Conception d'un neurone analogique Leaky Integrate-and-Fire robuste et compact en technologie 65 nm

Antoine JOUBERT
CEA-LETI, Minatec
17 rue des Martyrs
38054 GRENOBLE Cedex 9

Rodolphe Héliot
CEA-LETI, Minatec
17 rue des Martyrs
38054 GRENOBLE Cedex 9

Email : joubert.antoine@cea.fr

Résumé

Les contraintes de robustesse et de consommation deviennent prépondérantes dans les technologies deca-nanométriques. Les architectures neuromorphiques ont à la fois la capacité de s'affranchir de ces problèmes et d'effectuer des applications de type traitement de signal. Elles requièrent l'implémentation à grande échelle de neurones robustes aux défauts de fabrication, compacts et faible consommation. Nous proposons l'implémentation d'un neurone de type « Leaky Integrate-and-Fire » (LIF) dans la technologie ST CMOS 65 nm. Sa surface est de 100 um² et sa précision, lorsqu'il est soumis à des variations de fabrication, est de 35 dB.

1. Introduction

Les architectures de calcul sont en proie à de nouvelles contraintes de consommation et de variabilité qui forcent les concepteurs à développer de nouveaux types d'architectures. L'utilisation de systèmes hétérogènes d'accélérateurs matériels permet de réduire la consommation [1], mais il y a un réel besoin d'accélérateurs robustes à la variabilité des technologies déca-nanométriques [2].

Grâce à leur fonctionnement massivement parallèle, les réseaux de neurones sont de bons candidats pour résoudre ces problèmes. Ils effectuent des calculs à l'aide d'opérateurs simples, les neurones, et des mécanismes d'apprentissage leur permettent d'être plus robustes aux défauts de fabrication. Depuis les travaux de C. Mead [3], de nombreux travaux ont été effectués dans le domaine des réseaux de neurones [4-5]. En particulier, le développement du protocole AER (Adress-Event Representation) [6] a permis l'interconnexion d'un grand nombre de neurones.

Nous avons proposé dans [7], un accélérateur robuste et faible consommation basé sur des neurones analogiques à impulsions. Il peut effectuer un large spectre de tâches de type traitement du signal en décomposant celles-ci en opérations élémentaires réalisables par des neurones. Nous avons choisi d'utiliser une implémentation analogique plutôt que numérique des neurones pour des considérations de surface et de consommation. En effet, un neurone analogique peut tirer pleinement parti des équations physiques de l'électronique (intégration du courant aux bornes d'une capacité, sommation des courants, courant de fuite). De plus, un neurone analogique est tout à fait adapté pour effectuer des calculs à faible résolution. Le bruit -souvent problématique en analogique- 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 réduire sa surface et sa consommation tout en restant efficace sur un plan calculatoire [8]. Dans [7], nous avions choisi d'utiliser le modèle de neurone Leaky Integrate-and-Fire dont le comportement est décrit par :

$$\dot{V}_i = -\frac{V_i}{\tau_i} + \sum_{j=1}^{n} W_{ij} s_j(t - \Delta_{ij} t) \qquad (1)$$

avec pour la sortie et la remise à zéro du neurone :

$$\text{si } V_i < V_{th}, \text{ alors } s_i = 0 \qquad (2)$$

$$\text{si } V_i \geq V_{th}, \text{ alors } \begin{cases} s_i = 1 \\ V_i = 0 \end{cases} \qquad (3)$$

où V_i est le potentiel du neurone i, V_{th} est la tension de seuil du neurone, s_j est la sortie du neurone j, τ_i est la constante de temps de la fuite, W_{ij} est le poids synaptique entre le neurone j et i et Δ_{ij} est le délai synaptique. Par la suite, les poids sont normalisés par rapport à V_{th} et sont compris dans l'intervalle [-1 ; 1].

