

DURÉE LIBRE

FORMULE INTÉGRALE

Au lieu de
~~8,20 €~~
par mois

Partie réservée au service abonnement Ne rien inscrire

L'ESSENTIEL

> En exploitant le spin, une propriété quantique des électrons, il est possible de concevoir des composants électroniques aux caractéristiques inédites.

> Grâce à la spintronique, il est envisageable de placer des unités de mémoire au plus près des processeurs, à l'image

des synapses et des neurones du cerveau.

> Les premiers prototypes ont démontré que les dispositifs reposant sur la spintronique sont peu énergivores et ont des performances comparables aux algorithmes de réseaux de neurones usuels.

LES AUTEURS



JULIE GROLIER
chercheuse à l'unité mixte de physique CNRS/Thales, à Palaiseau



DAMIEN QUERLIOZ
chercheur au Centre de nanosciences et de nanotechnologies (université Paris-Saclay/CNRS), à Palaiseau

Quand la spintronique s'inspire du cerveau

En termes d'énergie, le cerveau est beaucoup plus efficace que les ordinateurs actuels pour mémoriser, calculer ou apprendre. Est-il possible de s'inspirer de cet organe pour réaliser des dispositifs plus performants ? Oui, en s'appuyant sur le spin des électrons...

Grâce aux jonctions tunnel magnétiques, il est possible de concevoir des neurones et des synapses artificiels. Économes en énergie, ils pourraient avantageusement remplacer les algorithmes de réseaux de neurones.



Les algorithmes d'intelligence artificielle ont fait des progrès considérables ces dernières années. Ils sont de plus en plus présents dans notre quotidien et sont utilisés pour analyser le contenu de textes ou d'images, pour reconnaître des instructions vocales, et souvent avec une efficacité supérieure aux humains. Ainsi, en 2017, le programme *AlphaGo* a battu le champion du monde coréen Ke Jie au jeu de go, un exploit qu'aucun expert n'attendait avant au moins dix ans.

Cependant, pour parvenir à de telles performances, ces algorithmes consomment des quantités très élevées d'électricité, beaucoup plus que notre cerveau pour des tâches équivalentes. Par exemple, les modèles d'intelligence artificielle de traitement du langage humain, comme BERT (ou sa version française FlauBERT) nécessitent une énergie de mille kilowattheures pour être entraînés. C'est autant que consomme le cerveau humain en six ans !

L'intelligence artificielle moderne, qui s'appuie sur des algorithmes réalisant des réseaux de neurones artificiels dits « profonds », s'inspire pourtant partiellement de la structure du cerveau (*voir l'encadré ci-contre*). Pourquoi donc un tel contraste dans les performances ? Les programmes d'intelligence artificielle tournent aujourd'hui sur des ordinateurs classiques, ou sur leurs cousins les cartes graphiques, dont l'architecture et le principe de calcul sont bien différents de ceux du cerveau.

Pour résoudre ce défi énergétique, de nombreux chercheurs et industriels développent de nouvelles approches visant une électronique dont le fonctionnement serait vraiment calqué sur celui du cerveau. C'est le cas de la spintronique, où l'on exploite certaines propriétés quantiques des électrons. Des dispositifs reposant sur cette technique sont capables de simuler le comportement des neurones du cerveau et des synapses qui les relient. Nous sommes encore loin du stade de l'industrialisation d'une solution complète fondée sur la spintronique, mais les premiers résultats sont déjà très prometteurs.

UNE SÉPARATION CALCUL-MÉMOIRE ÉNERGIVORE

Dans les ordinateurs actuels, l'obstacle majeur pour économiser de l'énergie et gagner en vitesse est lié au fait que les informations sont stockées dans une mémoire physiquement séparée des processeurs, où les données sont transformées en suivant les instructions du programme. Or les algorithmes d'intelligence artificielle comportent des millions, voire des milliards, de paramètres. Lorsqu'un tel programme opère, il doit sans cesse chercher des informations en mémoire puis les additionner ou les multiplier dans le processeur, avant de

les replacer dans la zone de stockage. Ce flux permanent de données par une voie de communication unique consomme une quantité d'électricité considérable.

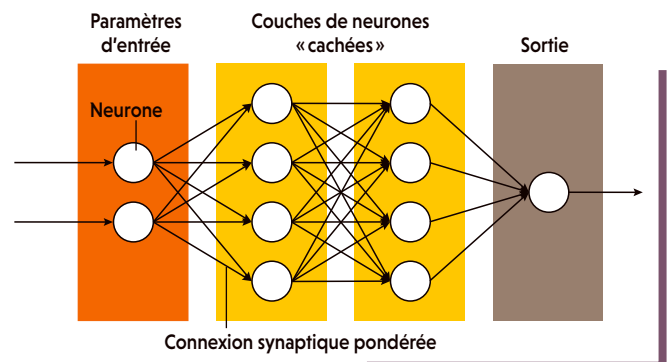
Dans le cerveau, le fonctionnement est très différent. La mémoire est partout : elle est distribuée et stockée dans les synapses, les connexions entre les neurones qui, eux, exécutent des opérations sur les données. Chaque neurone est connecté en moyenne à dix mille synapses, ce qui permet notamment de réaliser des calculs en parallèle. Cette juxtaposition entre calcul et mémoire, doublée d'un traitement parallèle, est d'une redoutable efficacité énergétique.

