Elle serait plutôt à privilégier dans le cas d'une mémoire numérique à haute densité.

Le modèle comportemental utilisé dans les simulations correspond aux caractérisations effectuées sur des mémoires de types GeS_2 (61). Ce modèle est écrit en Verilog A et permet une première approche de conception neuromorphique à base de mémoires résistives. Les temps, les tensions et les résistances moyennes (off/on) sont connus et inclus dans le modèle. Les limites actuelles du modèle résident dans la faible connaissance du dispositif en réponse à des stimuli courts ($<$ 500 ns) et la non prise en compte des variations du composant. Il a en effet été observé de grandes disparités en termes de valeur de résistance et de temps de basculement.

Des études (63) ont montré la robustesse de systèmes neuromorphiques grâce notamment à la STDP. Nous avons cependant privilégié un moyen d'adressage apportant maitrise de la programmation et opportunité de caractérisation. Selon la colonne et la ligne sélectionnées, un bloc de portes logiques accède à une unique CBRAM rendue alors modifiable. Ce scénario est à mettre en rapport avec les mémoires classiques pour lesquelles la sélection d'une ligne et d'une colonne permet de modifier directement l'état d'une cellule. Grâce au bloc *spk_sharper*, des caractérisations de la CBRAMs soumises à des impulsions de très courtes durée ($<$ 10 ns) seront possibles et pourront compléter les mesures déjà réalisées.

b) Description générale de l'architecture

L'architecture développée est présentée sur la figure III.17. Elle est composée de deux neurones et d'une matrice de six CBRAMs. Deux d'entre elles sont utilisées pour faire entrer une impulsion dans le circuit. Les quatre restantes forment une matrice de 2*2 pour connecter les neurones entre eux. D'autres blocs de contrôle ne sont pas présents sur cette figure et seront présentés par la suite.

Bien qu'un système neuromorphique soit conceptuellement asynchrone, l'architecture comporte une horloge. Elle apportera une souplesse additionnelle lors de la pro-

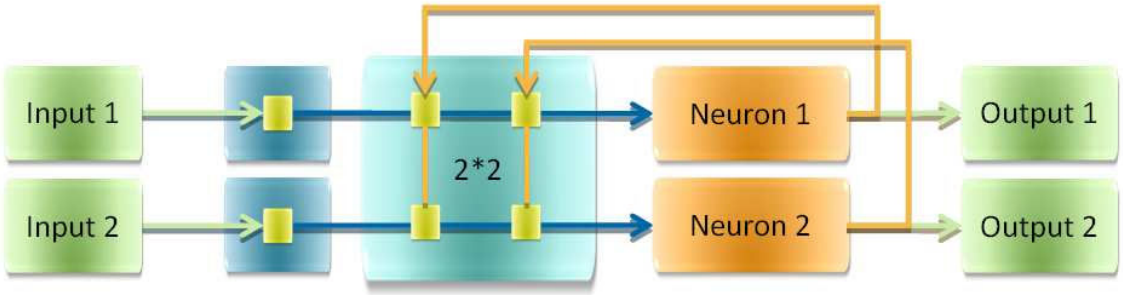


Figure III.17: Architecture implémentée pour évaluer le potentiel des CBRAMs. Deux neurones, une matrice 2*2 de CBRAM pour leur interconnexion et deux CBRAMs pour la stimulation des neurones de l'extérieur.

grammation des CBRAMs mais également un contrôle du temps de charge de la capacité de membrane des neurones. Elle a également permis de réduire le temps nécessaire à la conception des blocs responsables de l'émission du potentiel d'action et de la remise à zéro du potentiel de membrane grâce à l'utilisation de bascules numériques. Il est tout à fait envisageable de retirer cette horloge lorsque les CBRAMs auront des spécifications bien déterminées.

c) Conception et caractéristiques des différents blocs

Nous détaillons ici les différents blocs composant l'architecture.

Bloc *spk_sharper* : il a pour but de fournir un créneau long pour les phases de set/reset de la CBRAM et des impulsions brèves pour les phases de lecture. On distingue deux phases dans les architectures neuromorphiques : la phase d'apprentissage ou de programmation et la phase d'utilisation. Pour un créneau court inférieur à 500 ns, il est de plus en plus difficile de contrôler précisément la largeur de l'impulsion de l'extérieur du circuit. Les capacités et résistances parasites agissent comme un filtre passe-bas et peuvent devenir un frein à la transmission du signal. Les caractéristiques en termes de bande passante de la structure de test vers la puce n'étant pas clairement définies à l'heure de la conception, il semble judicieux de pouvoir contrôler les durées d'accès ou d'écriture de la CBRAM depuis l'intérieur du circuit.

La largeur d'impulsion générée est fonction de trois bits d'entrée selon le tableau III.8. Le signal, noté *spk_shp*, est amplifié puis envoyé vers les blocs de portes logiques

d'accès des CBRAM. Sa largeur maximale sera alors égale à la demi-période de l'horloge mais en opposition de phase avec celle-ci. La lecture d'une CBRAM sur un intervalle de quelques nanosecondes sera possible et permettra d'étudier, par exemple, l'influence de lecture à répétition en conservant un potentiel identique à celui de la programmation.

A⟨2⟩	A⟨1⟩	A⟨0⟩	Largeur d'impulsion (ns)
0	0	0	$1/(2 * f)$
0	0	1	229.5
0	1	0	105.6
0	1	1	54.2
1	0	0	19.1
1	0	1	9.3
1	1	0	4.6
1	1	1	2.7

Table III.8: Caractéristiques du générateur programmable d'impulsions : cas nominal.

Bloc *neuron_core* : il contient l'intégration, la comparaison du neurone et la remise à zéro du neurone. Il est constitué d'une capacité, d'un trigger de Schmitt et de 2 bascules D. La valeur de la capacité permettant l'intégration du potentiel de membrane du neurone vaut 5 pF. Cette valeur est déterminée en fonction de R_{on} et R_{off} des CBRAMs ainsi que des largeurs d'impulsions générées par le bloc *spk_sharper*.

Le réglage des tensions haute et basse de basculement du trigger de Schmitt est réalisé à l'aide de deux potentiels de polarisation V_pbias et V_nbias. Il est relié à deux bascules D dont les sorties permettent le retour de la capacité au potentiel de repos ainsi que l'envoi d'une impulsion dont la largeur est égale à la période de l'horloge. Lorsque l'on inverse V_{anode} et $V_{cathode}$ en vue de reprogrammer une CBRAM, les sorties de ces bascules sont forcées à 0 par un signal *inhib*, ce qui évite ainsi un retour involontaire des points mémoires à une résistance R_{off}.

Un bloc d'inhibition est constitué de trois inverseurs analogiques et d'une porte ET. Les inverseurs permettent de revenir dans le domaine de tension du numérique. Il permet de bloquer la génération d'impulsions en sortie des neurones lors de la remise à zéro des CBRAMs.

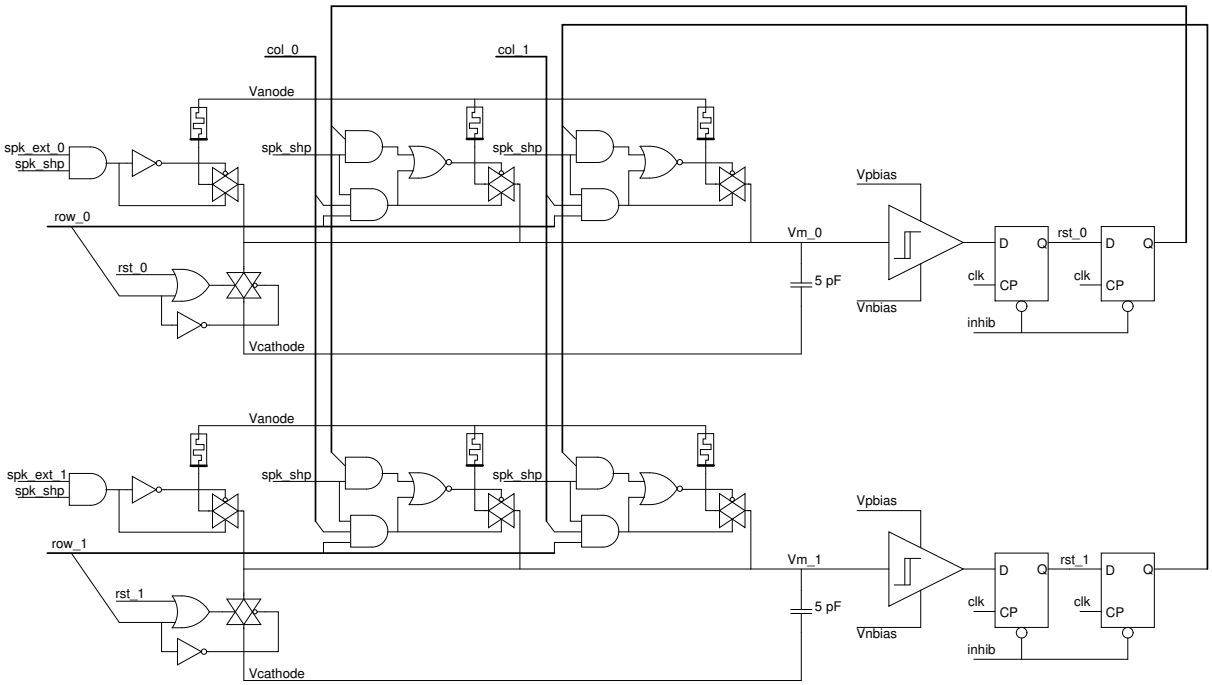


Figure III.18: Détails de l'architecture implémentée. On retrouve les six CBRAMs, et les deux neurones. Ces derniers sont constitués d'une capacité, d'un comparateur, de deux bascules et d'une remise à zéro.

d) Fonctionnement

L'architecture ainsi conçue est présentée sur la figure III.18 dont on présente ici le fonctionnement. La période de l'horloge est fixée à 120 us soit le double du temps nécessaire à la programmation d'une CBRAM. V_pbias et V_nbias sont choisis à 0.5 V et 1 V et correspondent à un déclenchement du neurone lorsque son potentiel de membrane atteint 1.2 V et une remise à zéro à 0.2 V. Les tensions V_{anode} et $V_{cathode}$ valent respectivement 1.5 V et 0 V lors de la programmation. Les différentes CBRAMS peuvent ensuite être programmées à l'aide des signaux spk_ext_i, col_i et row_j.

Avant la phase de stimulation du réseau, les bits de contrôle du bloc *spk_sharper* sont choisis de sorte que la largeur d'impulsion τ_{shp} respecte les conditions suivantes :

$$\tau_{shp} < \tau_{set} \tag{III.2}$$

$$R_{on} * C_m \sim \tau_{shp} \ll R_{off} * C_m \tag{III.3}$$

Ceci sous-entend que la différence de potentiel entre V_{anode} et $V_{cathode}$ est la même

que lors de la programmation et que l'on estime que la lecture dynamique de la CBRAM (temps de lecture $\ll$ temps d'écriture) n'influe pas sur la donnée stockée.

