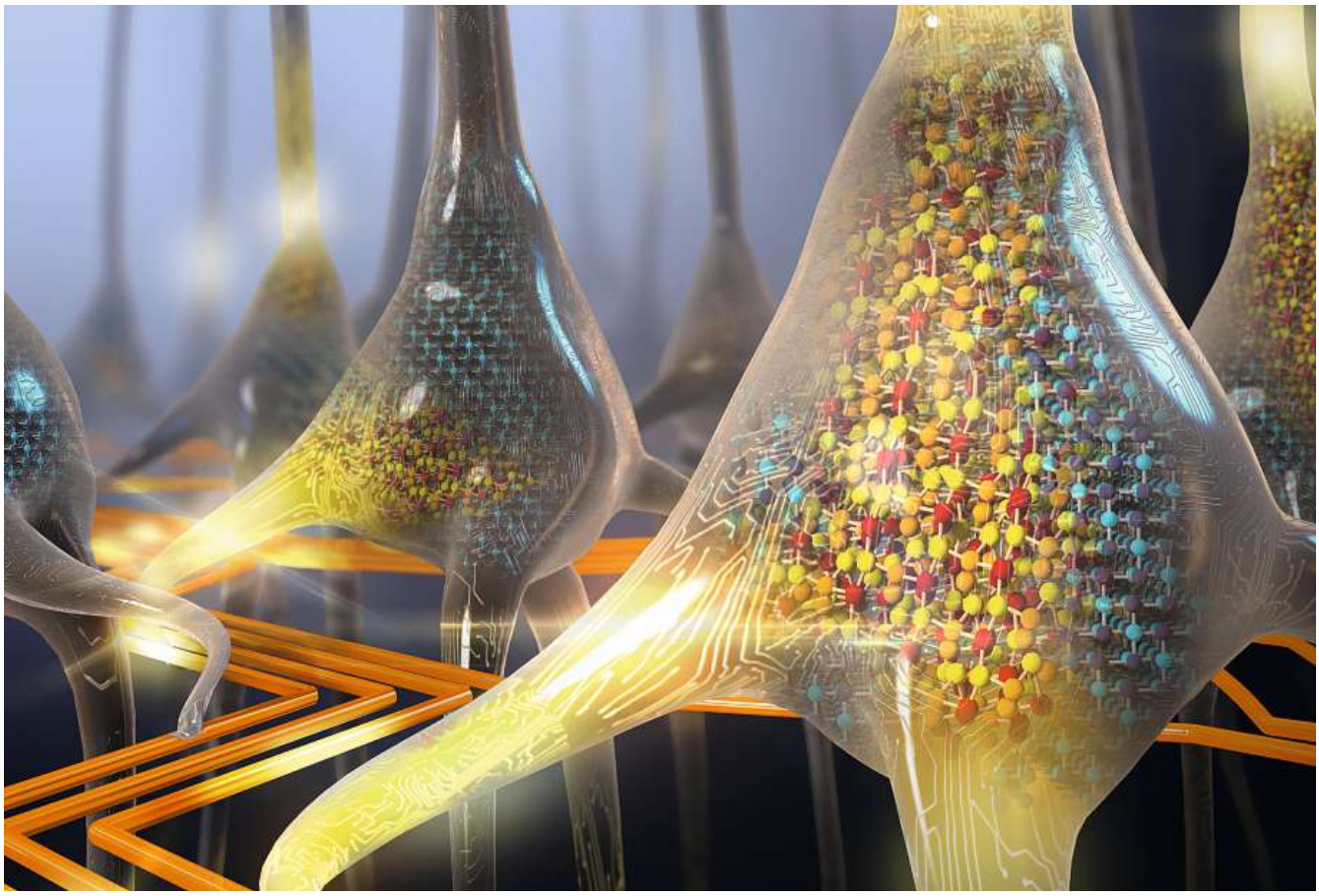

L'ère de l'informatique cognitive est arrivée

L'efficacité énergétique, la clé du progrès | L'architecture actuelle des ordinateurs cognitifs s'est révélée inefficace en raison de la séparation physique entre calcul et mémoire. De nouveaux dispositifs basés sur des matériaux à changement de phase permettent désormais de réaliser des ordinateurs neuromorphiques, inspirés par le cerveau, qui consomment beaucoup moins d'énergie.