Le modèle LIF permet un grand nombre d'opérations, à savoir l'addition ou la soustraction (par le biais de poids positifs ou négatifs), or encore la multiplication grâce à la fuite [9]. Il a donc été choisi pour ses capacités calculatoires, et sa simplicité qui permet une implémentation compacte et l'analyse de réseaux comportant un grand nombre de neurones. En vue d'être intégrés à grande échelle, les neurones LIF devront avoir une petite taille, une faible consommation et enfin être robuste à la variabilité du procédé de fabrication. D'autres contraintes sur la conception proviennent de l'architecture et des applications envisagées. Il a été déterminé que le poids synaptique des neurones doit être encodé sur 7 bits. Certaines applications nécessitent une fréquence de fonctionnement du neurone élevée afin de diminuer le temps de calcul. La fréquence maximale visée est donc 1 MHz. La fréquence basse correspond à des limites

matérielles détaillées dans la partie 3.1.1. Nous proposons ici un neurone analogique LIF dans une technologie CMOS 65 nm ayant toutes ces caractéristiques. Après la description d'un neurone LIF dans la partie 2, nous détaillons le circuit adopté pour le neurone ainsi que son mode de fonctionnement dans la partie 3. Les résultats de simulations présentent, dans la partie 4, le fonctionnement du neurone ainsi que sa robustesse aux variations de fabrication.

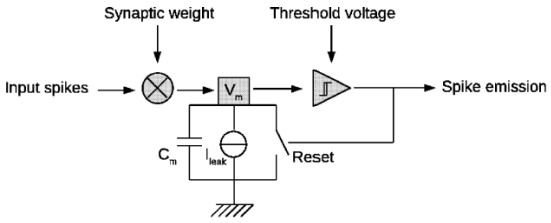

Figure 1. Structure d'un neurone LIF

2. Structure d'un neurone LIF

Un neurone LIF est constitué de cinq parties (fig. 1) à savoir: l'injection des poids synaptiques, le potentiel de membrane (V_M), une fuite, la comparaison avec une valeur seuil et une remise à zéro. L'injection correspond à la modulation pondérée d'une impulsion entrante dans le neurone. Selon le signe du poids, le potentiel V_M stocké dans la capacité C_M augmente ou diminue. Ce potentiel diminue en l'absence d'entrée à cause du courant de fuite I_{leak}. Lorsqu'un nombre suffisant d'impulsions positives sont consécutivement injectées dans le neurone, le potentiel de membrane V_M dépasse la valeur seuil égale à $V_{threshold}$; une impulsion est alors émise vers le(s) neurone(s) post-synaptiques alors que le potentiel V_M est remis au potentiel V_{reset}.

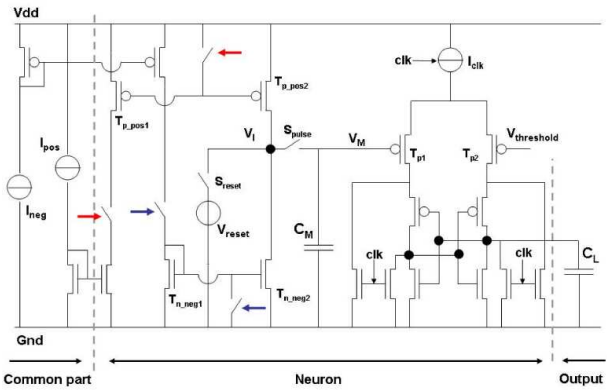

Figure 2. Schéma électronique du neurone

3. Circuit du neurone

Les avancés technologiques dans le domaine de la microélectronique ont eu pour conséquence la réduction de la longueur de grille du transistor. L'utilisation de technologies de dimensions de plus en plus réduites résulte en une augmentation de la variabilité et des courants de fuite, et en une diminution de la tension d'alimentation [10]. Par conséquent, une attention toute particulière a été portée sur le design du neurone pour conserver la robustesse de l'architecture.

3.1 Topologie du circuit

Lors de la conception du neurone, la première étape consiste à déterminer le type et la taille de capacité selon les fréquences de fonctionnement désirées et les contraintes technologiques. On choisira une injection du courant dans la capacité qui réponde aux différents critères : aptitude au calcul, présence d'un étage de comparaison, faible tension d'alimentation (V_{dd}=1.2V) et compacité. Enfin l'étage de comparaison est choisi selon des considérations de vitesse et de consommation.