Si toutefois des problèmes d'écriture indésirable surviennent, la différence de potentiel entre V_{anode} et $V_{cathode}$ devra être diminuée. L'amplitude entre la tension seuil de déclenchement du neurone et son potentiel de repos sera réduit et nécessitera un ajustement des potentiels V_pbias et V_nbias.

e) Résultats de simulation

Les résultats de simulation de l'architecture présentés sur la figure III.18 sont visibles sur les chronogrammes de la figure III.19.

Dans un premier temps, celui de la programmation, le signal Spk_{shp} suit celui de l'horloge ($A < i > = \ 0$). La sélection de la ligne 0 par le signal row_0 permet le basculement, suivant une notation $(ligne, colonne)$, des CBRAMs $(0, ext)$ puis $(0, 1)$. Les signaux row_1 et col_0 activent quant à eux la CBRAM $(1, 0)$. La phase de programmation est alors terminée, $A < 0 >$ passe à 1 et diminue ainsi la largeur de l'impulsion.

L'oscillateur, créé entre le neurone 0 et 1 et stimulé à l'aide d'une première impulsion porté par le signal Ext_{spk_0}, est activé.

A 9 ms, les tensions d'anode et de cathode sont inversées. La CBRAM $(0, 1)$ est remise à sont état R_{off} l'oscillateur ne fonctionne dorénavant plus. Une impulsion stimule le neurone 0 puis 1 et le potentiel de membrane des neurones retourne à un potentiel flottant causé par les courants de fuites des composants.

f) Réalisation du layout

Le schéma proposé sur la figure III.18 et le bloc *spk_sharper* ont été dessinés pour fabrication. Le layout obtenu est présenté sur la figure III.20. L'emploi de cellules standards, disponibles dans les librairies proposées par le fabricant, a permis l'alignement des différents blocs du neurone (interrupteurs, bascules, etc) près de la capacité. Les différents blocs de contrôle du circuit se trouvent sur la partie gauche du circuit. Celle de droite présente un axe horizontal de symétrie, il s'agit des deux neurones et de leur capacité (en rouge). La surface qu'elles occupent met en évidence le besoin futur d'une capacité surfacique plus importante.

Il ne s'agit cependant que d'une implémentation de deux neurones et donc de quatre synapses d'interconnexions auxquelles s'ajoutent les deux synapses d'entrées. En effet,

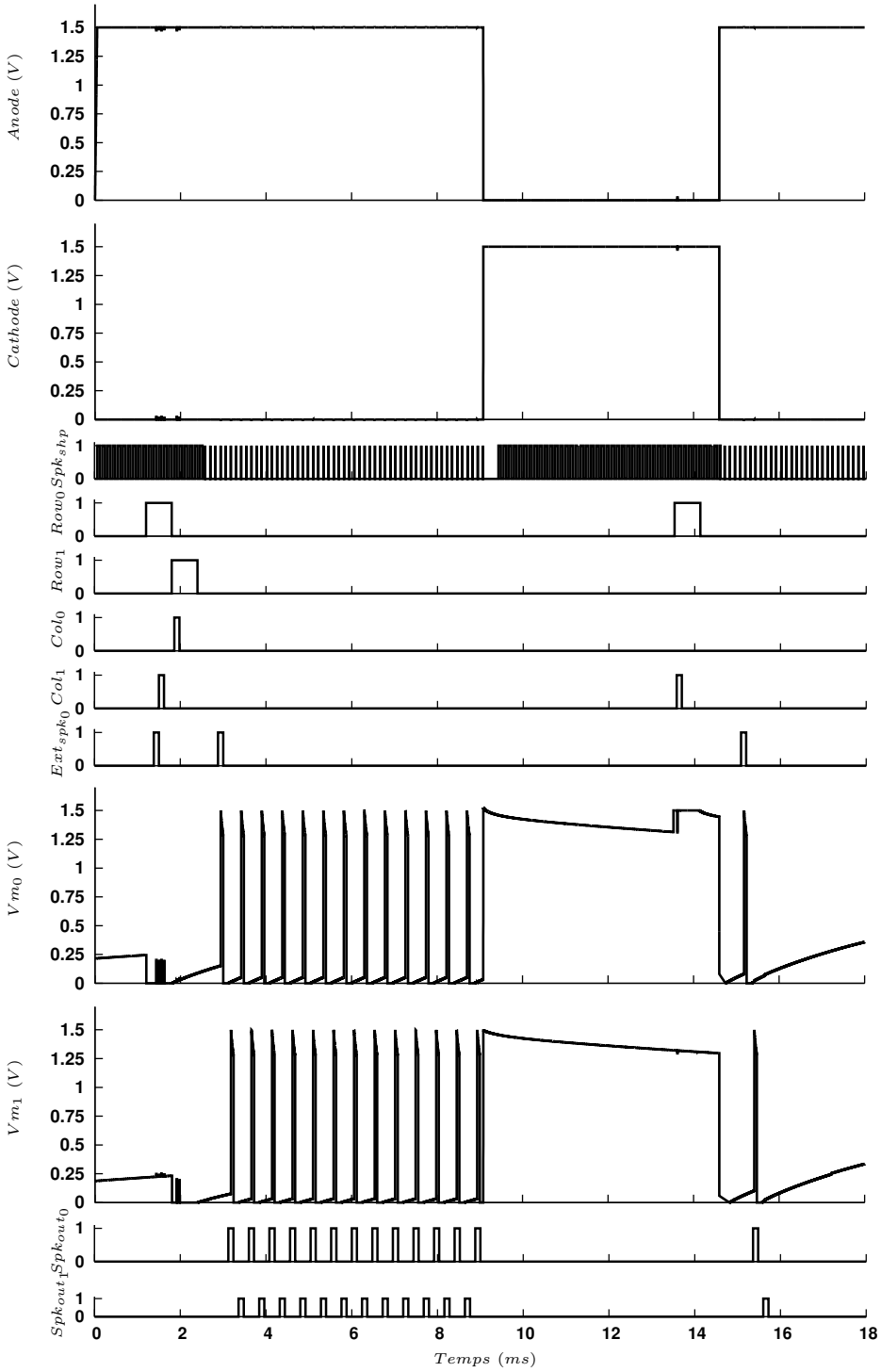


Figure III.19: Chronogramme des signaux de contrôles et de sortie des neurones : programmation des CBRAMS, fonctionnement en oscillateur, remise à zéro d'une CBRAM, transmission du PA du neurone 0 au neurone 1

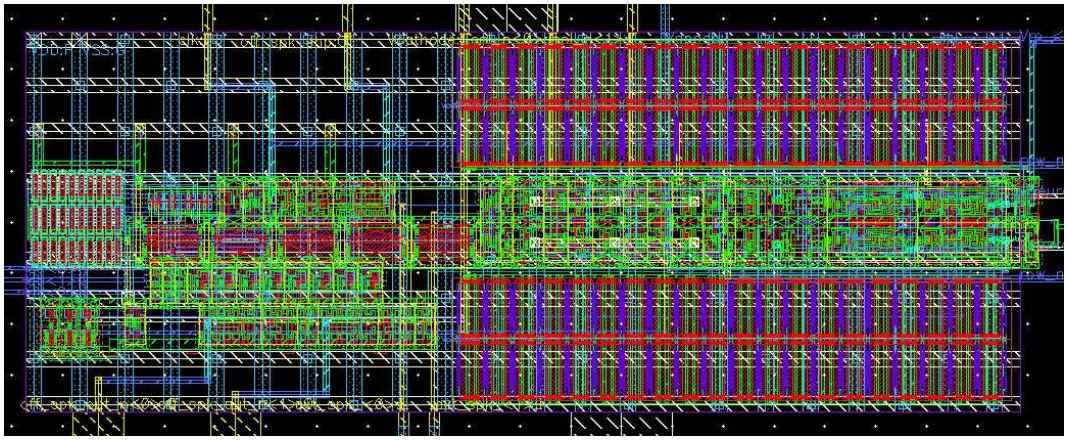


Figure III.20: Layout envoyé en fabrication pour des tests préliminaires. Il contient les deux neurones, les six CBRAMs d'interconnexions et de stimulation ainsi que le bloc *spk_sharper*.

à la vue de cette implémentation, le coût en surface de silicium entre les capacités et les synapses deviendrait équivalent lors d'une implémentation de 20 neurones. On retournerait alors à une limite d'intégration principalement causée par les synapses. Ceci provient de la partie de contrôle nécessaire à leurs commandes. Si elles étaient toutefois mutualisées ou encore empilées, l'intégration de la capacité serait de nouveau le centre du problème. Les futures intégrations d'architectures de mémoires résistives devraient permettre de cerner laquelle des deux sera le problème majeur des architectures neuromorphiques dans des nœuds très agressifs.

Ce circuit a été envoyé mais n'est pas encore revenu de fabrication et donc n'a pu, à ce jour, faire l'objet de tests. Il permettra de vérifier la programmabilité de CBRAMs et leur emploi potentiel dans une matrice d'interconnexion de neurones.

On a montré ici l'utilisation des CBRAMs pour une application de poids synaptiques binaires. Le comportement d'écriture, lecture et remise à zéro des connections a été validé en simulation. D'après les caractérisations effectuées, il sera possible d'utiliser les CBRAMs comme résistances multivaluées. Ce comportement n'était toutefois pas encore intégré dans le modèle du composant CBRAM au moment des simulations. Sur le démonstrateur réalisé, on pourra envisager d'adapter la tension d'alimentation lors de la phase de programmation.

Cœur du neurone				
Techno. / Carac.	MIM ST (Réf.)	Capafil	PCM	TSV
Surface	$\simeq$	++	++	+
Maturité	++	–	+	-
Fiabilité	++	-	$\simeq$	$\simeq$
Rétention	-	+	++	-
Résolution	++	++	-	+
Intérêts futurs	$\simeq$	++ (Gain de surface)	+ (Rétention)	+ (Connectivité 3D)

Synapse			
Techno. / Carac.	DAC (Réf.)	PCM	CBRAM
Surface	-	+	+
Maturité	++	$\simeq$	-
Apprentissage	-	++	+
Consommation	-	+	++
Multivalué	++	+	-
Intérêts futurs	-	+ (apprentissage)	+ (programmabilité)

Table III.9: Impact des technologies émergentes sur les futures architectures neuromorphiques. Comparaison avec la capacité MIM et la structure de DAC utilisée dans les circuits *Reptile* et *Spider*.

Conclusion

Dans cette partie, nous avons présenté plusieurs dispositifs émergents. L'étude de leur intégration ainsi que de leur impact au sein d'une architecture neuromorphique a été détaillée, poursuivant de ce fait, deux objectifs distincts. La première était d'optimiser le neurone lui-même en diminuant sa taille ou encore en améliorant sa connectivité. La seconde approche consistait en l'utilisation de mémoires résistives pour permettre l'interconnexion des neurones. Le but recherché alors, était de réduire la surface du circuit habituellement employée au stockage des poids synaptiques. Les apports des différentes technologies et leurs intérêts futurs sont résumés dans le tableau III.9.

