

> des spins portés par les électrons du courant électrique avec des matériaux magnétiques.

Concrètement, les briques de base de la spintronique sont des petits cylindres d'une dizaine de nanomètres de diamètre formant des «jonctions tunnel magnétiques». Ces jonctions sont constituées de deux couches de matériaux magnétiques (des nanoaimants) séparées par une couche isolante. Quand les deux nanoaimants sont orientés dans le même sens, un courant électrique traverse facilement (par l'«effet tunnel», un effet quantique) la couche isolante de la jonction: la résistance est faible. Si les orientations sont opposées, le courant circule difficilement: la résistance est alors élevée.



L'apprentissage d'un réseau de neurones comme BERT consomme 1000 kilowattheures

Ces propriétés permettent de stocker de l'information dans un format binaire («0» ou «1») selon les deux cas d'orientation possible (voir l'encadré page 34). Les composants de mémoire ainsi formés sont nommés MRAM (*magnetoresistive random-access memories*). Ils ont commencé à être développés vers le milieu des années 1980. Cependant, pour changer la valeur d'un bit, il fallait appliquer un champ magnétique externe afin d'inverser l'orientation d'un nanoaimant. C'était un obstacle à la miniaturisation de ces composants.

En 1996, John Slonczewski, chercheur chez IBM, et Luc Berger, de l'université Carnegie-Mellon, aux États-Unis, ont indépendamment proposé une amélioration importante de ces composants. Ils ont mis en évidence un nouvel effet, le transfert de spin. Lorsque les électrons d'un courant traversent un nanoaimant de la jonction, leur spin interagit avec celui des électrons dans la couche: le spin des électrons du courant se polarise et s'aligne avec l'aimantation de la couche. Dans une jonction, l'aimantation d'un des deux nanoaimants est fixe tandis que l'autre peut être modifiée. Quand un courant polarisé traverse cette dernière couche, elle réoriente l'aimantation. Selon le sens de circulation du courant, il est possible d'écrire un bit «0» ou un bit «1», et cela sans faire appel à un champ magnétique extérieur.

Cette découverte a ouvert la porte à la conception de MRAM intégrées.

Depuis, plusieurs fabricants de systèmes microélectroniques produisent des dispositifs de mémoire qui exploitent cette technique. Ils peuvent intégrer un milliard de ces composants sur une puce de silicium. Ces mémoires trouvent déjà des applications en électronique usuelle, car elles associent les qualités des deux grandes familles de mémoires informatiques: la vitesse de lecture et d'écriture des mémoires vives et la capacité des mémoires de stockage à conserver durablement les informations.

On se rapproche de l'idée de puces neuromorphiques, car il est possible d'insérer facilement des jonctions tunnel magnétiques aux endroits les plus utiles pour stocker des données sous forme de bits au sein des circuits en silicium qui réalisent les opérations de calcul. Comme dans le cerveau, où les synapses sont au contact des neurones, ces synapses artificielles seront ainsi au plus près du cœur de calcul en silicium. Si l'on veut utiliser un algorithme de réseau de neurones dans ce dispositif, il suffit donc de configurer toutes ces synapses avec les bonnes valeurs pour définir la tâche du programme. Selon le jeu de valeurs mis en place, le réseau sera en mesure d'identifier des éléments dans des images, du son...

En réduisant drastiquement les échanges de données entre mémoire externe et processeur, il est possible d'exécuter des algorithmes d'intelligence artificielle en dépensant moins d'énergie. Cette piste est activement explorée dans les laboratoires académiques comme industriels. Elle n'est cependant pas entièrement satisfaisante: il faut connaître à l'avance les bonnes valeurs à mémoriser par les synapses pour effectuer une tâche. Le dispositif n'a pas les capacités d'adaptation d'un vrai cerveau. On souhaite, idéalement, un circuit qui soit en mesure d'apprendre à réaliser de nouvelles opérations.

INITIALISER LES SYNAPSES

Dans les algorithmes de réseaux de neurones, on a souvent une première phase, dite «d'apprentissage», où le réseau ajuste la valeur des synapses afin d'exécuter de mieux en mieux la tâche qui lui est assignée. Malheureusement, la technique par laquelle l'algorithme identifie les bonnes valeurs des synapses dans les réseaux de neurones artificiels est fondée sur une méthode mathématique complexe, la «rétropropagation du gradient» (voir l'encadré page 30). Celle-ci exige des calculs nombreux, très précis et donc gourmands en énergie. Elle est bien adaptée sur un ordinateur, mais pas nécessairement pour un système inspiré du cerveau fonctionnant avec des calculs moins précis.

Pour rester dans la logique des systèmes neuromorphiques, au lieu de reproduire cette méthode énergivore, il serait plus judicieux de

LES DIFFÉRENTES APPROCHES NEURO- MORPHIQUES

Dans un ordinateur usuel, les données sont stockées dans une mémoire physique séparée du processeur où sont exécutés les calculs ; on parle d'« architecture de von Neumann ». Ce système simple à mettre en œuvre est cependant énergivore et sa vitesse est limitée. Pour rapprocher les mémoires des processeurs, la spintronique est une piste, mais il existe d'autres possibilités.



SPINTRONIQUE

Cette technique repose sur des jonctions tunnel magnétiques. La résistance d'une jonction, qui dépend de l'orientation de l'aimantation de ses deux couches magnétiques, encode des bits « 0 » et « 1 ». Ces composants peuvent être configurés pour agir comme des synapses ou des neurones artificiels.

- + Physique particulièrement riche et comprise
- Fabrication difficile



TRANSISTORS

Il est possible de concevoir des neurones et des synapses avec des transistors au silicium, mais pour réaliser un unique neurone, il faut des centaines de transistors. Cette idée a été mise en application dans les puces TrueNorth, d'IBM, et Loihi, d'Intel.

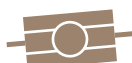
- + Accessible dès aujourd'hui
- Mémoires extrêmement coûteuses



MEMRISTORS

Ces composants électroniques, par exemple en dioxyde de titane, ont une résistance qui dépend du courant qui les traverse. Leur état est fonction de l'historique des courants qui ont déjà été appliqués. Ces mémoires artificielles exhibent ainsi une certaine capacité de plasticité synaptique.

- + Fabrication simple
- + Composants bon marché
- Fiabilité réduite



MÉMOIRES À CHANGEMENT DE PHASE

Ces composants contiennent des matériaux, tels que des verres de chalcogénure, qui basculent d'une phase cristalline à amorphe sous l'effet de la chaleur. La résistance électrique des phases amorphes est supérieure à celle des phases cristallines, ce qui permet d'encoder des bits « 0 » et « 1 ».

- + Technologie mature
- Consommation énergétique



OPTIQUE

Dans un ordinateur optique, les photons émis par des lasers remplacent les électrons. Le signal lumineux est directement traité grâce à des composants optiques qui sont plus rapides que les transistors électriques. Il est également possible de concevoir des dispositifs de mémoire avec ces composants optiques.

- + Calculs très rapides
- Mise en œuvre complexe (miniaturisation difficile des composants)
- Forte consommation énergétique