La nature même du calcul est une autre source de gain considérable d'énergie. Nos ordinateurs manipulent des données très précises, chacune étant codée sous la forme de bits d'information. Les opérations sur ces données

LES ALGORITHMES DE RÉSEAUX DE NEURONES

Les « réseaux de neurones » de l'intelligence artificielle, en particulier ceux dits « d'apprentissage profond », ne sont pas des structures physiques, mais des algorithmes dont le principe s'inspire du fonctionnement du cerveau. Ils représentent des nœuds, les neurones, reliés entre eux par des synapses. Dans le cas de l'apprentissage profond, les neurones sont organisés en couches. Le nombre de couches et de neurones par couche dépend de la tâche à accomplir par le réseau : analyse d'images, de textes, de sons... Le signal d'entrée est envoyé aux neurones de la première couche, qui transmettent un signal à ceux de la suivante, etc.

Au niveau de chaque neurone, les signaux issus de la couche précédente sont additionnés suivant une somme pondérée par les coefficients portés par les synapses, coefficients qui définissent la « fonction synaptique ». Cette dernière est configurée grâce à une phase d'apprentissage du réseau de neurones. On soumet des données en entrée du réseau, qui effectue un calcul et compare la solution obtenue à celle qu'on lui fournit. Prenons par exemple un réseau de neurones formé à reconnaître des images de chats. Si, pour une photo donnée, le réseau se trompe, les pondérations synaptiques sont corrigées par une méthode mathématique nommée « rétropropagation du gradient ». Celle-ci s'appuie sur des outils statistiques pour calculer la variation d'une erreur au niveau d'un neurone, de la dernière couche vers la première. Après un certain nombre d'itérations, la fonction synaptique est alors correctement configurée pour que le réseau de neurones reconnaisse les images de chats avec un taux d'erreurs très faible.



se font suivant des opérations logiques d'une rigueur implacable. Cette précision, assurément cruciale pour de nombreuses applications comme envoyer une fusée dans l'espace, ne l'est pas pour d'autres, comme reconnaître un visage familial. Or le fait d'imposer à un circuit électronique un comportement parfait où l'aléa n'est pas toléré a un coût élevé en énergie, car il faut éliminer tous les bruits parasites et corriger toutes les erreurs.

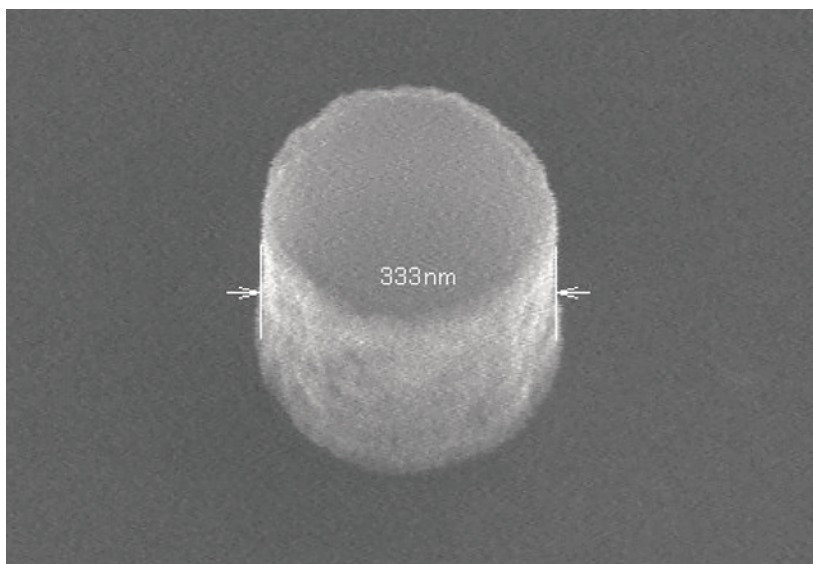
Au cours de l'évolution, le cerveau, qui ne dispose que d'une source d'énergie limitée – nos apports caloriques quotidiens – a développé une tout autre stratégie. Il fonctionne sur la base d'un compromis entre la fiabilité du traitement de l'information et la dépense énergétique. Les synapses et les neurones ont ainsi des comportements stochastiques: en intégrant une certaine marge d'aléatoire, ils ne donnent pas toujours un résultat identique même s'ils reçoivent des informations similaires.

En contrepartie, pour assurer tout de même un traitement fiable de l'information, le cerveau met en œuvre de multiples stratégies. Par exemple, comme chaque neurone, pris individuellement, n'est pas très fiable, plusieurs de ces cellules cérébrales sont parfois chargées de traiter le même signal. Un résultat cohérent et stable est ensuite déduit de l'ensemble de leurs réponses. La redondance est, en fin de compte, moins coûteuse en énergie qu'un système trop rigide qui corrige systématiquement les erreurs.