Pour revenir sur l'avenir d'un neurone analogique, il est judicieux de s'intéresser

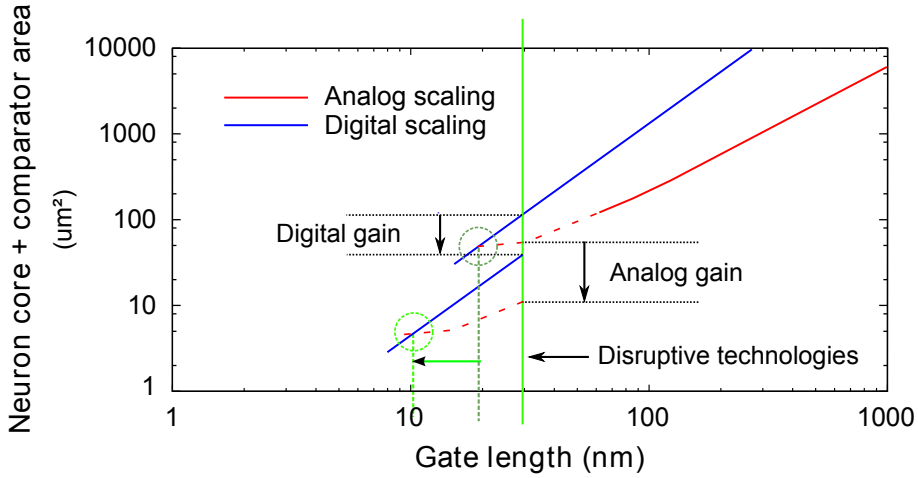


Figure III.21: Passage à l'échelle envisagé de neurones LIF : évolution probable des implémentations analogiques et numériques selon le nœud technologique. L'impact des technologies émergentes permet de repousser le point d'équivalence selon les gains procurés aux différentes implémentations.

aux gains apportés par l'utilisation de technologies émergentes. Les technologies en rupture, illustrées sur la figure III.21, entrainera des gains en termes de surface pour les implémentations de neurones. Comme ces technologies profiteront davantage à une implémentation analogique, on peut par conséquent envisager que le point d'équivalence énoncé dans la partie III-A. sera repoussé. Selon s'ils sont plus enclins à une implémentation analogique ou numérique, ces gains permettront un décalage plus ou moins important du point d'équivalence.

IV

Conclusion & perspectives

Le désordre est bien puissant quand il s'organise.

André Suarès, *Idées et Visions*, 1913.

Sommaire

A. Contributions de cette thèse 120
1. Caractéristiques du neurone 120
2. Exemple d'opérations typiques 121
B. Des perspectives à court terme. 121
1. La programmabilité 121
2. Les circuits intégrés à caractériser 121
C. Les architectures neuromorphiques, une technologie en devenir . . . 123
1. Un pas vers le marché industriel... 123
2. ... en étant attentif aux découvertes en neurobiologie 125

Ce chapitre recense les différents objectifs menés à terme lors de cette thèse et ouvre des perspectives d'avenir pour les architectures neuro-inspirées. Nous les rappelons brièvement et les détaillerons par la suite.

- Réalisation d'un neurone analogique en technologie 65 nm, étude de variabilité inter et intra puce.
- Envoi de 3 circuits (*Reptile*, *Spider*, *NeuroCBRAM*) en fabrication + caractérisation de l'un d'entre eux.
- Aperçu de l'impact des technologies émergentes.

A. Contributions de cette thèse

Les travaux menés pendant cette thèse ont permis la conception d'un neurone analogique en technologie 65 nm. Les différentes phases de son développement ont été détaillées et auront permis son intégration dans deux circuits en technologie 65 nm, *Reptile* et *Spider*. Le premier a fait l'objet de tests à l'échelle du neurone permettant sa caractérisation en termes de consommation et de variabilité. Il aura également rendu possible le réajustement de certains paramètres pour l'intégration dans le second circuit dont les capacités calculatoires sont largement étendues.

1. Caractéristiques du neurone

Différentes caractéristiques du neurone réalisé ont été étudiées. Nous résumons, dans le tableau IV.1, quelques exemples de neurones implémentés dans la littérature. Le neurone conçu pour le circuit *Reptile* se place au niveau de l'état de l'art mondial en termes de surface et d'énergie.

Référence	Techno (um)	Modèle	E_{spike}	E_{tot}	Surface (um^2)
(87)	0.35	Izhikevich	8.5-9.0 pJ	-	2800
(27)	0.5	IF	-	-	500
(33)	1.5	IF	3-15 nJ	-	2500
(35)	0.35	IF	-	900 pJ	500
(49)	0.35	$\sim$ aEIF	7-267 pJ	-	-
(54)	0.045 SOI	Digital IF	-	45-80 pJ	-
Neurone $\subset$ *Reptile*	0.065	IF	1.4-10 pJ	40 pJ	120

Table IV.1: Caractéristiques de neurones implémentés dans la littérature. En étant au niveau de l'état de l'art, le neurone intégré au circuit *Reptile* ouvre des perspectives nouvelles pour des architectures neuromorphiques basse consommation.

Il a également démontré la faisabilité d'un neurone analogique dans un nœud technologique avancé. La variabilité du neurone et de de son mécanisme d'injection a été caractérisée par un écart type égal à $\sigma = 6.25$ correspondant à un rapport signal sur bruit de 24 dB. La variabilité a été suffisament maitrisée dans la mesure où le neurone a pu pleinenement remplir son rôle d'opérateur mathématique élémentaire.

2. Exemple d'opérations typiques

Le neurone a ainsi montré ses performances calculatoires pour des opérations d'addition, de soustraction ou de multiplication. Ces premiers exemples de calculs sont illustrés sur la figure IV.1 En les interconnectant au sein d'un réseau, nous avons pu réaliser des opérations plus complexes, à savoir une détection de contour ou encore une dérivée temporelle.

La détection de contours utilise des trains d'impulsions $u_{i,j}$ dont la fréquence correspond à l'intensité du pixel $p_{i,j}$. Ces trains sont soustraits par l'intermédiaire d'un neurone en fonction de leur distance permettant ainsi l'ajustement de la finesse de détection.

B. Des perspectives à court terme.

1. La programmabilité

On a détaillé en introduction le besoin de programmabilité d'une architecture neuromorphique. Si elle peut se faire de manière similaire à celle d'un FPGA, d'autres critères pourraient être introduits lors de la programmation. Ils permettraient de privilégier certains aspects comme la robustesse ou plutôt la rapidité. Ceci est effectivement réalisable, et détaillé dans (28). On peut ainsi prévoir une marge sur les poids synaptiques ou distribuer le calcul sur un nombre plus élevé de neurones.

Un problème se pose lorsqu'une application n'est pas réalisable par le chainage d'opérations connues dans la librairie. Cependant, en stimulant un réseau soumis à des règles d'apprentissages, on peut accéder aux différents poids synaptiques constituant le réseau de neurones. On pourra préalablement optimiser ce jeu de connexions en vue de son implémentation en tant que boite noire. On connaitra en effet ses entrées/sorties et son comportement sans être pour autant capable de détailler précisément son fonctionnement. Si toutes les applications sont ainsi réalisables, il faudra cependant garder en mémoire qu'elles ne sont pas toujours prédisposées à un traitement neuro-inspiré.

2. Les circuits intégrés à caractériser

Le développement de la carte de test pour le circuit *Spider* permettra d'effectuer des mesures dès son retour de fabrication. La mise en place d'une caractérisation au-

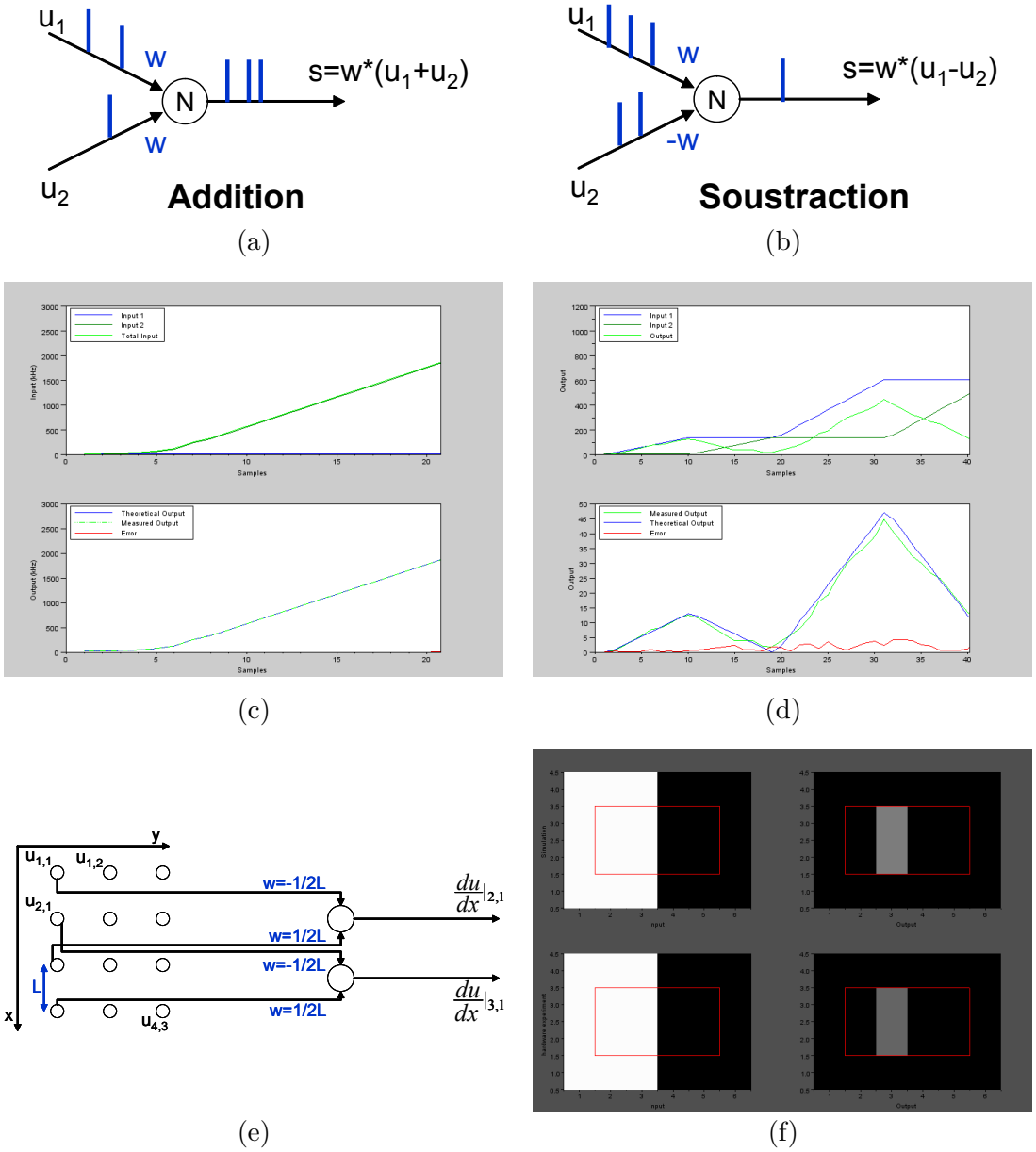


Figure IV.1: Les poids synaptiques d'un neurone permettent la réalisation d'une addition (a) et d'une soustraction (b). Pour chacune de ces opérations, les caractérisations du circuit sont comparées avec la valeur théorique afin d'estimer l'erreur produite (c) et (d). Un exemple plus complexe illustre la détection d'un contour réalisé par un filtrage sur une matrice de 8 pixels. La connectivité générique d'une telle implémentation est indiquée en (e), les résultats (f) sont en parfaite adéquation avec la simulation.

tomatique des 300 neurones facilitera le réglage optimisé du circuit. D'un point de vue applicatif, l'implémentation d'une transformée de Fourier sera possible et donnera un indicateur de performance à comparer avec les processeurs de signaux numériques existants. La topologie des circuits permet également leurs interconnexions pour étendre leur capacité.

