Introduction

L'invention du transistor à jonctions, en 1947, est certainement l'innovation la plus importante du XX^{e} siècle, tant l'ensemble de notre vie quotidienne en dépend. Depuis cette date, nous y reviendrons, le monde est devenu « numérique » et la presque totalité de l'information est traitée sous forme binaire par des microprocesseurs.

Pour arriver au monde numérique que nous connaissons, quelques étapes ont été essentielles comme la fabrication du premier circuit intégré en 1958. Il est rapidement apparu que les circuits intégrés permettaient non seulement de traiter les signaux analogiques, comme ceux utilisés en radio, mais aussi des signaux digitaux. De tels circuits digitaux ont été utilisés dans la mission Apollo XI qui a conduit l'Homme sur la lune, le 21 juillet 1969. Pour réaliser cet exploit fabuleux, les astronautes ne disposaient que de moyens de calcul très réduits. Le contrôleur de vol était une machine que nous pouvons considérer aujourd'hui comme très rudimentaire. Constituée de 2 800 circuits intégrés, comportant chacun deux portes « NON-OU » à trois entrées, d'une mémoire RAM[1] de 2 048 mots et d'une mémoire ROM[2] de 38 000 mots pour les programmes, elle travaillait à une fréquence d'horloge de 80 kHz et ne pesait pas moins de 32 kg pour 55 W de puissance consommée.

L'exploit a donc essentiellement reposé sur du traitement « humain » ou « cortical » de l'information : la puissance de calcul, trop souvent mise en avant aujourd'hui, n'est pas toujours la condition *sine qua non* du succès !

Afin de réduire le poids des systèmes de traitement, tout en améliorant leurs performances, il est nécessaire d'intégrer un grand nombre de portes logiques sur le même circuit. Cette voie de l'intégration a conduit à une véritable révolution en 1971 :

1. Où l'on peut lire et écrire.
2. Que l'on ne peut que lire.

l'élaboration du premier microprocesseur. Depuis cette date, la progression des technologies digitales de traitement de l'information a été formidable, tant sur le plan de leurs performances techniques que sur celui de leur impact dans la société.

Le monde dans lequel nous vivons est devenu celui d'un « déluge de données » (*data deluge*), terme inventé pour décrire la croissance massive du volume de données générées, traitées, stockées par les médias numériques (audio et vidéo), les transactions commerciales, les réseaux sociaux, les bibliothèques numériques, etc. À chaque minute par exemple, le réseau Internet gère près de 200 millions de courriels, 40 millions de messages vocaux, 20 millions de messages texte sans compter 500 000 « gazouillis » (*tweets*). En 2016, la taille de l'univers numérique, défini comme la quantité de données créées, numérisées et stockées par les êtres humains, était estimée à 16 ZB[3] (zettaoctets) et les prévisions sont d'un doublement tous les deux ans soit environ 44 ZB en 2020 et 160 ZB en 2025. Quel chemin parcouru en un demi-siècle !

Cette progression, symbolisée par la fameuse « loi » de Moore[4] qui prédisait un doublement de la puissance[5] des microprocesseurs tous les 18 mois, s'est effectuée à prix constant, c'est-à-dire que le prix d'un microprocesseur moderne est sensiblement le même que celui du microprocesseur de 1971 alors même que les performances ont été améliorées de plus de cinq ordres de grandeur.

Cette évolution remarquable n'a été possible que grâce à l'existence d'un modèle universel des machines de traitement de l'information, la machine de Turing, et d'une technologie apte à implémenter ces machines physiquement, celle des dispositifs à semi-conducteurs. Plus précisément, le triplet « codage binaire-architecture de von Neumann-technologie CMOS » constitue le modèle dominant des systèmes de traitement de l'information depuis le début des années 1970.

Cependant, deux limites sont aujourd'hui atteintes : celle de la miniaturisation, avec des dispositifs dont la taille ne dépasse pas quelques nanomètres, et celle de la puissance dissipée avec une barrière de l'ordre de la centaine de Watts lorsque le processeur travaille intensément.