De plus, les signaux stockés dans les synapses et traités par les neurones ne sont pas des «0» ou des «1» (codage binaire) comme dans l'électronique actuelle. Ces signaux présentent une grande gamme de valeurs différentes et très proches: ils contiennent ainsi bien plus d'information, mais sont plus sensibles aux perturbations, autrement dit au «bruit». Enfin, contrairement aux processeurs où chaque élément opère au rythme d'une horloge unique, assurant un fonctionnement cohérent, mais limitant la vitesse du traitement de l'information, le cerveau travaille de façon asynchrone: chaque neurone émet des signaux électriques selon son propre rythme et s'en sert pour communiquer efficacement avec d'autres neurones, qui se trouvent parfois dans une zone différente du cerveau.

LE CERVEAU, SOURCE D'INSPIRATION

Ces spécificités contribuent à rendre le cerveau plus économe en énergie. Est-il possible de s'en inspirer pour construire des composants électroniques ou puces capables de traiter des données avec très peu d'énergie? Avec des neurones et des synapses électroniques installés au plus près, de façon à rapprocher mémoire et calcul, il serait possible de reproduire



Les jonctions tunnel magnétiques (telle celle-ci, vue au microscope électronique à balayage) sont fabriquées en plusieurs étapes. On fait croître les couches de matériaux l'une après l'autre par pulvérisation cathodique dans une enceinte sous vide. Ensuite, pour obtenir des jonctions d'un diamètre inférieur au micromètre, le matériau est recouvert d'une résine électrosensible et d'un masque sur lequel est dessinée la forme de la jonction. L'ensemble est soumis à une source d'électrons (lithographie) et la résine exposée est conservée. Les parties non protégées par la résine du matériau sont gravées et retirées pour ne garder que la forme de la jonction.

l'architecture équivalente à celle du réseau de neurones que l'on veut faire fonctionner. Cette idée n'est pas récente. Dès la fin des années 1980, Carver Mead, chercheur à Caltech (l'institut de technologie de Californie) et l'un des pères fondateurs de la microélectronique moderne, a inventé le concept de circuits «neuromorphiques», dont l'objectif est de mimer l'architecture neurobiologique du cerveau. Cette approche a longtemps été délaissée, car les obstacles technologiques étaient nombreux.

Le premier obstacle concerne la mémoire: les techniques de l'électronique actuelle ne permettent pas aujourd'hui d'intégrer d'importantes quantités de mémoire au plus près du calcul. En effet, les seuls circuits de mémoire qui peuvent être intégrés dans les processeurs (les circuits de mémoire vive statique) sont extrêmement coûteux, car ils utilisent une grande et précieuse surface dans le processeur. C'est pour cette raison que les processeurs n'intègrent que de très faibles capacités de stockage, la plus grande part de la mémoire étant physiquement distribuée en dehors des zones de calcul. Plusieurs pistes sont néanmoins à l'étude pour rapprocher mémoire et processeurs (voir l'encadré page 33). Parmi elles, l'électronique de spin, ou spintronique, a connu des progrès spectaculaires ces dernières années.

Cette technique n'exploite pas uniquement les électrons au travers de leur charge électrique, mais utilise aussi leur spin, une propriété purement quantique, assimilable à un moment cinétique (on peut représenter le spin d'un électron comme une petite flèche orientée). Comme le spin est responsable du moment magnétique intrinsèque de l'électron, quand une majorité d'électrons ont un spin pointant dans une direction privilégiée dans un matériau, ce dernier a une aimantation non nulle. La spintronique exploite les interactions ➤

> 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 1 000 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.

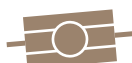
- + Accessible dès aujourd'hui
- Mémoires extrêmement coûteuses