Le circuit basé sur des synapses implémentées par des mémoires résistives de type CBRAM devra également être caractérisé dès son retour de fabrication. La co-intégration d'un neurone réalisé à l'aide d'un Design Kit industriel, couplé à des mémoires résistives, est effectivement une avancée notable dans le domaine du neuromorphique. Le circuit implémenté contient un nombre limité de neurones, mais qui permettra une étude précise des composants memristifs. Sa caractérisation pourra donner de précieuses indications quant à la marche à suivre pour une intégration à grande échelle des neurones et des synapses.

C. Les architectures neuromorphiques, une technologie en devenir

1. Un pas vers le marché industriel...

Dicté par les industriels, pour qui il est efficace et suffisant à l'heure actuelle, d'effectuer du calcul déterministe, l'approche *more Moore* reste la marche à suivre. A la vue du développement de technologies émergentes, les capacités de stockage à l'aide de mémoires ultra-denses vont indéniablement s'accroitre et l'énergie nécessaire à la communication entre les puces se réduire. Elles conféreront par conséquent de sérieux atouts aux futurs circuits pendant encore quelques années.

La figure IV.2 illustre les contraintes et les temps nécessaires au développement de circuits intégrés. Le temps t_p nécessaire à la maitrise de la variabilité, lors du développement d'un procédé de fabrication, engendre actuellement des investissements financiers de plus en plus importants. De la même façon, un temps t_a est également nécessaire à la conception et le développement d'une architecture.

Les temps de développement et de réalisation d'un circuit ($t_p + t_a$) ont été favorables au numérique. Les procédés de fabrication évoluaient en effet de manière à pouvoir

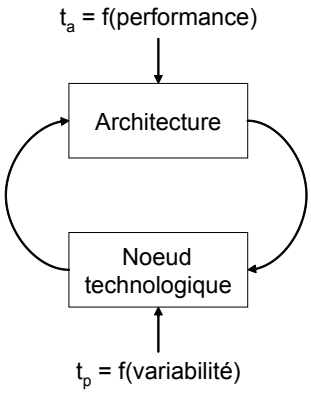


Figure IV.2: Temps de mise au point d'un procédé de fabrication et celui de la conception d'une architecture

bénéficier des fonctions déjà implémentées. Ceci équivaut à $t_a >> t_p$ se traduisant par un gain de performance sans d'importants changements au niveau de l'architecture.

Aujourd'hui la tendance semble plutôt tendre vers $t_a \sim t_p$ comme en témoigne l'approche d'Intel depuis 2007. Sa stratégie Tic-Tac emploie la même architecture sur deux noeuds technologiques parallèlement au développement d'un futur procédé de fabrication. Le gain de performance est alors causé successivement par un procédé de fabrication ou par une nouvelle architecture.

Et demain ? Les contraintes de conception (thermique, variabilité, etc) se complexifient, augmentant de ce fait t_a. Les procédés de fabrication laissent des opportunités qui pourraient parallèlement accroitre t_p. Il reste par conséquent encore quelques années pendant lesquelles la stratégie ci dessus fonctionnera. Cependant les grandeurs physiques ainsi que les conditions "normales" d'utilisations (température ambiante, utilisateur λ, etc) fixeront une limite aux procédés de fabrication et donc $t_p \longrightarrow \infty$. Le gain de performance sera donc très limité au niveau du procédé.

Afin de continuer l'amélioration des systèmes, il sera nécessaire d'envisager toutes sortes d'architectures, dont les architectures neuromorphiques. L'emploi de l'analogique couplé à ces technologies émergentes permettent un gain sérieux en termes d'énergie et de surface par rapport à une implémentation numérique. A l'aide des différentes caractérisations obtenues et présentées dans ce manuscrit, ce gain est illustré pour une architecture neuromorphique, sur la figure IV.3 selon différents degrés de précision à atteindre.

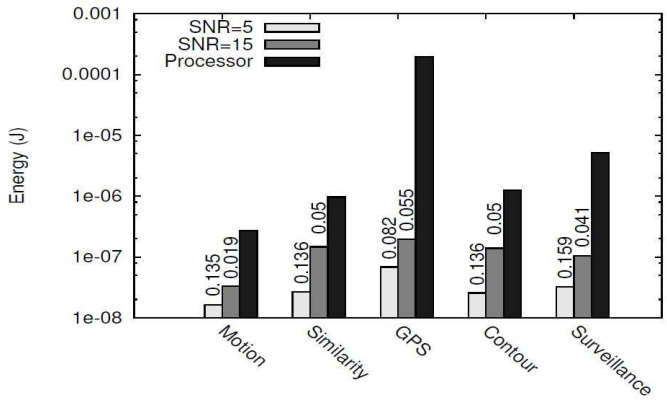


Figure IV.3: Comparaison de l'énergie consommée par un processeur et par une architecture neuromorphique basée sur la tuile du circuit *Spider*. Dans ce dernier cas, la détermination du rapport signal à bruit permet d'optimiser la consommation.

Ce gain sera de première importance lorsqu'il s'agira de convaincre un industriel d'utiliser une architecture neuromorphique pouvant lui paraitre "exotique". Dans le même ordre d'idées, le banc d'essai BenchNN (www.benchnn.org) implémente des tests de PARSEC de sorte qu'ils soient adaptés à la caractérisation des performances d'un réseaux de neurones. Lorsque le gain sera significatif d'un point de vue industriel, il faudra alors être en mesure de lui proposer une architecture neuromorphique capable de réaliser l'application qu'il désire.

Cette intégration d'accélérateurs neuronaux pourrait alors être réalisée progressivement. Un premier temps consisterait à intégrer des architectures neuromorphiques au sein d'un système sur puces. Elles seraient alors employées pour pallier le manque d'efficacité d'accélérateurs numériques face à une application. A plus long terme et sous réserve que la robustesse de ces derniers deviennent critique, des accélérateurs à base de neurones à impulsion pourraient alors devenir majoritaires.

2. ... en étant attentif aux découvertes en neurobiologie

Le cerveau est fascinant et si son fonctionnement est loin d'être entièrement connu, de nombreuses théories et découvertes en neurobiologies permettent d'éclaircir certains points. En électronique, les neurones et les synapses ont déjà largement été étudiés dans la littérature puisque leur intérêt calculatoire a rapidement été perçu. Les cellules

gliales jouent un rôle primordial dans le maintien de l'ordre du cerveau et sont pourtant très largement ignorées dans les implémentations électroniques. D'un point de vue biologique, elles permettent d'apporter des nutriments aux neurones ou d'évacuer les neurones déficients.

Une analogie électronique serait un bloc permettant :

- l'adaptation des neurones impactés lors de la fabrication
- la sélection et la déconnection des neurones hors spécification après adaptation.
- l'alimentation en courant/tension des neurones optimaux

Une fonction de la sorte est déjà employé dans les circuits numériques pour compenser les effets de la fabrication en modifiant, par exemple, la tension d'alimentation. L'implémentation d'un tel bloc pourrait devenir nécessaire dans des technogies très faiblement contrôlées et ceci même pour des architectures neuromorphiques massivement parallèles. Il sera alors judicieux d'estimer un rapport entre la variabilité du système et le coût nécessaire à sa maitrise.

Plusieurs millions d'années d'évolution ont été nécessaires à l'optimisation du cerveau. De manière similaire, il est légitime de penser que le développement de circuits efficaces et pérennes soit un processus long nécessitant, étape après étape, de relever une multitude de défis. Les procédés de miniaturisation constituent une de ces étapes et sont en passe d'être maitrisés. Cependant, outre ces évolutions, le vieillissement matériel d'un circuit est masqué par des problèmes logiciels et par un renouvellement souvent précoce. Si on tente d'adapter les dispositifs technologiques actuels dans une optique de durabilité, des phénomènes de vieillissement électroniques pourront fortement altérer leur fonctionnement.

Au contraire, les neurones électroniques intégrés dans un futur proche seront-ils infaillibles ? Évidemment non ; cependant, la robustesse inhérente aux réseaux de neurones permettra d'accroitre de manière significative la durée de vie d'un circuit neuro-inspiré.

A l'heure où les contraintes énergétiques s'accroissent et les ressources s'amenuisent, la longévité d'un circuit durable sera probablement l'impératif de demain.

Publications de l'auteur

Publications Scientifiques

Journal

- B. Belhadj, **A. Joubert**, O. Temam and R. Heliot, "Neuromorphic hardware as an alternative for real-world data processing", à soumettre dans IEEE Transaction on Circuits and Systems, (TCAS 2013)

Conférences avec actes

- **A. Joubert**, M. Duranton, B. Belhadj, O. Temam and R. Heliot, "Capacitance of TSVs in 3D Stacked Chips a problem ? - Not for Neuromorphic Systems ! ", Design Automation Conference, (DAC 2012)

  HiPEAC Paper Award 2012

- **A. Joubert**, B. Belhadj, O. Temam, R. Heliot, "Hardware spiking neurons design : analog or digital ?", IEEE International Joint Conference on Neural Networks, (IJCNN 2012)

- Belhadj, **A. Joubert**, O. Temam, R. Heliot, "Configurable Conduction Delay Circuits for High Spiking Rates", IEEE International Symposium on Circuits and Systems, (ISCAS 2012)

- **A. Joubert**, Bilel Belhadj and R. Heliot, "A robust and compact 65 nm LIF analog neuron for computational purposes", IEEE 9th International New Circuits and Systems Conference, (NEWCAS 2011)

- R. Heliot, **A. Joubert** and O. Temam, (2011) "Robust and Low-Power Accelerators based on Spiking Neurons for Signal Processing Applications", 3rd HiPEAC Workshop on Design for Reliability (DFR'11)

Brevets

- R. Heliot, M. Duranton, **A. Joubert**, "Utilisation des TSVs comme élément capacitif dans un neurone électronique"

Références

[1] Albea, C., Puschini, D., Vivet, P., Miro-Panades, I., Beigné, E. & Lesecq, S. (2011). Architecture and robust control of a digital frequency-locked loop for fine-grain dynamic voltage and frequency scaling in globally asynchronous locally synchronous structures. *Journal of Low Power Electronics*, **7**, 328–340. 69

[2] Alvado, L., Tomas, J., Saïghi, S., Renaud, S., Bal, T., Destexhe, A. & Le Masson, G. (2004). Hardware computation of conductance-based neuron models. *Neurocomputing*, **58-60**, 109–115. 17

[3] Bajolet, A., Giraudin, J., Rossato, C., Pinzelli, L., Bruyère, S., Cremer, S., Jagueneau, T., Delpech, P., Montès, L. & Ghibaudo, G. (2005). Three-dimensional 35 nf/mm^2 mim capacitors integrated in bicmos technology. In *Solid-State Device Research Conference, 2005. ESSDERC 2005. Proceedings of 35th European*, 121–124, IEEE. 97