TEXTE MANUEL LE GALLO-BOURDEAU

Depuis les premiers ordinateurs numériques inventés dans les années 1950, d'énormes améliorations technologiques ont été apportées afin de développer le concept jusqu'aux machines informatiques traditionnelles que nous utilisons couramment aujourd'hui. Basées sur l'architecture de von Neumann, ces dernières sont des ordinateurs à programme enregistré dans lesquels les unités de calcul et de mémoire sont physiquement séparées.

Aujourd'hui, nous entrons dans l'ère de l'informatique cognitive, dans laquelle intelligence et connaissance sont dérivées à partir d'énormes volumes de données. Les ordinateurs cognitifs sont des machines qui apprennent en fonction des données qu'elles reçoivent et à partir desquelles elles peuvent obtenir des informations et des connaissances pour complémenter notre capacité à raisonner. Dans les ordinateurs cognitifs actuels basés sur l'architecture de von Neumann, d'énormes quantités de données doivent être traitées de façon parallèle. Or, la séparation entre calcul et mémoire rend cette architecture inefficace pour cette tâche.

S'éloigner de l'architecture de von Neumann

L'une des premières démonstrations de l'informatique cognitive a été la participation de Watson, le programme informatique d'intelligence artificielle conçu

Figures: IBM Research

par IBM, à Jeopardy[1] en 2011, où il a pu battre deux des meilleurs joueurs de l'émission. Il s'agit là, sans aucun doute, d'une démonstration impressionnante de la puissance de l'informatique cognitive! Cependant, quand on se rend compte que Watson utilisait 80 kW alors que les joueurs humains avaient besoin chacun d'environ 20 W pour effectuer la même tâche de bout en bout, cela relativise la défaite de ces derniers.

Watson, comme tous les ordinateurs actuels, est basé sur l'architecture informatique de von Neumann, dans laquelle les unités de calcul (CPU) et de mémoire (DRAM) sont physiquement séparées. Cela signifie que les 20 To de données de connaissances humaines de Watson utilisées dans Jeopardy devaient constamment circuler entre le processeur et la mémoire vive pour effectuer les tâches de calcul, ce qui consomme une quantité d'énergie considérable.

Cet exemple montre qu'il existe clairement un besoin de nouveaux dispositifs et architectures pour l'informatique cognitive. Pour construire des ordinateurs cognitifs efficaces, il est en effet essentiel de passer à des architectures non von Neumann où la mémoire et la logique coexistent sous une forme ou une autre.

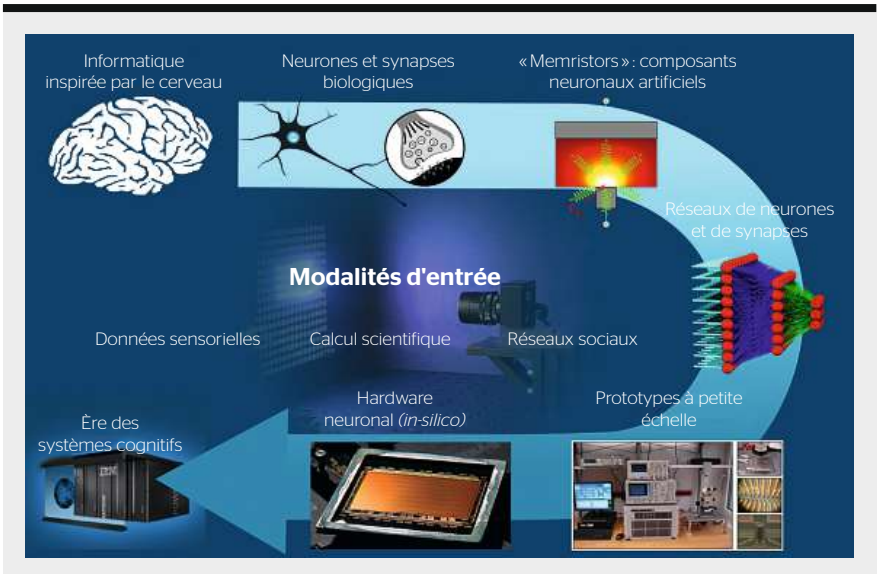

Figure 1 S'inspirant du fonctionnement interne du cerveau biologique, les ordinateurs neuromorphiques sont des systèmes construits à partir de neurones et de synapses artificiels connectés sous forme de réseau.