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

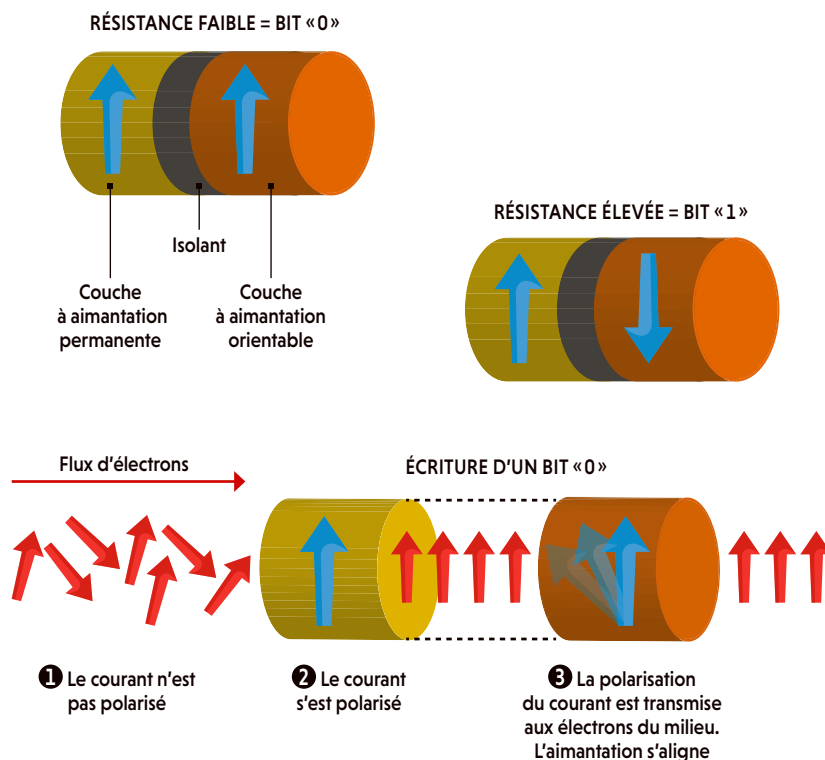
MANIPULER LES SPINS DES ÉLECTRONS

Dans les circuits électroniques actuels, la manipulation de la charge électrique de l'électron est au cœur du fonctionnement des dispositifs, à l'instar des transistors. En électronique de spin, ou spintronique, on utilise, en plus de la charge, une autre propriété quantique des électrons, le spin. Celui-ci correspond à un moment magnétique intrinsèque de la particule. Les interactions de ce spin avec certains matériaux magnétiques offrent la possibilité de concevoir des composants aux comportements très riches. Dans un matériau magnétique, comme le fer ou le cobalt, le spin des électrons est globalement orienté dans une même direction, ce qui confère au matériau son aimantation. Lorsqu'un courant électrique traverse un tel matériau, le spin des électrons qui constituent le courant s'oriente dans la direction de l'aimantation. Cet effet donne lieu à deux phénomènes fondamentaux en spintronique.

Le premier est le phénomène de magnétorésistance observé dans les « jonctions tunnel magnétiques ». Ces composants sont constitués de deux couches aimantées séparées par une couche isolante. Les matériaux utilisés dépendent des propriétés que l'on veut donner à la jonction. L'isolant est souvent un oxyde métallique et son épaisseur est d'environ 1 à 2 nanomètres. L'aimantation d'une couche magnétique est fixe tandis que l'autre peut être réorientée.

En 1975, Michel Jullière, de l'Insa de Rennes, a été le premier à mettre en évidence la magnétorésistance à effet tunnel. Lorsqu'un courant électrique traverse une jonction tunnel magnétique, la traversée de la couche isolante s'effectue grâce à l'effet tunnel (un phénomène purement quantique). Le courant passe plus facilement si les aimantations des deux couches magnétiques pointent dans la même direction, et plus difficilement si les aimantations sont orientées dans des directions opposées.

Ainsi, en mesurant la résistance électrique (faible ou élevée) d'une jonction tunnel magnétique, on peut « lire » la direction respective des aimantations. On associe ainsi un bit « 0 » à des aimantations parallèles et un bit « 1 » à des aimantations antiparallèles. Ces jonctions tunnel magnétiques peuvent ainsi servir d'élément de base dans des systèmes de mémoire pour stocker des bits d'information.



Une importante activité industrielle s'est développée autour des mémoires reposant sur ce principe et qualifiées de MRAM (*magnetoresistive random-access memory*).

La découverte d'un effet analogue dans des systèmes multicouches métalliques, la magnétorésistance géante, en 1988, a été décisive pour l'intérêt porté au développement de la spintronique, et a valu le prix Nobel de physique en 2007 à Albert Fert, de l'université Paris-Saclay, et Peter Grünberg, de l'université de Cologne. Cette technique et la magnétorésistance tunnel sont utilisées notamment dans les têtes de lecture des disques durs modernes.

Le fait d'avoir une couche dont l'aimantation est « mobile » permet de réorienter cette dernière pour que la jonction porte un bit « 0 » ou « 1 ». Pour écrire ce bit, l'approche initiale a été d'utiliser un champ magnétique extérieur, mais cette méthode ne permet pas une bonne réduction de la taille du système et son intégration à des échelles nanométriques.

En 1996, John Slonczewski, chercheur chez IBM, et Luc Berger, de l'université Carnegie-Mellon, aux États-Unis, ont mis en évidence le second phénomène crucial de la spintronique : le transfert de spin. Lorsqu'un courant électrique suffisamment intense traverse une jonction tunnel magnétique, il peut faire basculer l'aimantation de la couche à aimantation mobile pour obtenir une configuration parallèle ou antiparallèle. Pour écrire

un bit « 0 » (voir ci-dessus), le courant traverse d'abord la couche à aimantation fixe. Il se polarise : le spin des électrons du courant, qui n'avait initialement pas d'orientation particulière, s'aligne avec l'aimantation de la couche. Quand les électrons arrivent dans la seconde couche, ils interagissent avec les électrons du matériau et leur transfèrent leur polarisation. L'aimantation de cette couche devient progressivement parallèle à l'autre.