[4] Bernstein, K., Frank, D.J., Gattiker, A.E., Haensch, W., Ji, B.L., Nassif, S.R., Nowak, E.J., Pearson, D.J. & Rohrer, N.J. (2006). High-performance CMOS variability in the 65-nm regime and beyond. *IBM Journal of Research and Development*, **50**, 433–449. 39

[5] Bichler, O., Querlioz, D., Thorpe, S., Bourgoin, J. & Gamrat, C. (2011). Unsupervised features extraction from asynchronous silicon retina through spike-timing-dependent plasticity. In *Neural Networks (IJCNN), The 2011 International Joint Conference on*, 859–866, IEEE. 91, 106

[6] Bienia, C., Kumar, S., Singh, J.P. & Li, K. (2008). The parsec benchmark suite : characterization and architectural implications. In *Proceedings of the 17th international conference on Parallel architectures and compilation techniques*, PACT '08, 72–81, ACM, New York, NY, USA. 25

[7] Blanchard, P. (2011). First commercial demonstration of an emerging memory technology for embedded flash using cbram. In *Workshop on Innovative Memory Technologies, MINATEC, Grenoble - France*. 95

[8] Boahen, K. (2000). Point-to-point connectivity between neuromorphic chips using address events. *Circuits and Systems II : Analog and Digital Signal Processing, IEEE Transactions on*, **47**, 416 –434. 15, 137

[9] Brette, R. & Gerstner, W. (2005). Adaptive exponential integrate-and-fire model as an effective description of neuronal activity. *Journal of neurophysiology*, **94**, 3637–3642. 11

[10] Cadix, L., Fuchs, C., Rousseau, M., Leduc, P., Chaabouni, H., Thuaire, A., Brocard, M., Valentian, A., Farcy, A., Bermond, C. *et al.* (2010). Integration and frequency dependent parametric modeling of through silicon via involved in high density 3d chip stacking. *ECS Transactions*, **33**, 1–21. 100, 101

[11] Camilleri, P., Giulioni, M., Dante, V., Badoni, D., Indiveri, G., Michaelis, B., Braun, J. & Del Giudice, P. (2007). A Neuromorphic aVLSI network chip with configurable plastic synapses. In *Proceedings of the 7th International Conference on Hybrid Intelligent Systems*, 296–301, IEEE Computer Society. 15, 17

[12] Cauwenberghs, G. (1996). An analog VLSI recurrent neural network learning a continuous-time trajectory. *IEEE Transactions on Neural Networks*, **7**, 346–361. 15, 17, 88

[13] Chan, V., Liu, S.C. & van Schaik, A. (2007). AER EAR : A Matched Silicon Cochlea Pair With Address Event Representation Interface. *IEEE Transactions on Circuits and Systems I : Regular Papers*, **54**, 48–59. 15, 18

[14] Chicca, E., Indiveri, G. & Douglas, R. (2004). An event based VLSI network of integrate-and-fire neurons. In *Proc. IEEE Int. Symp. Circuits Syst*, vol. pp, 357–360, Citeseer. 15, 17, 88

[15] Choi, Y., Mosley, L., Min, Y. & Amaratunga, G. (2010). Carbon nanotube capacitors arrays using high-k dielectrics. *Diamond and Related Materials*, **19**, 221–224. 97

[16] Chua, L. (1971). Memristor-the missing circuit element. *Circuit Theory, IEEE Transactions on*, **18**, 507–519. 89

[17] Chua, L. & Kang, S. (1976). Memristive devices and systems. *Proceedings of the IEEE*, **64**, 209–223. 91

[18] Das, S., Fan, A., Chen, K., Tan, C., Checka, N. & Reif, R. (2004). Technology, performance, and computer-aided design of three-dimensional integrated circuits. In *Proceedings of the 2004 international symposium on Physical design*, 108–115, ACM. 98

[19] De Pittà, M., Volman, V., Berry, H. & Ben-Jacob, E. (2011). A Tale of Two Stories : Astrocyte Regulation of Synaptic Depression and Facilitation. *PLoS Comput Biol*, **7**, e1002293+. 4

[20] Detalle, M., Barrenetxea, M., Muller, P., Potoms, G., Phommahaxay, A., Soussan, P., Vaesen, K. & De Raedt, W. (2010). High density, low leakage Back-End 3D capacitors for mixed signals applications. *Microelectronic Engineering*, **87**, 2571–2576. 97

[21] Douence, V. (2000). *Circuits et systèmes de modélisations analogiques de neurones biologiques*. Ph.D. thesis, Bordeaux 1. 4

[22] Emery, R., Yakovlev, A. & Chester, G. (2009). Connection-centric network for spiking neural networks. In *Proceedings of the 2009 3rd ACM/IEEE International Symposium on Networks-on-Chip-Volume 00*, 144–152, IEEE Computer Society. 13, 17, 88

[23] Esmaeilzadeh, H., Sampson, A., Ceze, L. & Burger, D. (2012). Neural acceleration for general-purpose approximate programs. In *Proceedings of the 45th Annual IEEE/ACM International Symposium on Microarchitecture*. 25

[24] Farquhar, E. & Hasler, P. (2005). A bio-physically inspired silicon neuron. *IEEE Transactions on Circuits and Systems I : Regular Papers*, **52**, 477–488. 14

[25] Figueiredo, P.M. & Vital, J.C. (2009). *Offset Reduction Techniques in High-Speed Analog-to-Digital Converters : Analysis, Design and Tradeoffs*. Springer. 52

[26] Fitzhugh, R. (1961). Impulses and Physiological States in Theoretical Models of Nerve Membrane. *Biophysical journal*, **1**, 445–66. 10

[27] Goldberg, D., Cauwenberghs, G. & Andreou, A. (2001). Probabilistic synaptic weighting in a reconfigurable network of VLSI integrate-and-fire neurons. *Neural Networks*, **14**, 781–793. 15, 17, 88, 120

[28] Heliot, R., Joubert, A. & Temam, O. (2011). Robust and low-power accelerators based on spiking neurons for signal processing applications. In *The 3rd HiPEAC Workshop on Design for Reliability (DFR'11)*. 27, 34, 45, 121

[29] Hodgkin, A.L. & Huxley, A.F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology*, **117**, 500–44. 2, 6

[30] Hynna, K. & Boahen, K. (2007). Silicon neurons that burst when primed. In *IEEE International Symposium on Circuits and Systems, 2007. ISCAS 2007*, 3363–3366. 17

[31] IBM (2006). http://www.almaden.ibm.com/st/past_projects/nano_devices/phasechangememory/. 94, 95

[32] Ielmini, D., Lavizzari, S., Sharma, D. & Lacaita, A. (2007). Physical interpretation, modeling and impact on phase change memory (pcm) reliability of resistance drift due to chalcogenide structural relaxation. In *Electron Devices Meeting, 2007. IEDM 2007. IEEE International*, 939–942, IEEE. 105

[33] Indiveri, G. (2003). A low-power adaptive integrate-and-fire neuron circuit. In *Proceedings of the 2003 International Symposium on Circuits and Systems, 2003. ISCAS '03.*, vol. 4, IV–820–IV–823, IEEE. 76, 120

[34] Indiveri, G. & Fusi, S. (2007). Spike-based learning in VLSI networks of integrate-and-fire neurons. In *Proc. IEEE International Symposium on Circuits and Systems, ISCAS*, vol. 2007, 3371–3374, Citeseer. 15, 17

[35] Indiveri, G., Chicca, E. & Douglas, R. (2006). A VLSI array of low-power spiking neurons and bistable synapses with spike-timing dependent plasticity. *IEEE Transactions on Neural Networks*, **17**, 211–221. 15, 17, 88, 120

[36] ITRS (2011). http://www.itrs.net/Links/2011ITRS/Home2011.htm. 19, 23, 24, 37, 39

[37] Izhikevich, E. (2003). Simple model of spiking neurons. *IEEE Transactions on neural networks*, **14**, 15691572. 10

[38] James, D. (2012). Intel ivy bridge unveiled - the first commercial tri-gate, high-k, metal-gate cpu. In *Custom Integrated Circuits Conference (CICC), 2012 IEEE*, 1 –4. 23

[39] Jeannot, S., Bajolet, A., Manceau, J., Cremer, S., Deloffre, E., Oddou, J., Perrot, C., Benoit, D., Richard, C., Bouillon, P. *et al.* (2007). Toward next high performances MIM generation : up to 30fF/μm^2 with 3D architecture and high-κ materials. In *Electron Devices Meeting, 2007. IEDM 2007. IEEE International*, 997–1000, IEEE. 97

[40] Jung, R., Brauer, E.J. & Abbas, J.J. (2001). Real-time interaction between a neuromorphic electronic circuit and the spinal cord. *IEEE transactions on neural systems and rehabilitation engineering : a publication of the IEEE Engineering in Medicine and Biology Society*, **9**, 319–26. 17

[41] Khan, M., Lester, D., Plana, L., Rast, A., Jin, X., Painkras, E. & Furber, S. (2008). Spinnaker : mapping neural networks onto a massively-parallel chip multiprocessor. In *Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*, 2849–2856, IEEE. 12

[42] Koch, U. & Brunner, M. (1988). A modular analog neuron-model for research and teaching. *Biological cybernetics*, **312**, 303–312. 28

[43] Kozyrakis, C., Kansal, A., Sankar, S. & Vaid, K. (2010). Server Engineering Insights for Large-Scale Online Services. *IEEE Micro*, **30**, 8–19. 25

[44] Kuzum, D., Jeyasingh, R.G.D. & Wong, H.S.P. (2011). Energy efficient programming of nanoelectronic synaptic devices for large-scale implementation of associative and temporal sequence learning. In *2011 International Electron Devices Meeting*, 30.3.1–30.3.4, IEEE. 91, 105

[45] Le Masson, G., Renaud-Le Masson, S., Debay, D. & Bal, T. (2002). Feedback inhibition controls spike transfer in hybrid thalamic circuits. *Nature*, **417**, 854–858. 17

[46] Le Masson, S., Laflaquiere, A., Bal, T. & Le Masson, G. (1999). Analog circuits for modeling biological neural networks : design and applications. *IEEE transactions on biomedical engineering*, **46**, 638–645. 14, 17

[47] Lichtsteiner, P., Posch, C. & Delbruck, T. (2008). A 128 x 128 120 dB 15 μs Latency Asynchronous Temporal Contrast Vision Sensor. *IEEE Journal of Solid-State Circuits*, **43**, 566–576. 15, 18

[48] Linares-Barranco, B., Serrano-Gotarredona, T. & Serrano-Gotarredona, R. (2003). Compact low-power calibration mini-DACs for neural arrays with programmable weights. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, **14**, 1207–16. 49

[49] Livi, P. & Indiveri, G. (2009). A current-mode conductance-based silicon neuron for address-event neuromorphic systems. 2898–2901. viii, 43, 64, 76, 120

[50] Mahowald, M. & Douglas, R. (1991). A silicon neuron. *Nature*, **354**, 515–518. 14, 17

[51] Markram, H. (2006). The blue brain project. *Nature Reviews Neuroscience*, **7**, 153–160. 13