Inspirés par le cerveau

Les ordinateurs neuromorphiques inspirés par le cerveau sont actuellement étudiés comme étant une approche non von Neumann prometteuse. Ce paradigme s'inspire du fonctionnement interne du cerveau biologique afin de construire des neurones et synapses artificiels qui peuvent imiter les fonctions de leurs homologues biologiques en utilisant très peu d'énergie. De ce fait, des tâches typiques d'intelligence artificielle, par exemple reconnaître et classer de grandes quantités de données comme des images contenant des millions de pixels, des sons ou des vidéos, pourraient être réalisées beaucoup plus rapidement et de façon bien plus économe en énergie qu'avec des ordinateurs traditionnels. Contrairement à ces derniers, le cerveau traite l'information de façon parallèle à l'intérieur de réseaux de neurones, sans aucune séparation entre le calcul et la mémoire. En construisant des neurones et synapses artificiels et en interconnectant ceux-ci sous forme de réseaux de manière similaire au cerveau humain, il est envisageable de réaliser des ordinateurs neuromorphiques composés d'une multitude de ces réseaux aptes à réaliser certaines tâches d'intelligence artificielle de façon très efficace **(figure 1)**.

L'élément critique dans ce nouveau paradigme de calcul est un dispositif de mémoire à l'échelle nanométrique à haute densité, à faible puissance, variable, programmable et non volatile. Bien qu'il soit possible d'émuler les fonctions de synapses et de neurones en utilisant des transistors, les composants de base de l'informatique actuelle, cela requiert un nombre relativement élevé de transistors par synapse ou neurone et limite donc les possibilités d'avoir un grand nombre de ces derniers sur un même chip. De plus, la consommation énergétique de ces systèmes basés uniquement sur des

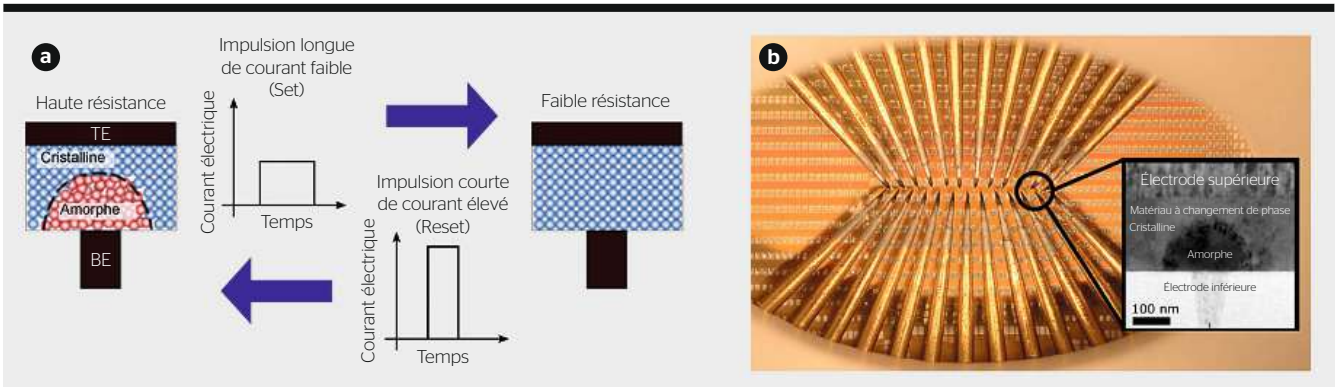

Figure 2 Les mémoires à changement de phase basculent entre une phase amorphe de haute résistance et une phase cristalline de faible résistance sous l'effet d'un courant électrique : principe d'opération d'une mémoire à changement de phase (a) et puce contenant des centaines de mémoires à changement de phase qui peuvent être connectées en utilisant une sonde multipointes (b).