Pour être plus précis, à chaque interface (entre une couche magnétique et l'isolant), une partie des électrons est réfléchi et une autre est transmise. Pour écrire un bit « 1 », le courant arrive par la couche à aimantation mobile. Il se polarise d'abord selon l'aimantation de cette couche. Mais les électrons qui sont réfléchis sur la couche à aimantation fixe prennent une polarisation un peu plus antiparallèle et transmettent cette orientation à la couche à aimantation mobile. Résultat : on obtient une configuration antiparallèle, correspondant à un bit « 1 ».

Depuis 2008, de nombreux industriels exploitent les jonctions tunnel magnétiques à transfert de spin pour concevoir des dispositifs de mémoire performants.

> s'inspirer encore une fois de la façon dont le cerveau apprend. En particulier, les synapses biologiques ne servent pas juste comme mémoire, elles sont aussi au cœur du mécanisme de l'apprentissage. Elles sont en effet «plastiques»: elles ajustent leur activité selon la quantité d'information à laquelle elles sont exposées. Par exemple, quand deux neurones connectés par une synapse sont actifs en même temps, cette synapse s'en trouve renforcée avec une activité accrue. Mais ce renforcement n'est pas toujours complet, sinon, d'une certaine façon, chaque nouveau signal effacerait la subtile configuration de la synapse élaborée au cours de toute son expérience passée.

En transposant ce principe aux dispositifs spintroniques, cela reviendrait à modifier la valeur des synapses artificielles en fonction du signal qu'elles reçoivent. Et effectivement, il est possible de reproduire une certaine plasticité avec ces systèmes. Quand une tension électrique est appliquée à une jonction tunnel magnétique, l'effet de transfert de spin met en mouvement l'aimantation d'un de ses nanoaimants. Si la tension est assez élevée et appliquée assez longtemps, ce mouvement entraîne l'inversion de l'aimantation du nanoaimant, et donc de la valeur stockée par la jonction tunnel magnétique. Mais ce renversement n'est pas toujours prédictible: si la tension n'est pas appliquée pendant une durée suffisante, parfois l'aimantation se retourne, parfois non. Plus la tension appliquée à la jonction est élevée, plus la probabilité de basculement de l'aimantation est importante, sans pour autant atteindre 100%.

UNE CAPACITÉ D'APPRENTISSAGE

Dans une mémoire usuelle, ce comportement incertain doit être évité à tout prix. Pour un système neuromorphique, on obtient des synapses qui, sous l'action des signaux qu'elles reçoivent, transforment progressivement la fonction du réseau de neurones dont elles font partie et adaptent la réponse du système à la nouvelle tâche sans avoir besoin d'utiliser des méthodes aussi lourdes que la rétropropagation du gradient.

En 2015, avec nos collègues, nous avons montré qu'un système équipé de mémoires spintroniques plastiques pouvait apprendre à compter des véhicules sur une autoroute (avec une précision de 95%) ou à reconnaître des chiffres manuscrits, de la même façon que l'aurait fait un algorithme usuel d'intelligence artificielle. Sur ces tâches relativement simples, les performances s'approchent des techniques plus traditionnelles d'apprentissage des réseaux de neurones (qui reposent cependant sur plus de synapses, avec une consommation plus élevée en énergie). En revanche, sur des tâches complexes (comme concurrencer un

spécialiste du jeu de go), les techniques d'apprentissage bio-inspirées sont pour l'instant loin de rivaliser avec les techniques plus conventionnelles.

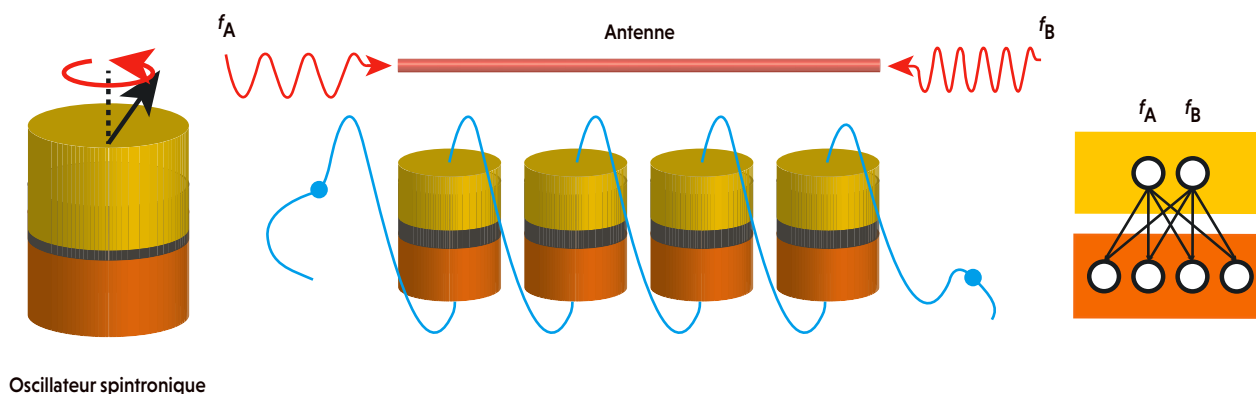
Si nous sommes capables de réaliser des synapses artificielles qui stockent de l'information et dont il est possible de moduler la réponse pour simuler la plasticité synaptique et le phénomène d'apprentissage, un système complet neuromorphique doit aussi inclure des neurones pour effectuer les calculs.