[52] Matolin, D., Schreiter, J., Schfiffny, R., Heittmann, A. & Ramacher, U. (2004). Simulation and implementation of an analog VLSI pulse-coupled neural network for image segmentation. In *The 2004 47th Midwest Symposium on Circuits and Systems, 2004. MWSCAS '04.*, vol. 2, II_397–II_400, IEEE. 17

[53] Mead, C. (1989). *Analog VLSI and neural systems*. VLSI systems series, Addison-Wesley. 2, 5, 6, 14

[54] Merolla, P., Arthur, J., Akopyan, F., Imam, N., Manohar, R. & Modha, D. (2011). A digital neurosynaptic core using embedded crossbar memory with 45pJ per spike in 45nm. 1 –4. 64, 76, 120

[55] Modha, D.S. & Singh, R. (2010). Network architecture of the long-distance pathways in the macaque brain. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 13485–90. 7

[56] Moore, G.E. (1965). Cramming More Components onto Integrated Circuits. *Electronics*, **38**, 114–117. 19

[57] Morel, P.H. (2011). *Etude de l'Intégration 3D et des Propriétés Physiques de Nanofils de Silicium obtenus par Croissance – Réalisation de Capacités Ultra-Denses*. Ph.D. thesis, Université de Grenoble. 96, 97

[58] Muller, M. (2010). Dark Silicon and the Internet. In *EE Times "Designing with ARM" virtual conference*. 26

[59] Nagumo, J., Arimoto, S. & Yoshizawa, S. (1962). An Active Pulse Transmission Line Simulating Nerve Axon. *Proceedings of the IRE*, **50**, 2061–2070. 10, 11

[60] Nere, A., Olcese, U., Balduzzi, D. & Tononi, G. (2012). A neuromorphic architecture for object recognition and motion anticipation using burst-STDP. *PloS one*, **7**, e36958. 28

[61] Palma, G., Vianello, E., Cagli, C., Molas, G., Reyboz, M., Blaise, P., De Salvo, B., Longnos, F. & Dahmani, F. (2012). Experimental investigation and empirical modeling of the set and reset kinetics of ag-ges2 conductive bridging memories. In *Memory Workshop (IMW), 2012 4th IEEE International*, 1 –4. 109

[62] Prodromakis, T., Toumazou, C. & Chua, L. (2012). Two centuries of memristors. *Nature Materials*, **11**, 478–481. 90

[63] Querlioz, D., Bichler, O. & Gamrat, C. (2011). Simulation of a memristor-based spiking neural network immune to device variations. In *Neural Networks (IJCNN), The 2011 International Joint Conference on*, 1775 –1781. 92, 109

[64] Rahman, A. & Reif, R. (2000). System-level performance evaluation of three-dimensional integrated circuits. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, **8**, 671–678. 98

[65] Renaud, S., Tomas, J., Bornat, Y., Daouzli, A. & Saighi, S. (2007). Neuromimetic ICs with analog cores : an alternative for simulating spiking neural networks. In *2007 IEEE International Symposium on Circuits and Systems*, 3355–3358, IEEE, New Orleans. 17

[66] Renaud, S., Tomas, J., Lewis, N., Bornat, Y., Daouzli, A., Rudolph, M., Destexhe, A. & Saïghi, S. (2010). PAX : A mixed hardware/software simulation platform for spiking neural networks. *Neural networks : the official journal of the International Neural Network Society*, **23**, 905–916. 6, 8, 10

[67] Rison, B. (1998). EE 308 Course. http://www.ee.nmt.edu/~rison/ee308_spr99/supp/990119/princeton.gif. 19

[68] Rosenblatt, F. (1958). The perceptron : A probabilistic model for information storage and organization in the brain. *Psychological Review*, **65**, 386–408. 2, 8

[69] Roy, K., Mukhopadhyay, S. & Mahmoodi-Meimand, H. (2003). Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits. *Proceedings of the IEEE*, **91**, 305–327. 49

[70] Saighi, S., Tomas, J., Bornat, Y. & Renaud, S. (2005). A Conductance-Based Silicon Neuron with Dynamically Tunable Model Parameters. In *2nd International IEEE EMBS Conference on Neural Engineering*, 1, 285–288, IEEEE. 17

[71] Sarpeshkar, R. (1998). Analog versus digital : extrapolating from electronics to neurobiology. *Neural Computation*, **10**, 1601–1638. 39, 40

[72] Schemmel, J., Bruderle, D., Meier, K. & Ostendorf, B. (2007). Modeling synaptic plasticity within networks of highly accelerated I&F neurons. In *IEEE International Symposium on Circuits and Systems, 2007. ISCAS 2007*, 3367–3370. 17

[73] Shibuya, A., Ouchi, A. & Takemura, K. (2010). A silicon interposer with an integrated thin film decoupling capacitor and through-silicon vias. *Components and Packaging Technologies, IEEE Transactions on*, **33**, 582 –587. 97

[74] Simoni, M., Cymbalyuk, G., Sorensen, M., Calabrese, R. & DeWeerth, S. (2004). A multiconductance silicon neuron with biologically matched dynamics. *IEEE Transactions on Biomedical Engineering*, **51**, 342–354. 17

[75] Smith, L. (2006). Implementing neural models in silicon. *Handbook of nature-inspired and innovative computing*, 433–476. 5

[76] Snider, G. (2008). Spike-timing-dependent learning in memristive nanodevices. In *Nanoscale Architectures, 2008. NANOARCH 2008. IEEE International Symposium on*, 85–92, IEEE. 91

[77] Song, S., Miller, K.D. & Abbott, L.F. (2000). Competitive Hebbian learning through spike-timing-dependent synaptic plasticity. *Nature Neuroscience*, **3**, 919–926. 5

[78] SPEC (2012). http://www.spec.org. 24

[79] Srinivasan, M. & Bernard, G. (1976). A proposed mechanism for multiplication of neural signals. *Biological Cybernetics*, **21**, 227–236. 45

[80] Strukov, D.B., Snider, G.S., Stewart, D.R. & Williams, R.S. (2008). The missing memristor found. *Nature*, **453**, 80–83. 90

[81] Thomas, M., Farcy, A., Gaillard, N., Perrot, C., Gros-Jean, M., Matko, I., Cordeau, M., Saikaly, W., Proust, M., Caubet, P. *et al.* (2006). Integration of a high density Ta_2O_5 MIM capacitor following 3D damascene architecture compatible with copper interconnects. *Microelectronic engineering*, **83**, 2163–2168. 97

[82] Tomas, J., Bornat, Y., Saighi, S., Levi, T. & Renaud, S. (2006). Design of a modular and mixed neuromimetic ASIC. In *2006 13th IEEE International Conference on Electronics, Circuits and Systems*, 946–949, IEEE. 17

[83] Topol, A., Tulipe, D., Shi, L., Frank, D., Bernstein, K., Steen, S., Kumar, A., Singco, G., Young, A., Guarini, K. *et al.* (2006). Three-dimensional integrated circuits. *IBM Journal of Research and Development*, **50**, 491–506. 98

[84] Valle, M., Caviglia, D. & Bisio, G. (1996). An experimental analog VLSI neural network with on-chip back-propagation learning. *Analog Integrated Circuits and Signal Processing*, **9**, 231–245. 17

[85] Van Schaik, A. (2001). Building blocks for electronic spiking neural networks. *Neural Networks*, **14**, 617–628. viii, 41, 51

[86] Vogelstein, R.J., Mallik, U., Vogelstein, J.T. & Cauwenberghs, G. (2007). Dynamically reconfigurable silicon array of spiking neurons with conductance-based synapses. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, **18**, 253–65. viii, x, 15, 16, 17, 41, 42, 49, 51, 138

[87] Wijekoon, J.H.B. & Dudek, P. (2008). Compact silicon neuron circuit with spiking and bursting behaviour. *Neural networks : the official journal of the International Neural Network Society*, **21**, 524–34. viii, 16, 17, 42, 64, 76, 120

[88] Zamarreño-Ramos, C., Camuñas-Mesa, L.A., Pérez-Carrasco, J.A., Masquelier, T., Serrano-Gotarredona, T. & Linares-Barranco, B. (2011). On spike-timing-dependent-plasticity, memristive devices, and building a self-learning visual cortex. *Frontiers in neuroscience*, **5**, 26. 91

[89] Zidenberg, T., Keslassy, I. & Weiser, U. (2012). MultiAmdahl : How Should I Divide My Heterogenous Chip ? *IEEE Computer Architecture Letters*, **11**, 65–68. 26

V

Annexes

A. *Adress Event Representation* - AER

L'Adresse Event Representation a été décrit par Kwabena A. Boahen dans (8) dont sont tirés les principes brièvement rappelés ici. L'AER est un protocole de communication asynchrone utilisant un multiplexage temporel. Il permet la communication de neurones entre plusieurs circuits ou, au sein d'un même circuit entre plusieurs clusters.

La génération d'un potentiel d'action d'un neurone source est détectée par un protocole de type poignée de main. Une partie d'arbitrage transcrit cette impulsion selon l'adresse du neurone source généralement décrite de façon matricielle selon ses coordonnées ligne/colonne. Ceci est décrit sur la figure V.1.

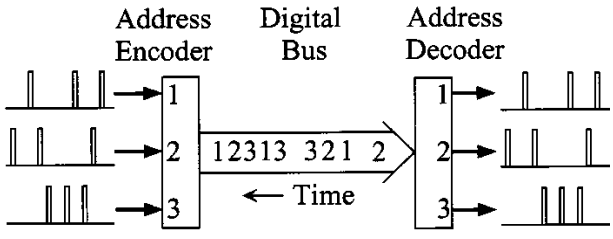


Figure V.1: Encodage des impulsions selon l'adresse du neurone source, d'après (8)

Il est ensuite acheminé vers le(s) neurone(s) de destination par le biais d'impulsions numérique pouvant fonctionner à des fréquences allant jusqu'à plusieurs MHz. La fréquence faible des neurones de l'ordre du kHz, permet l'utilisation d'un bus, d'un encodeur, d'un décodeur et d'un arbitre pour un nombre relativement élevé de neurones.

Ce nombre peut-être choisi en fonction des spécificités ou des contraintes à respecter. On peut en effet envisager d'éviter le traitement de potentiels d'actions simultanés ou être limité par la fréquence de fonctionnement d'un DAC. Un exemple d'intégration d'architecture neuromorphique avec un protocole AER est illustré sur la figure V.2.

