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 familier. 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.

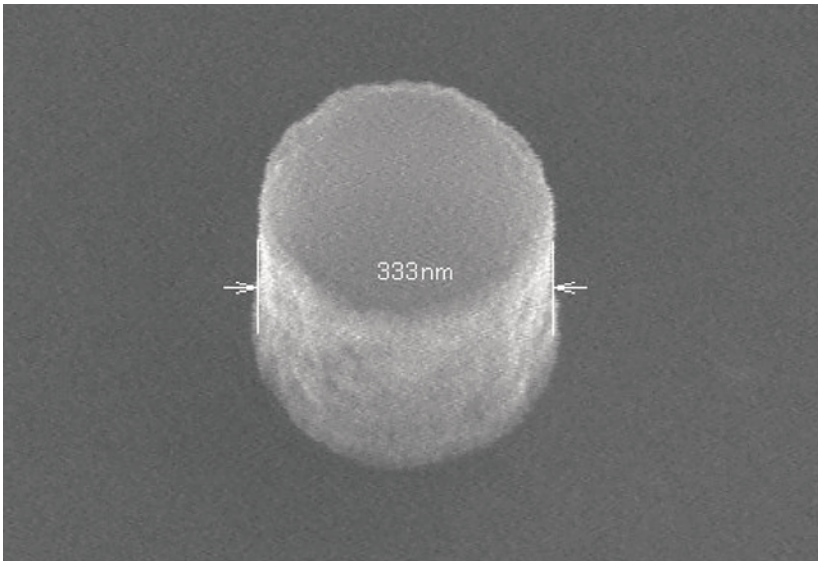
Au cours de l'évolution, le cerveau, qui ne dispose que d'une source d'énergie limitée – nos apports caloriques quotidiens – a développé une tout autre stratégie. Il fonctionne sur la base d'un compromis entre la fiabilité du traitement de l'information et la dépense énergétique. Les synapses et les 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 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.




**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.**

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 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 >