Pour réaliser un unique neurone artificiel avec des transistors, il en faudrait des centaines

Le cerveau humain contient près d'une centaine de milliards de neurones, et les algorithmes d'intelligence artificielle simulent aujourd'hui jusqu'à des dizaines de millions de neurones. Est-il envisageable de réaliser des neurones avec les techniques actuelles de transistors au silicium? D'abord, si l'on veut autant de composants dans une puce électronique de la taille d'un pouce, il faudrait que chaque neurone ait une largeur bien inférieure à un micromètre, ce qui est impossible, car réaliser même un unique neurone artificiel par cette approche nécessite des centaines de transistors.

Par exemple, la puce TrueNorth conçue par IBM en 2014 est une démonstration impressionnante d'électronique neuromorphique qui ne fait pas appel à la spintronique. Ce système comporte cinq milliards de transistors (un nombre considérable), ce qui permet d'implémenter un million de neurones et des centaines de millions de synapses. En rapprochant ainsi centre de calcul et mémoire, IBM affirme que la puissance consommée par la puce est inférieure de plusieurs ordres de grandeur à celle d'une puce usuelle.

En réalité, l'architecture de la puce TrueNorth ne contient pas un million de neurones. Un neurone artificiel de ce type occupe une trop grande surface pour en mettre un million sur une puce. En revanche, cette puce est extrêmement rapide; dès lors, un seul circuit >



Oscillateur spintronique

La jonction tunnel magnétique peut aussi être utilisée comme un neurone qui traite l'information. Pour cela, les caractéristiques de la jonction sont telles que le courant qui la traverse fait tourner de façon permanente l'aimantation de la couche supérieure (à gauche). La jonction émet alors un signal impulsif comme un neurone biologique.

Ce composant se comporte aussi comme un émetteur-récepteur d'ondes radio. Dans une expérience, les chercheurs ont associé quatre neurones artificiels qui synchronisaient aussi leurs signaux grâce à une antenne (au centre). Ce réseau de neurones (schématisé à droite) est capable de reconnaître des voyelles prononcées (dont la signature était réduite à deux fréquences caractéristiques f_A et f_B).

> numérique implémente plusieurs neurones artificiels (ceux qui seraient physiquement situés autour), en faisant successivement les calculs associés à chacun d'eux. C'est une petite, mais très pragmatique, entorse à l'idée de base du neuromorphisme de faire des neurones et synapses artificiels. En 2017, Intel a développé une puce analogue, mais plus modeste, Loihi, qui n'implémente que 130 000 neurones. Ces dispositifs restent volumineux et il est difficile de penser qu'il sera possible de rivaliser avec le cerveau humain en suivant cette approche.

La spintronique fournit là encore une piste. Il est possible d'utiliser une jonction tunnel magnétique pour imiter les fonctions clés des neurones du cerveau. En particulier, ces derniers communiquent en émettant des impulsions électriques. Ces impulsions sont toutes identiques, mais leur nombre dépend de la sollicitation du neurone. Si un neurone reçoit beaucoup d'impulsions pendant un temps court, la fréquence avec laquelle il va engendrer des impulsions sera alors également élevée. S'il est stimulé par peu d'impulsions, il en émettra lui-même un nombre réduit.

Avec de l'électronique traditionnelle, il est difficile de réaliser des neurones impulsifs de taille nanométrique. En effet, produire des impulsions à un certain rythme nécessite de créer des boucles de rétroaction dans le circuit électronique, ce qui occupe de la place. Il est en revanche possible de fabriquer des jonctions tunnel magnétiques qui reproduisent ce comportement.

Une solution consiste à prendre les signaux d'entrée sous la forme de courants électriques qui sont additionnés (exactement comme dans un neurone biologique) avant d'être injectés dans la jonction. Ensuite, on choisit les caractéristiques du dispositif de telle sorte que le spin transféré par les électrons du courant entrant ne conduit pas à un basculement complet de l'aimantation du nanoaimant, mais la fait osciller et tourner en

permanence, telle une boussole dont l'aiguille pivoterait sur elle-même. Comme la configuration respective des deux aimantations (celle fixe et celle en rotation) change périodiquement, la résistance électrique de la jonction (on parle d'effet de magnétorésistance) s'en trouve affectée et se traduit par un signal électrique périodique. Plus le courant entrant est important, plus la vitesse de rotation sera grande et la fréquence des impulsions engendrées sera élevée. On obtient bien un comportement très similaire à celui d'un neurone. En outre, ce phénomène est réalisable à température ambiante et est très stable pour des dispositifs nanométriques. Il peut donc être exploité pour des dispositifs concrets.