3.1.1 Choix de la capacité

Les évolutions technologiques ont eu pour conséquence la diminution de l'oxyde de grille. Par conséquent les capacités de type MOS ont une fuite intrinsèque gênante en basses fréquences. Les capacités classiques de type métal/métal ou inter-digitée ont une capacité surfacique faible nécessitant une grande surface. La capacité de type MIM (Métal-Inter-Métal) est une capacité constituée d'une couche d'un matériau à grande permittivité (high-k) entre deux électrodes métalliques. L'ensemble se trouve entre deux couches de métal supérieures du procédé de fabrication. Cette capacité a une faible fuite et également une grande capacité surfacique. En utilisant une armature métallique sous la capacité, les transistors du neurone pourront être placés en dessous de celle-ci. De ce fait, la capacité n'utilise pas de surface supplémentaire dans le design du neurone.

La taille de la capacité a été choisie en considérant les courants de fuite responsables de sa décharge. Il a été choisi une fréquence basse de fonctionnement à 1 kHz pour éviter une taille de capacité trop importante. On peut voir sur la figure 2 les deux principales causes des courants de fuites : le courant de grille dans le transistor T_{p1} et le courant de fuite à travers l'interrupteur utilisé pour l'injection (S_{pulse}). La précision δV est liée à l'amplitude de l'excursion de V_M et la résolution en bits. Elle donne un ordre de grandeur sur la surface (W*L) de T_{p1} et T_{p2}, transistors de la paire différentielle de l'étage de comparaison.

$$\delta V = \frac{V_{threshold} - V_{reset}}{2^{resolution}}$$

Pour que l'injection des poids positifs soit symétrique avec celle de poids négatifs, les transistors T_{p_pos2} et T_{n_neg2} doivent rester en saturation. On considère alors que la dynamique de V_M vaut 0.6 V. Comme la résolution des poids est de 7 bits, la précision sur le comparateur doit-être de 0.6/128 = 4 mV. On peut par conséquent déduire un ordre de grandeur de la surface du transistor T_{p1} grâce au paramètre A_{Vt} (représentant l'erreur faite sur la tension de seuil d'un transistor en mV/µm) et par conséquent estimer son courant de grille.

Pour l'interrupteur S_{pulse}, le courant en dessous du seuil n'est pas considéré. En effet pour des raisons de robustesse à la variabilité, la longueur minimale L de la grille est fixé à 0,2 µm. Par conséquent seules les fuites des diodes polarisées en inverse des transistors sont

considérées et sont proportionnelles à la largeur W des transistors.

Les courants de fuite sont évalués à quelques pA et la capacité à 500 fF. Sa densité surfacique étant de 5 fF/µm², elle occupera une surface de 100 µm². Ceci deviendra par la suite une contrainte de surface pour l'ensemble des transistors du neurone qui seront placé sous la capacité.

3.1.2 Injection

L'injection se fait par le biais d'envoi d'impulsions de courant dans la capacité à l'aide de l'interrupteur S_{pulse} fonctionnant à une fréquence de 500 MHz. Une impulsion correspond au poids minimal que l'on peut injecter dans la capacité ($1/128 = 8.10^{-3}$). Cette injection modulée par une durée d'impulsion est plus robuste à la variabilité (cf précision obtenue par l'utilisation de mini DAC dans [11]). D'autre part, l'interrupteur S_{pulse} est composé de transistors high-Vt pour réduire les fuites lorsqu'il est bloqué.

On a vu que les transistors T_{p_pos2} et T_{n_neg2} doivent être dans leur régime de saturation pour assurer la validité des opérations. En connaissant la taille de la capacité et la fréquence de fonctionnement, on en déduit le courant qu'ils doivent fournir à savoir 1,5 µA. Comme la surface du comparateur dépend de la dynamique de V_M, on polarise les transistors en inversion modérée afin d'obtenir un compromis entre leur taille et la dynamique. La surface est augmentée pour améliorer leur appariement.

3.1.3 Comparaison

Différentes topologies ont été utilisées dans des neurones comme des OTA [4] ou bien des inverseurs [5]. Cependant un meilleur appariement est possible lorsque deux transistors de même type sont utilisés. Ceci privilégie les topologies basées sur une paire différentielle comme les OTA. En considérant les contraintes de consommation, il apparait un grand problème lorsque la valeur à comparer V_M se trouve juste de la valeur seuil $V_{threshold}$. La sortie du comparateur se trouve alors autour de la valeur intermédiaire Vdd/2. Ceci résulte, pour un circuit numérique, à la création d'un chemin de conduction entre Vdd et la masse. Comme un neurone requiert de la précision et une certaine rapidité dans la décision, nous avons choisi d'utiliser un comparateur à bascule. Les transistors d'entrée de la paire différentielle sont de type PMOS pour leur meilleur appariement et leur plus faible courant de grille.