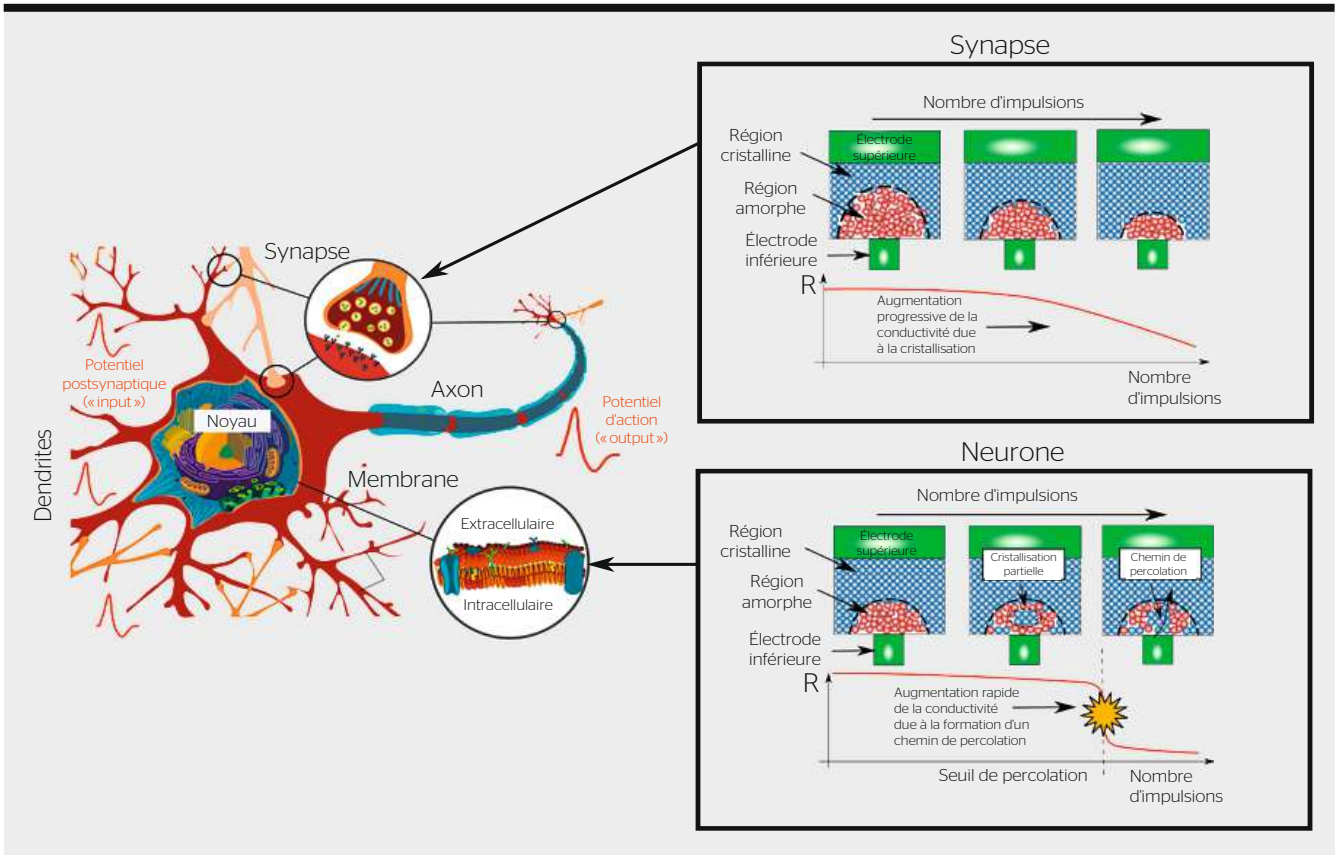

Figure 3 Une mémoire à changement de phase peut être utilisée pour imiter les fonctions à la fois de neurones et de synapses biologiques en utilisant des impulsions courtes de courant répétées qui vont cristalliser la phase amorphe.

transistors risque d'être relativement élevée, et ce, surtout pour des systèmes réalisant de l'apprentissage automatique, tâche clé de l'intelligence artificielle, de façon in situ et non volatile comme le cerveau humain. De ce fait, de nouveaux dispositifs de mémoire qui ne sont pas basés sur des transistors, mais sur des matériaux inédits et pas encore conventionnels dans l'informatique actuelle, doivent être envisagés pour réaliser des neurones et synapses artificiels ayant une très haute densité et qui ne consomment que très peu d'énergie.