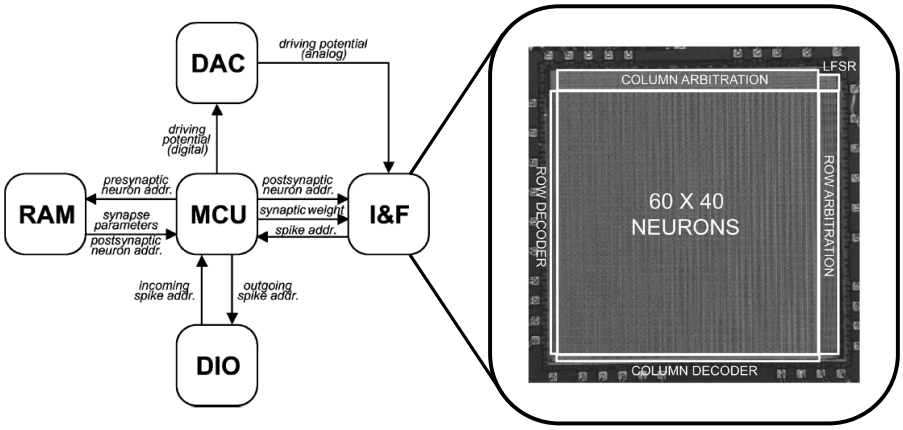


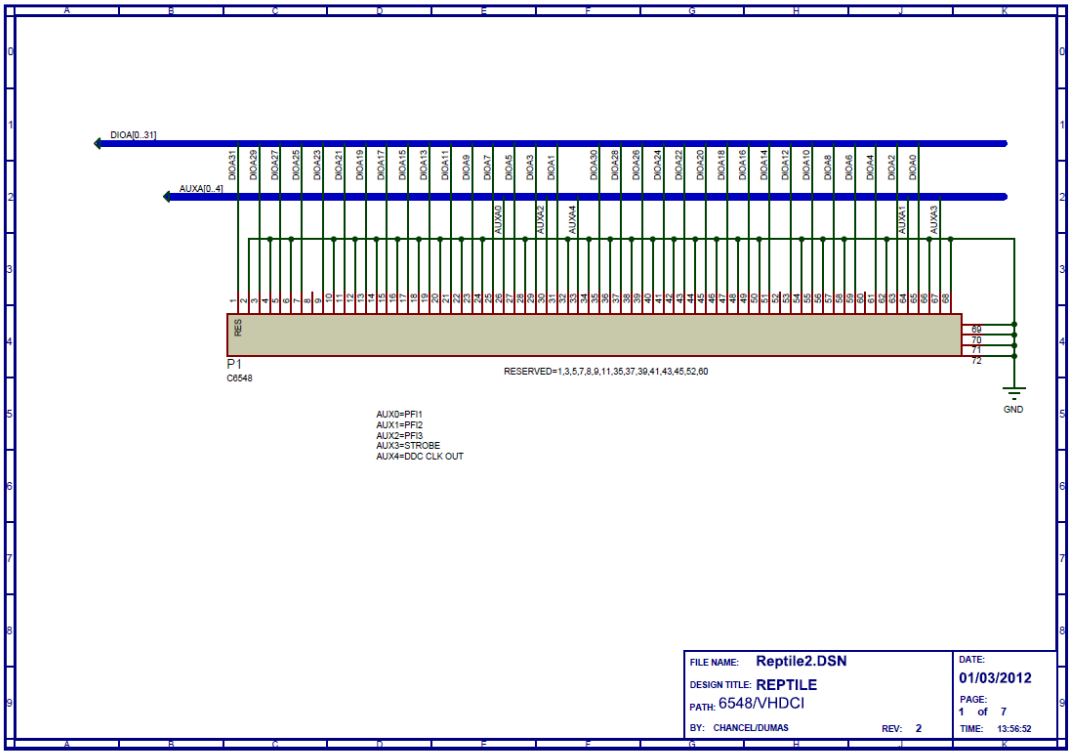
Figure V.2: Architecture AER décrit dans (86)

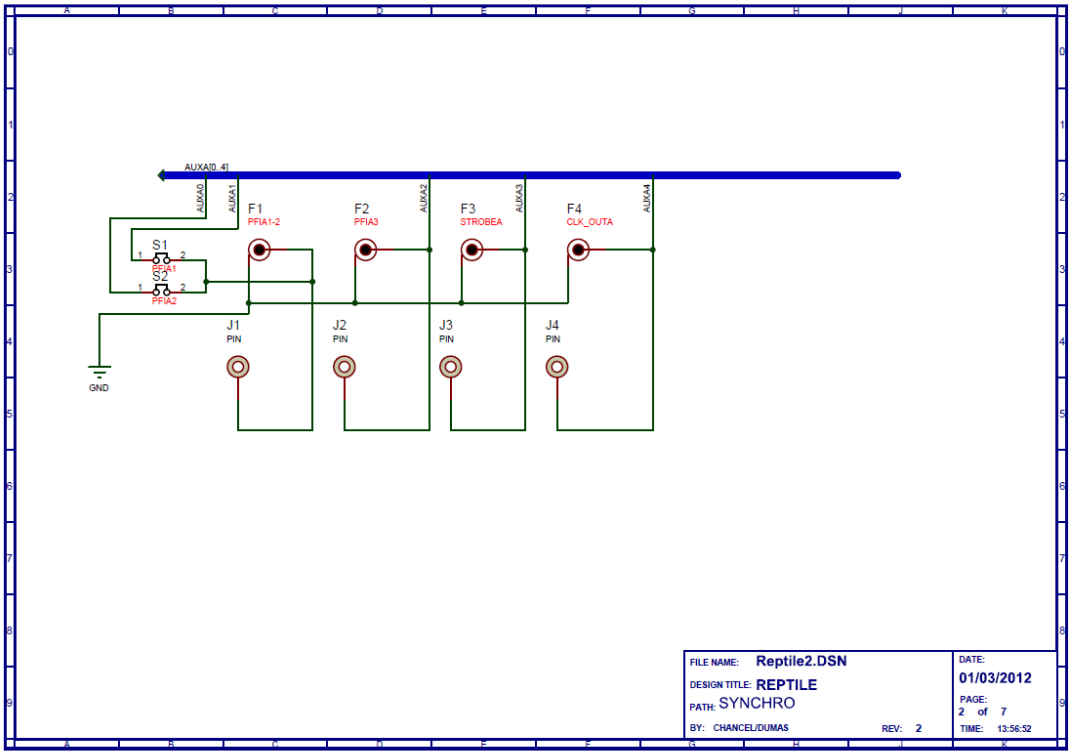
Un contrôleur (MCU) commande une matrice de neurone qui recherche dans une mémoire (RAM) les neurones de destinations et les poids synaptiques correspondants. Une fois ceux-ci obtenu, l'injection est réalisé à l'aide du CNA (DAC) et de la sélection du neurone de destination par le contrôleur. Le bloc DIO permet les entrées et sorties d'impulsion vers d'autres puces.

B. Carte de test

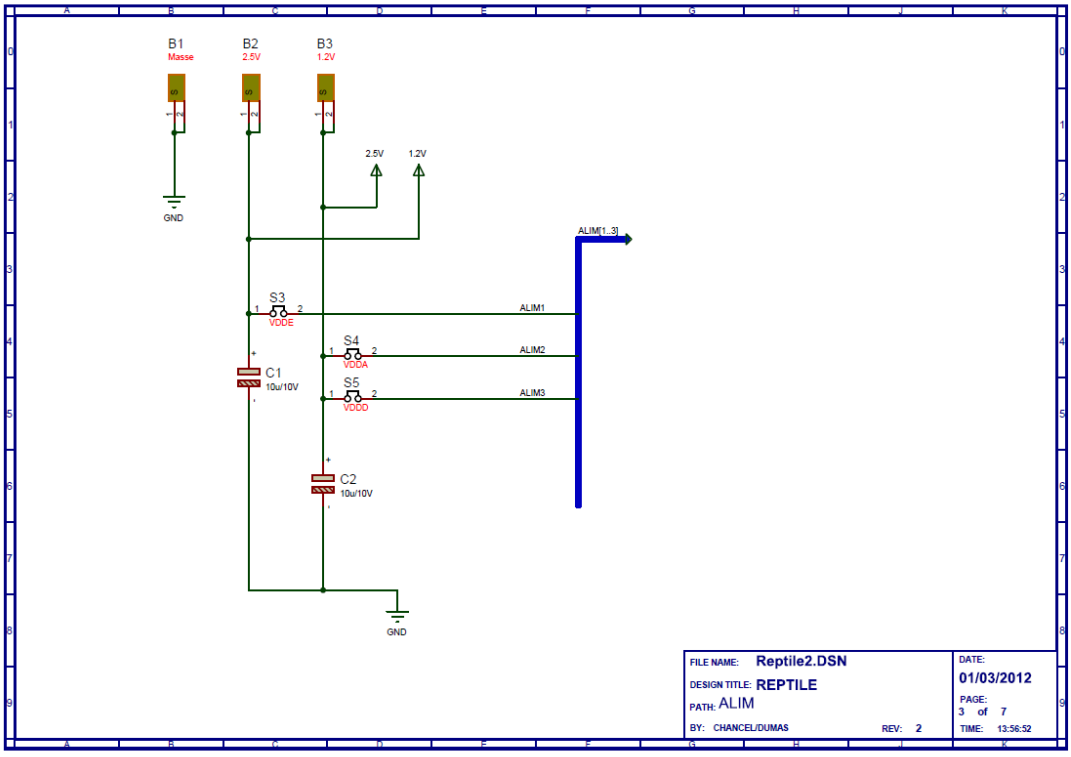
La carte de test a permis la caractérisation du circuit *Reptile*. Elle a pour but d'interfacer la banc de contrôle et les sources d'alimentation avec le circuit. Les courants et tensions de polarisations peuvent également être réglés. La carte fournit des points de mesure pour l'observation des signaux à haute fréquence de la FLL ou du neurone instrumenté. Son schéma est décrit sur les images des pages suivantes.