3.2 Mode de fonctionnement

Lorsqu'une impulsion arrive à un neurone, un des deux miroirs de courant (composés de T_{n_neg} ou T_{p_pos} sur la figure 2) est activé suivant le signe du poids (respectivement négatif ou positif) grâce aux interrupteurs commandés par les flèches (bleues ou rouges). Quand le miroir est stable, l'interrupteur S_{reset} est ouvert et l'injection commence dans la capacité via S_{pulse}. 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, un événement logique est généré si $V_M > V_{threshold}$, et envoyé vers d'autres neurones via un mécanisme de routage qui remet le potentiel V_M à la valeur V_{reset}. La comparaison est activée lors d'injection de poids positifs puisque le dépassement de $V_{threshold}$ ne peut avoir lieu lorsque le potentiel V_M est diminué.

4. Résultats

4.1 Exemple d'injection de poids

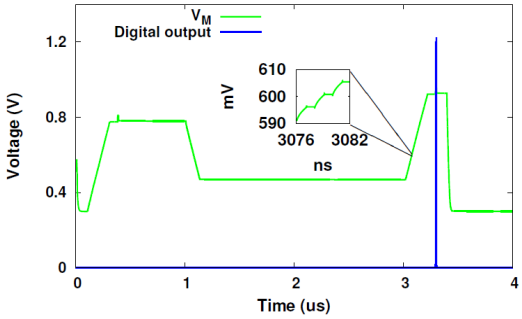

Figure 3. Evolution du potentiel de membrane après injection des poids 0.8, -0.5, 0.8 et sortie logique

La simulation de la figure 3 montre un exemple d'utilisation du neurone. On envoie 3 impulsions successives ayant pour poids 0.8, -0.5 et 0.8, respectivement. Les poids étant encodés sur 7 bits, ceci correspond à un nombre d'impulsions égal respectivement à 102, 64 et 102. La figure 3 montre l'évolution du potentiel V_M et la sortie du neurone. Au début de la simulation une étape d'initialisation a lieu pour mettre le potentiel V_M à V_{reset}. La première impulsion (poids : 0.8) augmente le potentiel V_M jusqu'à une valeur inférieure à $V_{threshold}$. Une comparaison est effectuée, visible par un faible pic de tension sur le potentiel membranaire, mais sans effet sur le fonctionnement du neurone. Une deuxième impulsion (poids : -0.5) arrive et, du fait de son poids négatif, diminue le potentiel aux bornes de V_M. Enfin une troisième impulsion (poids : 0.8) arrive et permet de dépasser la valeur seuil (0.8–0.5+0.8=1.1). Il y a, après comparaison, génération d'un événement logique à 3.3 µs.

4.2 Robustesse à la variabilité des procédés de fabrication.

Afin d'étudier la robustesse du neurone face à la variabilité du procédé de fabrication, nous souhaitons analyser les variations globales et locales. 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. Ces deux types de variations sont étudiés dans les deux parties suivantes.

4.2.1 Variations locales dans le corner nominal (TT)

L'histogramme de la figure 4 montre l'influence des variations locales entre transistors dans un cas où celles globales sont nominales. Pour l'obtenir, des simulations ont été réalisées pour tenir compte de la variabilité (simulations Monte-Carlo). 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 une sortie lorsque le poids vaut 1 (128 impulsions). En pratique on observe une distribution des neurones dont la sortie est active centrée autour de 128. La distribution peut être modélisée par une loi normale dont l'espérance μ = 128.2 et l'écart type σ = 3.9. Ceci correspond à un coefficient de variation de 3% qui est équivalent à un rapport signal à bruit de 34.91 dB. En choisissant des poids adaptés lors de la programmation de l'architecture, on peut réduire l'impact des variations locales (voir [7] pour plus de détails sur ce sujet).