Une mémoire à changement de phase

Chez IBM Research - Zurich, les dispositifs à mémoire à changement de phase (PCM, Phase Change Memory) basés sur des matériaux à changement de phase de type chalcogénure, tels que le $Ge_2Sb_2Te_5$ sont étudiés pour construire des synapses et neurones artificiels. Ces matériaux sont les mêmes que ceux utilisés dans les disques DVD et Blu-Ray pour stocker l'information. Une différence cependant: dans un composant PCM, le stockage et la lecture de l'information s'effectue en appliquant des pulses de courant électrique, alors qu'un laser est utilisé dans les dispositifs optiques. Les PCM font partie d'une classe de dispositifs appelés communément « memristors ». Ces derniers stockent l'information dans leur résistance électrique, laquelle peut être changée de façon permanente sous l'application d'un courant électrique. De ce fait, le memristor est un résistor variable qui reflète l'histoire du courant qui l'a traversé par sa valeur de résistance et, de cette manière, peut être utilisé comme une mémoire qui traite l'information de façon similaire aux neurones et aux synapses biologiques.

La mémoire PCM exploite le comportement de matériaux dits à changement de phase qui peuvent être basculés de manière réversible entre des phases amorphes et cristallines de résistivité électrique différente. La phase amorphe a une résistivité électrique élevée, tandis que la phase cristalline présente une faible résistivité, parfois inférieure de trois ou quatre ordres de grandeur. Ce grand contraste de résistance est utilisé pour stocker des informations: l'état de haute résistance peut représenter un « 0 » logique tandis que l'état de faible résistance peut représenter un « 1 » logique. Ainsi, un dispositif PCM consiste essentiellement en une couche de matériau à changement de phase prise en sandwich entre deux électrodes métalliques.

La phase amorphe est un verre thermodynamiquement instable dont le temps de cristallisation à température ambiante est très long. Cependant, en chauffant le matériau amorphe à une température suffisamment élevée mais inférieure à la température de fusion, il cristallisera rapidement. Pour transformer le matériau en phase amorphe, il faut le chauffer au-dessus de sa température de fusion, puis le refroidir rapidement. Ce refroidissement rapide va « geler » la structure atomique dans un état désordonné. Dans un dispositif PCM, la chaleur est produite par le passage d'un courant électrique à travers le matériau à changement de phase (effet Joule). L'impulsion électrique utilisée pour amener le dispositif à l'état amorphe de haute résistance est appe-

lée « impulsion Reset » et l'impulsion utilisée pour ramener le dispositif à l'état cristallin de faible résistance est appelée « impulsion Set » **(figure 2)**.

Une démarche neuromorphique

Le fait qu'un dispositif PCM puisse être programmé dans des états de résistance intermédiaires (entre les états de haute et de basse résistance) constitue l'une de ses propriétés-clés. Plus précisément, la région amorphe peut être progressivement cristallisée en appliquant des impulsions de courant répétitives. Cette propriété peut être utilisée pour imiter les fonctions de synapses et de neurones avec des dispositifs PCM. En effet, la plasticité d'une synapse, qui est sa faculté à se renforcer ou à se déprécier pour mieux relier les neurones importants, peut être émulée par un dispositif PCM en utilisant cet effet de cristallisation progressive. De même, la fonction « intègre et tire » (en anglais Integrate-and-Fire) d'un neurone, qui fait référence à la génération d'un potentiel d'action lorsqu'il est soumis à suffisamment d'excitations, peut aussi être émulée par un dispositif PCM, et ce, toujours avec le même effet de cristallisation progressive **(figure 3)**.