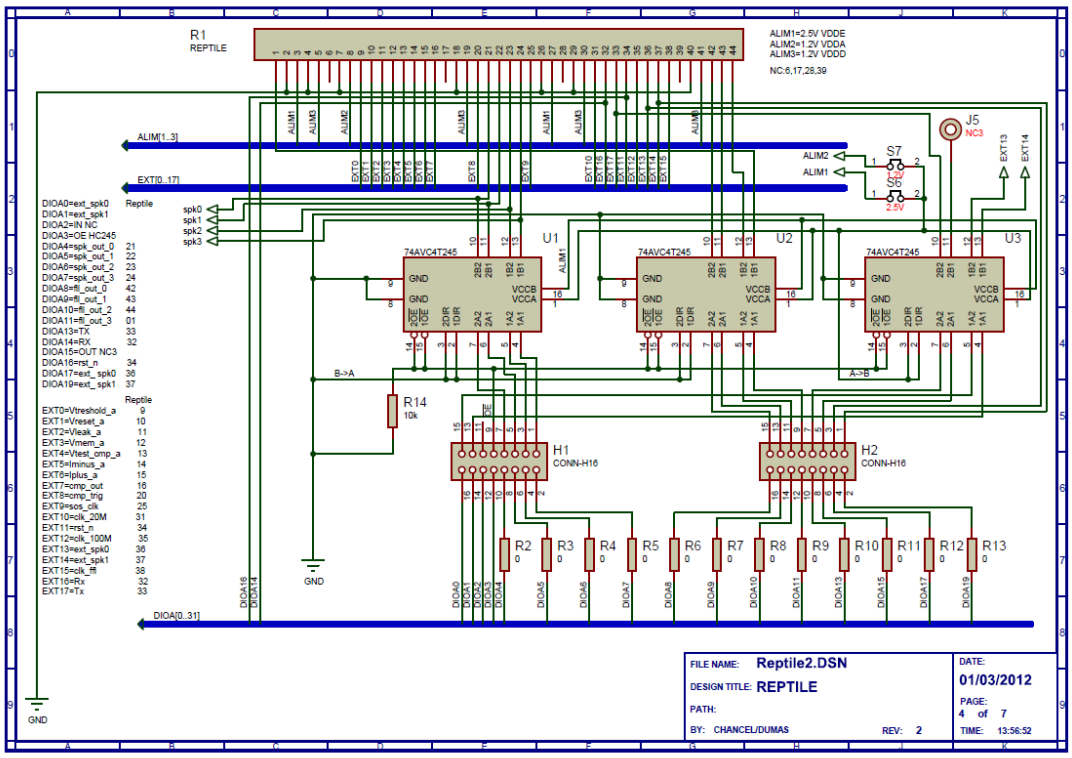
V. ANNEXES



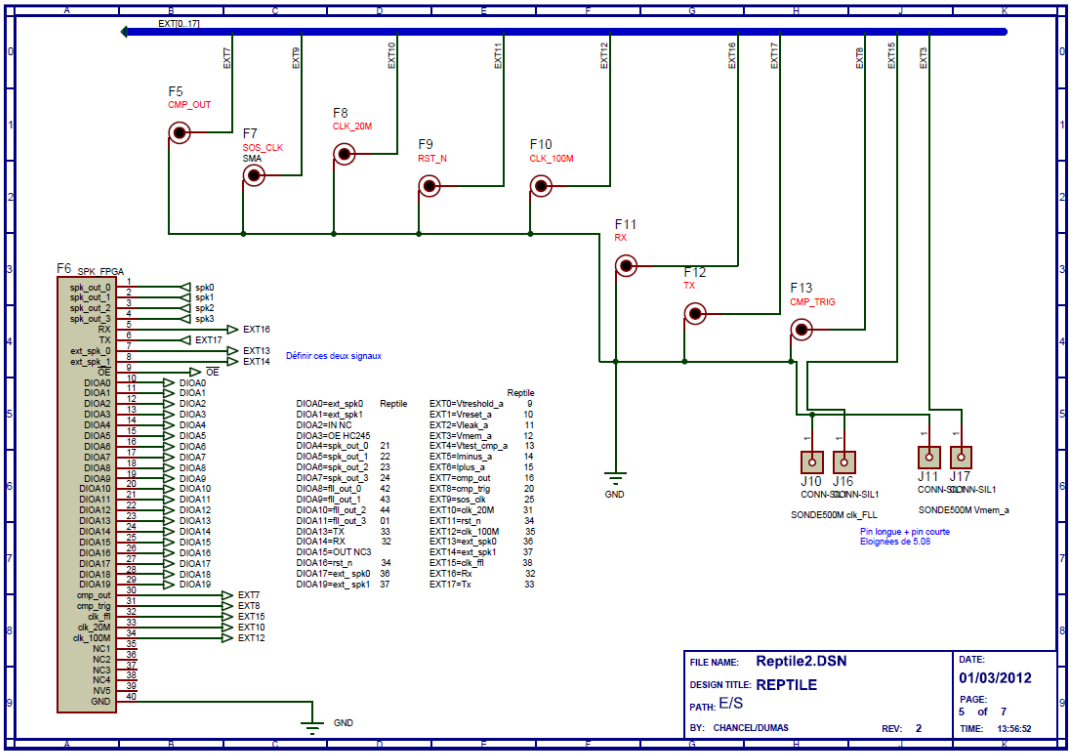


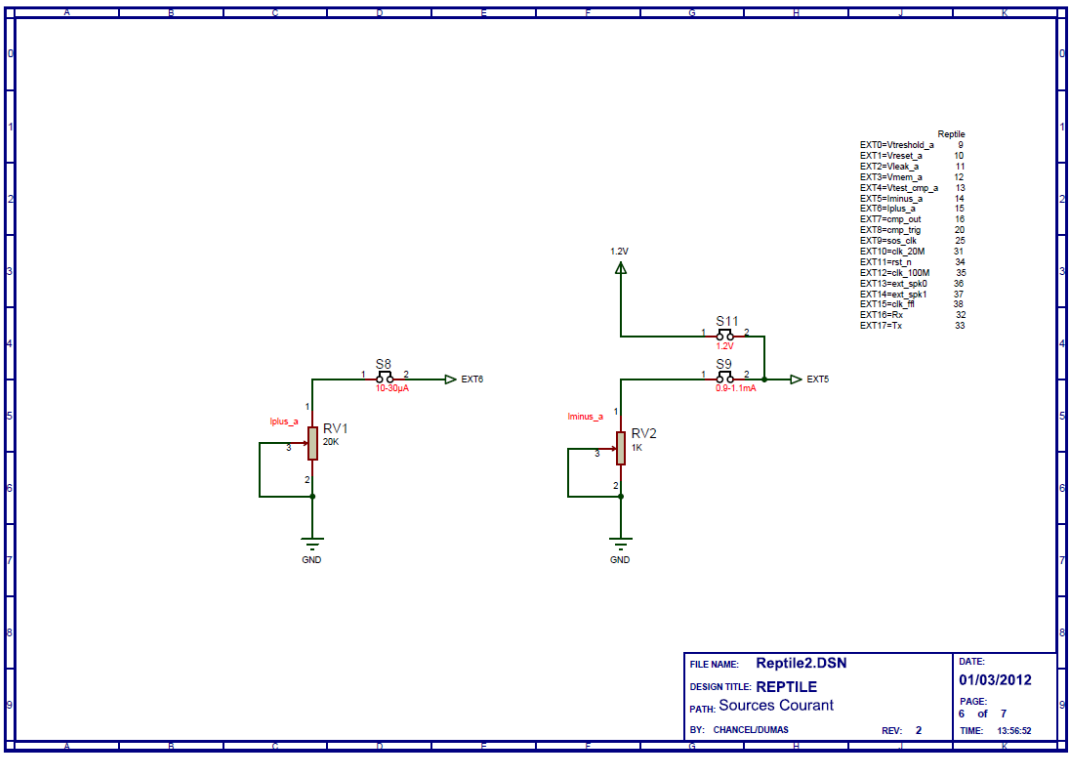
En haut : brochage pour connecteur au châssis de caractérisation. En bas : signaux de synchronisation.



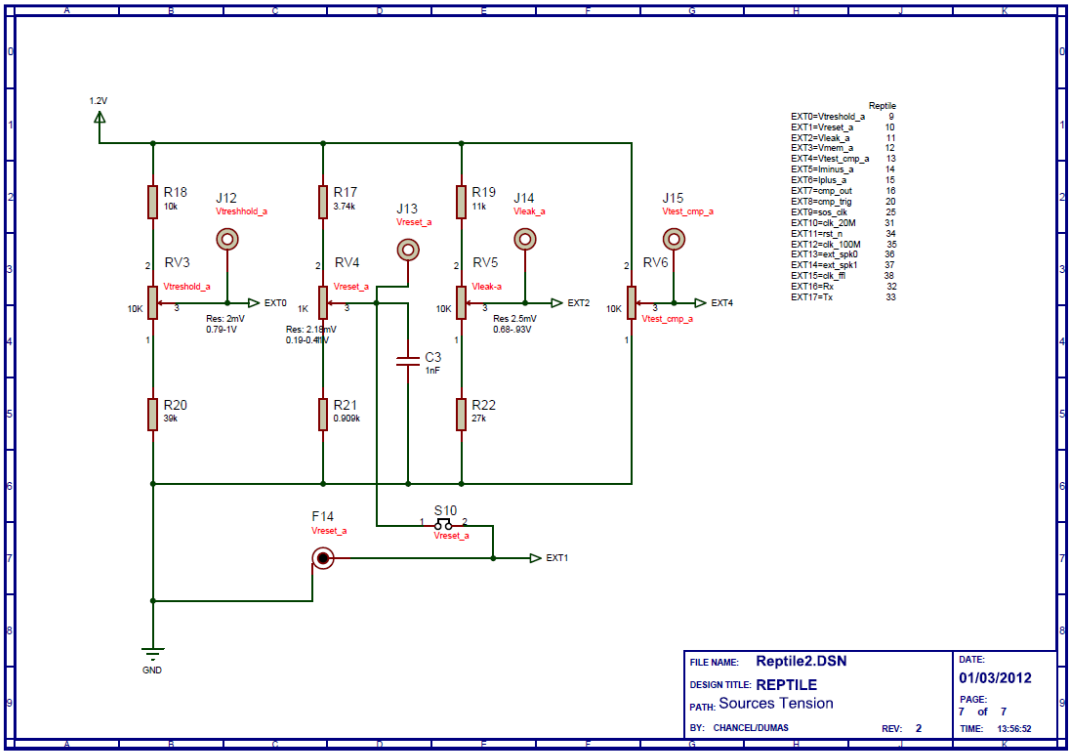


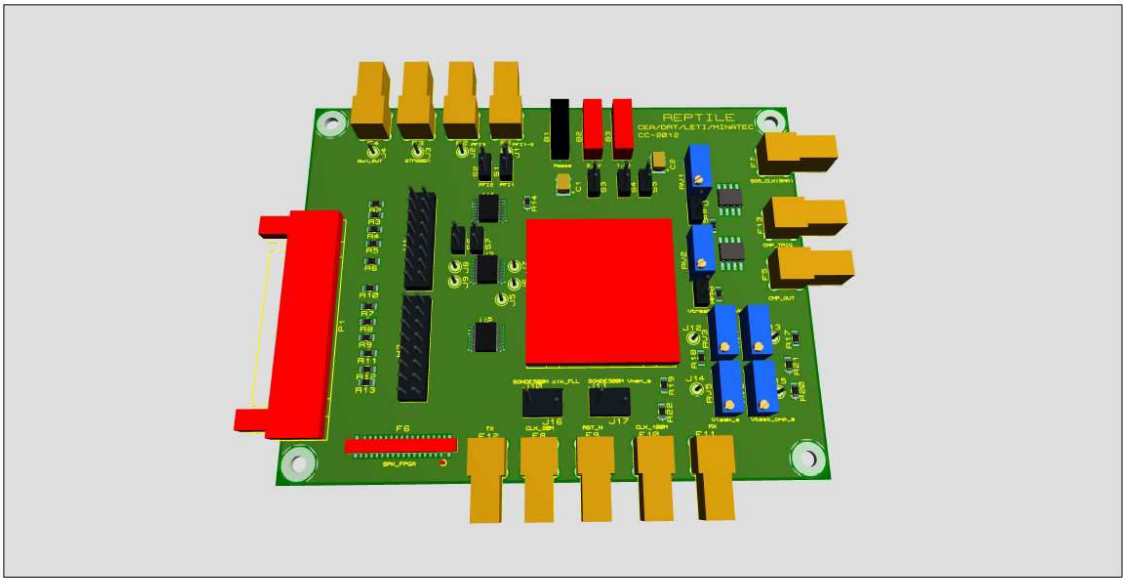
En haut : Alimentations des parties numérique, analogique et de la couronne de plot.
En bas : vue générale de la carte, connections au circuit *Reptile*.





En haut : Entrées/Sorties et signaux analogiques HF. En bas : Sources de courant.





En haut : Tensions de polarisation. En bas : Vue 3D de la carte conçue.