Tant que les performances s'amélioraient régulièrement, la recherche de nouveaux paradigmes de traitement de l'information n'était pas prioritaire. Avec la saturation prévisible, à moyen terme, des performances des processeurs, mais aussi avec l'émergence de nouveaux domaines applicatifs, comme les objets connectés ou l'Intelligence artificielle, la question d'un paradigme de traitement de l'information possédant à la

3. Zetta = 10^{21} et un octet est un ensemble de 8 bits.

4. Gordon Moore est l'un des fondateurs de la compagnie Intel en 1968.

5. Représentée par le nombre de portes logiques par circuit.

fois *i)* une grande efficacité énergétique et *ii)* des performances supérieures aux systèmes actuels pour résoudre certains types de problèmes, réapparaît de façon prégnante.

Cet ouvrage, consacré au traitement neuro-inspiré[6] de l'information, s'inscrit dans le cadre de cette réflexion. Son objectif est de donner aux étudiant(e)s ou aux chercheur(e)s intéressé(e)s par ce thème passionnant, une vue générale des connaissances et de l'état de l'art sans oublier de les sensibiliser aux innombrables questions qui se posent et aux problèmes qui restent ouverts.

Associant les neurosciences, l'informatique, la physique des semi-conducteurs, la conception de circuits, mais aussi les mathématiques et la théorie de l'information, le sujet abordé est fortement pluridisciplinaire.

Pour permettre une lecture fluide du texte, les notions de base sont souvent rappelées ou renvoyées à des documents donnés en référence. Chaque fois que possible, des modèles mathématiques des phénomènes étudiés sont proposés, afin de permettre une analyse, certes simplifiée, mais quantitative de l'influence des différents paramètres. Ce support à la réflexion à l'aide de formulations analytiques est à notre sens *la* condition à une bonne compréhension de la physique des phénomènes en jeu.

L'ouvrage est organisé en 4 chapitres pratiquement indépendants :

– le premier présente les notions de base du traitement électronique de l'information, en particulier le codage, la mémorisation, l'architecture des machines, et la technologie CMOS qui constitue le support matériel de ce traitement. Comme l'un des objectifs de l'ouvrage est d'approfondir le lien entre traitement de l'information et consommation énergétique, différentes voies d'amélioration des performances des systèmes actuels sont présentées, en particulier le traitement neuro-inspiré, sujet central de l'ouvrage. Une comparaison assez générale des principes de fonctionnement et des performances d'un microprocesseur moderne et du cerveau est également présentée dans ce chapitre ;

– le second chapitre est consacré aux principes connus du fonctionnement du cerveau et en particulier ceux du cortex cérébral également appelé « matière grise ». Dans cette partie, l'approche est descendante ou *top-down*, c'est-à-dire que le cortex est d'abord vu de façon globale et fonctionnelle avant l'étude de son organisation en unité de base du traitement que sont les colonnes corticales. Un exemple emblématique, celui de la vision et du cortex visuel, est également décrit afin d'illustrer ces différents aspects fonctionnels ;

6. Également appelé « bio-inspiré ».

– le troisième chapitre explore en détail les neurones et les synapses, qui sont les briques de base du traitement de l'information dans le cortex. À partir d'une analyse approfondie des principes physiques régissant les propriétés des membranes biologiques, différents modèles mathématiques de neurones sont décrits allant des plus complexes aux modèles phénoménologiques les plus simples. À partir de ces modèles, la réponse des neurones et synapses à divers stimuli est également décrite. Ce chapitre explore également les principes de la propagation des potentiels d'action ou *spike* le long de l'axone et il précise comment peuvent être introduites certaines règles d'apprentissage dans les modèles de synapses ;

– le quatrième et dernier chapitre traite des réseaux de neurones et synapses artificiels. Les deux grandes approches de réalisation de ces réseaux, logicielle ou matérielle, sont présentées ainsi que leurs performances respectives. Un état de l'art de chaque approche est également donné. Nous montrons dans ce chapitre l'intérêt de la voie matérielle pour la conception et la fabrication de réseaux de neurones et synapses artificiels à ultra faible puissance et énergie consommée, et des exemples de réseaux de neurones artificiels allant du plus simple au plus complexe sont décrits.