Dans les récents travaux effectués chez IBM Research - Zurich, dont l'un a été présenté en couverture de la revue Nature Nanotechnology en août 2016, il a été démontré expérimentalement comment détecter des corrélations dans plusieurs flux d'événements à l'aide de neurones et de synapses basés sur des dispositifs PCM prototypes **(figure 4)**. Ces événements pourraient être par exemple des données Twitter ou tout type de données sensorielles. En utilisant un neurone connecté à plusieurs synapses plastiques recevant les événements, implémentés en utilisant ces dispositifs PCM, il est possible de détecter ces corrélations de manière complètement non supervisée (les corrélations sont apprises directement à partir des données d'entrée sans connaissance préalable du système, **figure 5)**. L'énergie dissipée dans les dispositifs PCM est d'environ 5 pJ (10^{-12} J) par mise à jour synaptique/neuronale, alors que les ordinateurs standard actuels présentent une consommation de l'ordre de 1 nJ (10^{-9} J) par opération en virgule flottante. Bien qu'il ne s'agisse pas d'une comparaison

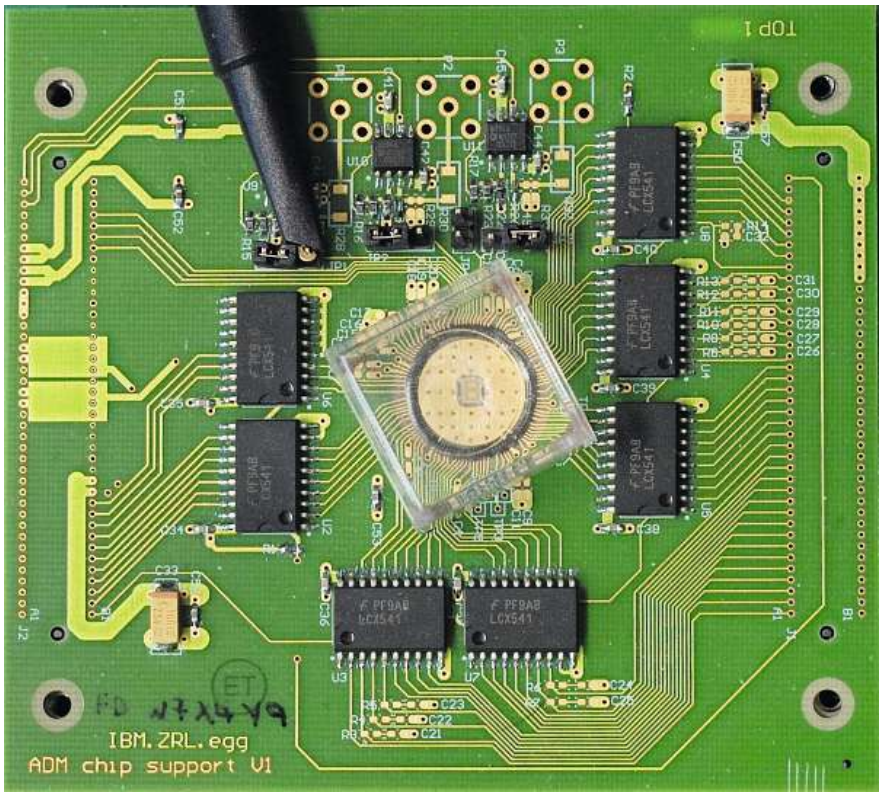

Figure 4 Puce électronique prototype contenant plusieurs millions de mémoires à changement de phase. La puce, qui se trouve sous la protection de plastique transparent, a une dimension de 2,75 mm x 2,2 mm et est connectée par fils à une plateforme de support.