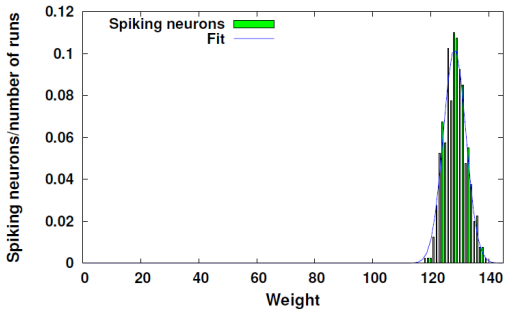

Figure 4. Histogramme vert : nombre de neurones ayant émis une sortie à un poids donné. Courbe bleu : fit par une loi normale (μ = 128.2 et σ = 3.9)

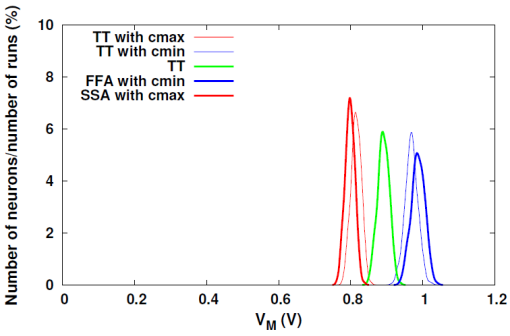

Figure 5. Distribution du potentiel de membrane avec différents corners

4.2.2 Comparaison entre différents corners : SSA-TT-FFA

Dans la figure 5, la forme de chaque courbe représente les variations du potentiel membranaire V_M lorsque le neurone reçoit une impulsion pondéré d'un poids égal à 1 (correspondant à 128 impulsions de S_{switch}). 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.

5. Conclusion

Nous avons proposé ici une implémentation d'un neurone analogique de type Leaky Integrate-and-Fire dont le but est d'être inclus dans une architecture de calcul robuste à la variabilité. Il est à la fois compact (100 μm^2), supporte l'injection de poids positifs et négatifs, et a un rapport signal à bruit de 35 dB en présence de variations due au procédé technologique. L'implémentation d'une fuite programmable est prévue afin d'augmenter les possibilités d'utilisation du neurone comme élément de calcul. Comme le neurone est inclus dans un circuit mixte analogique et numérique, des écarts de température peuvent exister, qui nécessiteront une étude poussée du comportement du neurone en fonction de la température.

Remerciements

Les auteurs tiennent à remercier D. Morche, A. Peizerat, M. Belleville et A. Valentian pour leurs avis et leurs précieux conseils.

Références

[1] M. Muller, "Dark silicon and the internet," in *Designing with ARM, EE Times Virtual Conference*, March 2010.

[2] S. Borkar, "Designing Reliable Systems from Unreliable Components: The Challenges of Transistor Variability and Degradation," *IEEE Micro*, vol. 25, no. 6, pp. 10–16, Nov. 2005.

[3] C. Mead, *Analog VLSI and Neural Systems*. Addison-Wesley, 1989.

[4] A. Van Schaik, "Building blocks for electronic spiking neural networks," *Neural Networks*, vol. 14, no. 6-7, pp. 617–628, 2001.

[5] R. J. Vogelstein, U. Mallik, J. T. Vogelstein, and G. Cauwenberghs, "Dynamically reconfigurable silicon array of spiking neurons with conductance-based synapses." *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, vol. 18, no. 1, pp. 253–65, 2007.

[6] K. A. Boahen, "Point-to-point connectivity between neuromorphic chips using address events," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 47, no. 5, pp. 416–434, May 2000.

[7] R. Heliot, A. Joubert, and O. Temam, "Robust and low-power accelerators based on spiking neurons for signal processing applications," in *The 3rd HiPEAC Workshop on Design for Reliability (DFR'11)*, 2011.

[8] R. Sarpeshkar, "Analog versus digital: extrapolating from electronics to neurobiology," *Neural Computation*, vol. 10, no. 7, pp. 1601–1638, 1998.

[9] M. Srinivasan and G. Bernard, "A proposed mechanism for multiplication of neural signals," *Biological Cybernetics*, vol. 21, no. 4, pp. 227–236, 1976.

[10] A. Annema, B. Nauta, R. van Langevelde, and H. Tuinhout, "Analog circuits in ultra-deep-submicron CMOS," *IEEE Journal of Solid-State Circuits*, vol. 40, no. 1, pp. 132–143, Jan. 2005.

[11] B. Linares-Barranco, T. Serrano-Gotarredona, and R. Serrano- Gotarredona, "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*, vol. 14, no. 5, pp. 1207–16, Jan. 2003.