RECONNAISSANCE DE CHIFFRES : 99,6 % DE RÉUSSITE

En 2017, avec nos équipes, nous avons montré expérimentalement qu'une jonction tunnel magnétique pouvait effectivement imiter un neurone. Nous sommes même allés plus loin : nous avons utilisé une jonction unique pour imiter un réseau entier de quatre cents neurones grâce à une stratégie dite « de multiplexage temporel », la jonction jouant tour à tour le rôle de chaque neurone. Nous avons ensuite utilisé ce réseau de neurones spintroniques pour identifier des chiffres prononcés par différentes personnes.

Pour ce faire, nous avons converti les signaux audio à reconnaître en courant électrique injecté à travers la jonction tunnel magnétique. Ce courant injecté, qui correspond aux chiffres prononcés mille fois plus vite (pour être dans la bonne gamme de fréquences du dispositif), provoque une rotation de l'aimantation d'un nanoaimant de la jonction. Ce faisant, grâce à l'effet de magnétorésistance, le neurone convertit les modulations du courant en variation de la tension électrique à ses bornes. Nous avons enregistré ces variations de tension électrique provenant des neurones spintroniques, que nous avons

transmises à un ordinateur qui a simulé l'ensemble du réseau de neurones avec ses fonctions synaptiques. Ces dernières ont alors été configurées grâce à une phase d'apprentissage pour reconnaître les chiffres parlés. Puis, soumis à de nouveaux échantillons de paroles, le réseau identifiait correctement les chiffres avec un taux de succès de 99,6%, ce qui représente une performance exceptionnelle. Cette expérience a démontré la capacité des jonctions tunnel magnétiques impulsives à imiter les neurones de façon fiable.

Les travaux dans le domaine de la spintronique à des fins neuromorphiques sont encourageants. Notre futur défi sera de réaliser des réseaux de neurones complets comportant des millions de jonctions tunnel magnétiques jouant le rôle de neurones et de synapses. Comme on l'a vu, l'industrie sait déjà les intégrer par milliards sur des puces et les relier à des circuits électroniques classiques pour faire des mémoires, ce qui constitue une base solide pour aller plus loin.

La spintronique n'a pourtant pas fini de nous surprendre. Les jonctions tunnel magnétiques seraient capables d'imiter d'autres propriétés du cerveau, comme la synchronisation de neurones distants. En effet, dans le cerveau, des impulsions conduisent des neurones parfois très éloignés à accorder leurs réponses, en les déclenchant au même moment. Cette synchronisation a de nombreux avantages. Des neurones qui émettent des impulsions en même temps auront plus d'importance dans le traitement du signal. S'ils sont connectés au même neurone, celui-ci reçoit un très grand nombre d'impulsions dans un bref laps de temps et, ainsi, il les transmet très efficacement à son tour. Des populations entières de neurones se coordonnent par ce phénomène, au point de créer des ondes cérébrales qui, grâce à leur puissance, modifient le comportement de neurones dans d'autres aires cérébrales, parfois lointaines.

DE MINUSCULES ÉMETTEURS-RÉCEPTEURS

Les jonctions tunnel magnétiques ayant un fonctionnement très rapide, elles produisent des impulsions électromagnétiques à une fréquence pouvant dépasser quelques centaines de mégahertz. Le traitement de l'information se fait ainsi des millions de fois plus vite que dans le cerveau. Mais, autre conséquence, ces dispositifs se comportent comme des nanoémetteurs radio. Or ces jonctions sont aussi très sensibles aux ondes électromagnétiques environnantes, lesquelles peuvent modifier leur génération d'impulsions : les jonctions sont donc aussi de très bons récepteurs radio. En ajustant correctement les caractéristiques d'un ensemble de jonctions, il est possible de les synchroniser entre elles grâce aux

signaux qu'elles émettent et reçoivent... comme les neurones biologiques!

Induire ce type de comportement synchronisé dans des réseaux de neurones artificiels est une piste passionnante pour coordonner matériellement et efficacement des réseaux de neurones spécialisés dans des tâches différentes. Dans cette optique, en 2018, avec nos collègues, nous avons étudié un système de quatre jonctions tunnel magnétiques dont la tâche était de reconnaître des voyelles prononcées oralement (voir la figure page 36). Le signal vocal était réduit

Grâce au multiplexage, une jonction a imité quatre cents neurones

à deux fréquences (par analyse de Fourier) et accéléré 100 000 fois pour être émis par une antenne aux quatre nanooscillateurs. De façon très intéressante, dans cette expérience, nous avons constaté que la capacité des jonctions à coordonner leurs rythmes augmentait les performances de classification des signaux vocaux.

Pour aller plus loin, nous cherchons désormais à réaliser des réseaux dont les neurones communiquent par les ondes électromagnétiques qu'émettent et reçoivent les jonctions. Il s'agit là d'une piste prometteuse pour interconnecter les neurones dans un réseau dense, ce qui est une condition indispensable pour l'implémentation de tâches complexes. À terme, un système complet associera ces jonctions jouant le rôle de neurones à des synapses artificielles à base de MRAM, pour lui donner une capacité d'apprentissage.