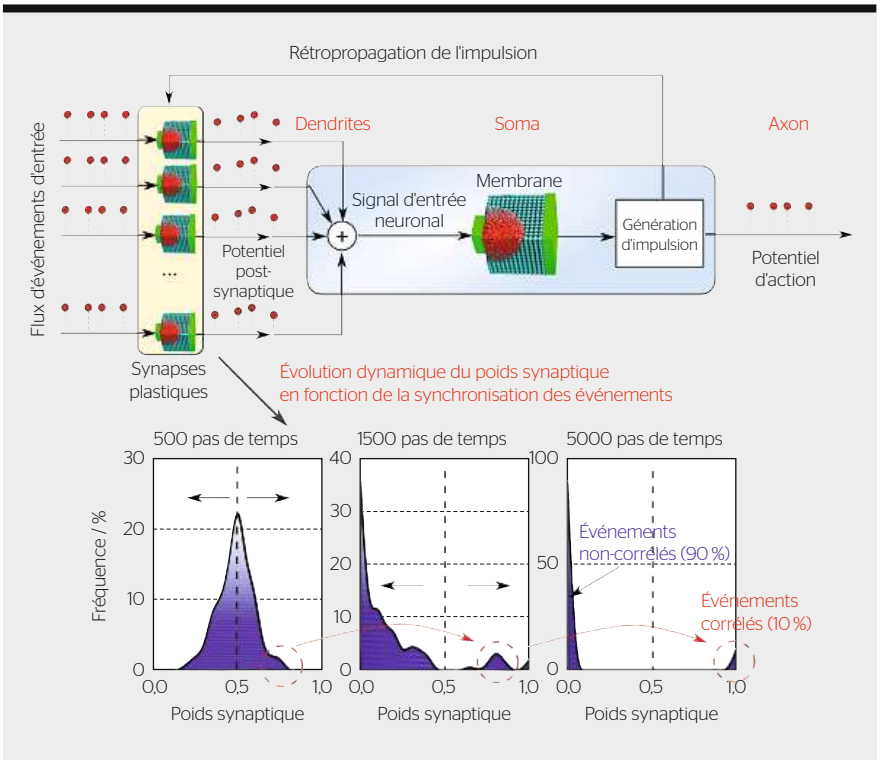

Figure 5 Le poids synaptique des synapses implémentées avec des mémoires à changement de phase évolue en fonction des données d'entrées de façon à pouvoir discriminer les événements corrélés de ceux qui ne le sont pas de manière complètement non supervisée.

directe, l'économie d'énergie réalisée par l'approche neuromorphique est tout de même prometteuse.

De multiples applications

Il existe de multiples applications potentielles pour les ordinateurs neuromorphiques basés sur des mémoires à changement de phase. Outre accélérer les tâches d'apprentissage automatique qui sont communément exécutées aujourd'hui sur des accélérateurs graphiques (GPU), ces systèmes neuromorphiques pourraient aussi potentiellement être intégrés sous forme de capteurs «intelligents» capables de traiter les données collectées localement en temps réel en utilisant une très faible puissance. Cela pourrait être bénéfique dans de nombreux domaines, par exemple pour le traitement de la pollution.

Dans la société actuelle, la qualité de l'air est devenue un problème majeur dans certaines parties du monde. Par exemple, respirer l'air de Pékin est comparable à fumer 30 cigarettes par jour. Il y a quelques années, IBM a commencé à travailler avec le gouvernement chinois pour prédire les niveaux de pollution et, finalement, réduire considérablement les polluants. Les capteurs à travers le pays recueillent des données et peuvent aujourd'hui générer des prévisions de pollution à une haute résolution de 1 km² 72 h à l'avance. Alors que les données sont trop complexes pour que les analystes humains soient capables de détecter des modèles, l'intelligence artificielle est en mesure, grâce à l'apprentissage automatique, de digérer de grandes quantités de données de capteurs afin de repérer les tendances. Avec les capteurs, l'intelligence artificielle et son évaluation humaine, il est possible d'identifier les pollueurs et d'aider à concevoir des villes plus intelligentes, des autoroutes et des méthodes d'adaptation.

Des ordinateurs aux capteurs neuromorphiques

Les capteurs d'aujourd'hui sont gourmands en énergie, ce qui signifie que leur batterie doit être remplacée fréquemment, et leur capacité de traitement est très limitée. Par conséquent, toutes les données collectées doivent être envoyées au cloud pour être analysées, ce qui nécessite beaucoup de temps et d'énergie et conduit à une énorme accumulation de données sur les serveurs cloud. En revanche, les capteurs neuromorphiques disposeraient de réseaux intégrés de neurones artificiels et de synapses qui pourraient traiter les données reçues en temps réel d'une manière complètement non supervisée, tout comme notre cerveau. Parce que ces neurones et synapses artificiels consomment très peu d'énergie, la batterie du capteur durerait des années et des capacités de traitement semblables à celles du cerveau assureraient que seules des données «pertinentes» soient envoyées au cloud. Cela entraînerait non seulement des prévisions de pollution plus rapides, mais également une détection immédiate des tendances afin d'identifier d'où proviennent les polluants en temps réel, par exemple.

Ces travaux de recherche sur les ordinateurs neuromorphiques permettent d'espérer qu'à l'avenir, il sera possible de construire de grands réseaux neuronaux avec ces appareils pour effectuer des tâches de calcul complexes avec des budgets et des densités de puissance comparables à ceux de la biologie, et de contribuer ainsi à une nouvelle ère de l'informatique cognitive.

Auteur
Manuel Le Gallo-Bourdeau est doctorant en génie électrique à l'École polytechnique fédérale de Zurich (EPFZ) et chercheur dans le département Memory & Cognitive Technologies chez IBM Research - Zurich.
→ IBM Research – Zurich, 8803 Rüschlikon
→ ANU@zurich.ibm.com

[1] Jeopardy est un jeu télévisé créé en 1964 aux États-Unis au cours duquel les candidats doivent, à partir d'indices formulés sous la forme de réponses à une question, trouver la question correspondante.

IN KÜRZE

Das Zeitalter der kognitiven Informatik ist eingeläutet

Energieeffizienz – der Schlüssel zum Fortschritt

Eine der ersten Demonstrationen im Bereich der kognitiven Informatik war die Teilnahme von Watson – einem von IBM entwickelten, auf künstlicher Intelligenz basierenden Computerprogramm – an der Quizsendung Jeopardy im Jahr 2011: das Programm besiegte am Ende zwei der bis dato besten Spieler. Um diese Aufgabe zu bewerkstelligen, verbrauchte Watson jedoch 80 kW – die beiden Spieler aus Fleisch und Blut hingegen gerade mal je 20 W. Dies liegt daran, dass Watson, wie alle gegenwärtigen Computer, auf der Von-Neumann-Architektur basiert, bei der Rechen- und Speichereinheit physisch voneinander getrennt sind: 20 TB Wissensdaten mussten bei Watson permanent zwischen Prozessor und RAM hin und her fliessen, was einen beträchtlichen Energieverbrauch mit sich bringt.

Neuromorphe Computer werden derzeit als vielversprechender Gegenentwurf zur Von-Neumann-Architektur untersucht. Dieses Paradigma ist von den inneren Mechanismen des biologischen Gehirns inspiriert, das Informationen innerhalb von neuronalen Netzwerken parallel und ohne Trennung zwischen Rechen- und Speichereinheit verarbeitet. Kritisches Element ist ein variables, programmierbares, nicht flüchtiges, nanoskaliges Speichermodul mit hoher Dichte und geringem Leistungsverbrauch. Im IBM-Forschungslabor bei Zürich werden Phase Change Memories (PCM) zum Aufbau von künstlichen Synapsen und Neuronen entwickelt. PCM speichern Informationen in deren elektrischem Widerstand, der stabil ist, aber reversibel verändert werden kann, indem man die Phase des Materials (kristallin oder amorph) durch einen elektrischen Stromimpuls ändert. Zudem können PCM durch fortschreitende Kristallisation in Zwischenwiderstandszuständen programmiert werden, um die Funktionen von Synapsen und Neuronen noch naturgetreuer nachzuahmen.

PCM-basierte neuromorphe Computer werden es künftig erlauben, Energie zu sparen und gleichzeitig maschinelle Lernaufgaben zu beschleunigen. Zusätzlich könnten diese neuromorphen Systeme unter Umständen in Form von «intelligenten» Sensoren integriert werden, die in der Lage sind, lokal gesammelte Daten in Echtzeit mit sehr geringem Leistungsverbrauch zu verarbeiten. CHE