Que pourront alors réaliser de tels systèmes? Notre objectif n'est pas d'obtenir une intelligence capable de rivaliser avec l'intelligence humaine : notre connaissance du cerveau est bien trop partielle pour un tel résultat. En revanche, nous souhaitons concevoir une intelligence artificielle plus économe en électricité et qui soit plus rapide que celles disponibles actuellement. Ainsi, nous réduirons la consommation énergétique considérable des centres de calcul qui traduisent des textes ou interprètent ce que nous disons à nos enceintes intelligentes, et nous aurons la possibilité d'embarquer de tels dispositifs dans nos téléphones portables ou dans des véhicules autonomes pour nous assister dans notre vie quotidienne. ■

BIBLIOGRAPHIE

J. Grollier, D. Querlioz et al., **Neuromorphic spintronics**, *Nature Electronics*, 2 mars 2020.

M. Romera et al., **Vowel recognition with four coupled spin-torque nano-oscillators**, *Nature*, vol. 563, pp. 230-234, 2018.

J. Torrejon et al., **Neuromorphic computing with nanoscale spintronic oscillators**, *Nature*, vol. 547, pp. 428-431, 2017.

N. Locatelli et al., **Spintronic devices as key elements for energy-efficient neuroinspired architectures**, *IEEE Xplore : Design, Automation & Test in Europe Conference & Exhibition*, 2015.

CNRS FORMATION ENTREPRISES

Le CNRS propose un catalogue de plus de 270 formations technologiques à destination des entreprises pour l'année 2021

L'organisme de formation continue du CNRS présente un large panel de stages courts (3 jours en moyenne) dans les domaines suivants :

DONNÉES, CONNAISSANCES, APPRENTISSAGE	PHYSIQUE ET INSTRUMENTATION	BIOLOGIE MOLÉCULAIRE ET BIOCHIMIE
BIOINFORMATIQUE	CARACTÉRISATION DES MATÉRIAUX	BIOLOGIE ANIMALE ET FORMATIONS RÉGLEMENTAIRES
GÉNIE LOGICIEL ET SYSTÈMES D'INFORMATION	CHIMIE, SYNTHÈSE, PROCÉDÉS	QUALITÉ ET SÉCURITÉ
TERRITOIRE, PATRIMOINE ET ENVIRONNEMENT	CHIMIE ANALYTIQUE	SOCIOLOGIE, SCIENCE POLITIQUE ET ÉCONOMIE
SCIENCES DE L'INGÉNIEUR	MICROSCOPIE ET IMAGERIE	COGNITION ET COMPORTEMENT
	BIOLOGIE CELLULAIRE ET MICROBIOLOGIE	

Cet outil de transfert des savoirs et savoir-faire du CNRS permet de **former plus de 1600 professionnels chaque année** issus des secteurs privé (PME et grands groupes) et public.

LES NOUVEAUTÉS PROPOSÉES PAR CNRS FORMATION ENTREPRISES EN 2021

DONNÉES, CONNAISSANCES, APPRENTISSAGE - Python et le module Pandas pour gérer et analyser des données du 27 au 29/01/2021 MONTPELLIER - 2,5 jours 1 400€ - Deep learning pour le traitement automatique des langues du 31/05 au 02/06/2021 GIF-SUR-YVETTE - 3 jours 1 500 € - Machine learning et deep learning pour la vision par ordinateur du 14 au 16/06/2021 LYON - 2,5 jours 1 600 € - Deep learning pour le traitement et l'analyse d'images médicales du 16 au 18/06/2021 VILLEURBANNE - 2,5 jours 1 600 €	BIOINFORMATIQUE - Bioinformatique pour la métagénomique du 10 au 12/05/2021 MONTPELLIER - 2,5 jours 1 200 € GÉNIE LOGICIEL ET SYSTÈMES D'INFORMATION - Gestion d'applications Cloud avec l'orchestrateur Kubernetes (concepts et pratique) du 26 au 27/05/2021 GRENOBLE - 2 jours 1 000 € TERRITOIRE, PATRIMOINE ET ENVIRONNEMENT - LiDAR et photogrammétrie pour la numérisation de terrain : instrumentation, mesure et modélisation du 08 au 11/03/2021 CAEN - 3,5 jours 1 200 €	SCIENCES DE L'INGÉNIEUR - Acquérir et dialoguer en Python du 22 au 26/03/2021 GIF-SUR-YVETTE - 4 jours 1 600 € - Visualiser en temps réel et interagir avec ses données en Python du 05 au 08/04/2021 GIF-SUR-YVETTE - 3 jours 1 350 € PHYSIQUE ET INSTRUMENTATION - Fabrication additive : dépôt par jet d'encre du 30/03 au 01/04/2021 TOULOUSE - 3 jours 1 400 € - Micro- et nano-électronique : évolution, état de l'art et perspectives des technologies CMOS, Beyond-CMOS et Mémoires du 29/09 au 01/10/2021 GRENOBLE - 2 jours 1 000 €
--	--	---