dossier.

Mit der Zukunft rechnen

Das Ende des mooreschen Gesetzes motiviert | Die Miniaturisierung der konventionellen Halbleitertechnologie stösst an ihre Grenzen. Da eine weitere Skalierung kaum mehr möglich ist, arbeitet man intensiv an neuen Rechner-Konzepten.

Compter sur l'avenir

La fin de la loi de Moore motive | La miniaturisation de la technologie des semi-conducteurs atteint ses limites. Afin de continuer à accroître la puissance de calcul, les scientifiques travaillent intensément sur de nouveaux concepts d'ordinateurs.

Das Gehirn imitieren
Mit Phase Change Memory Devices (PCM) erforscht Manuel Le Gallo neuromorphe Rechner, die künftig Muster in grossen Datenmengen viel schneller und effizienter erkennen sollen.

Imiter le cerveau
Manuel Le Gallo travaille à la conception d'ordinateurs neuromorphiques basés sur des mémoires à changement de phase qui, à l'avenir, devraient reconnaître beaucoup plus rapidement et efficacement des patterns dans de grandes quantités de données.

TEXT RADOMÍR NOVOTNÝ

Vor einem halben Jahrhundert hat der Intel-Mitgründer Gordon Moore vorausgesagt, dass sich – vereinfacht gesagt – die Anzahl Transistoren auf einem Chip alle 12 Monate verdoppeln würde. 1975 passte er die Zeitspanne auf 24 Monate an, da seine Datenbasis nun solider war. Was zunächst eine Beobachtung der damaligen Entwicklung war, wurde von den Chipherstellern als Zielvorgabe übernommen.

Um dies zu erreichen, sah man bereits in den 2000er-Jahren, dass es nicht ausreicht, nur die Geometrie zu verkleinern. Zunächst wurden neue Materialien zur Erhöhung der Kapazität der Transistoren eingeführt und dann Tri-Gate-Transistoren bei 22 nm – eine Grösse, die der menschlichen DNA entspricht. Nun zeichnet sich ab, dass die Vorgabe von Moore nicht mehr erreicht werden kann, denn das heutige fotolithografische Verfahren schafft kleinste Strukturen von 14 nm mit Licht mit einer Wellenlänge von 193 nm. Noch kleinere Strukturen können zwar durch den Einsatz von extremem UV-Licht geschaffen werden, dies ist aber technisch schwierig und teuer. Zudem sinkt die Zuverlässigkeit der Transistoren, denn sie sind nun so nahe beieinander, dass sie sich gegenseitig beeinflussen. Es treten zunehmend auch unerwünschte Quanteneffekte und Leckströme auf.

Die ökonomischen Folgen dieser Herausforderungen können mit dem «Gegenspieler» des mooreschen Gesetzes ausgedrückt werden: Das Gesetz von Arthur Rock sagt, dass sich die Kosten einer Chipfabrik alle vier Jahre verdoppeln. Es kommt also ein Zeitpunkt, an dem die Investitionskosten so hoch sind, dass sich das mooresche Wachstum nicht mehr finanzieren lässt.

Wie weiter?

Die Datenmengen wachsen ungebremst weiter. Mit Voice-over-IP, dem Internet der Dinge, den Social Media und Video-Streams kommen Anforderungen, die nicht ignoriert werden können. Zudem steigen die Ansprüche an die Rechenleistung. Das Rechnen wird intensiver, Stichwort Cognitive Computing: Aus riesigen unstrukturierten Datenmengen möchte man nützliche Informationen gewinnen.

An neuen Hardware-Technologien wird bereits gearbeitet, aber bis Nano-Schalter und Ähnliches in grossem Massstab verfügbar sein werden, wird noch viel Zeit vergehen. Wie schliesst man diese Lücke? Einen Lösungsansatz skizziert Bruno Michel, der sich seit 14 Jahren als Forscher bei IBM mit Computerchiptechnologien und deren Kühlung beschäftigt. Bei seiner Perspektive steht der Wärmewiderstand im Fokus. Je kleiner er in Rechnern ist, desto dichter können diese gebaut werden. Und da 99% der Energie für den Datentransport eingesetzt werden und nur 1% für das eigentliche Rechnen, ist es naheliegend, dass dichter integrierte Computersysteme dank kürzerer Kommunikationswege deutlich weniger Energie brauchen. Für die gesamte Computerbranche gilt deshalb: Je kleiner die Systeme, desto effizienter sind sie. Bruno Michel vergleicht dies mit dem menschlichen Gehirn, das rund 10 000 Mal dichter und um

Il y a un demi-siècle, le cofondateur d'Intel Gordon Moore prédisait que le nombre de transistors dans une puce doublerait tous les 12 mois. En 1975, il ajustait la période à 24 mois, sa base de données étant désormais plus solide. Ce qui au départ était une observation du développement est devenu l'objectif des fabricants de puces.

Pour atteindre cet objectif, il est déjà apparu dans les années 2000 que la réduction de la géométrie ne suffirait pas. De nouveaux matériaux visant à augmenter la capacité des transistors ont d'abord été introduits puis les transistors Tri-Gate en 22 nm – une taille qui correspond à l'ADN humain. Mais il est de plus en plus clair que l'objectif de Moore ne peut plus être atteint, les plus petites structures pouvant être réalisées à l'aide des processus photolithographiques actuels avec une longueur d'onde de 193 nm étant limitées à 14 nm. Certes, il est possible de créer des structures encore plus petites avec une lumière UV extrême, mais il s'agit d'un procédé techniquement difficile et coûteux. De plus, cela réduirait la fiabilité des transistors, car ils seraient si proches les uns des autres qu'ils s'influenceraient mutuellement. Des effets quantiques et des courants de fuite indésirables apparaîtraient également.

Les conséquences économiques de ces défis peuvent être exprimées par « la contrepartie » de la loi de Moore : la loi d'Arthur Rock affirme que les frais d'une usine de fabrication de puces doublent tous les quatre ans. Il arrive donc un moment où les frais d'investissement sont si élevés qu'il devient impossible de financer la croissance relative à la loi Moore.

Comment poursuivre ?

Les quantités de données augmentent sans cesse. Voice-over-IP, l'Internet des objets, les médias sociaux et le streaming de vidéos génèrent une demande qu'il est impossible d'ignorer. Les exigences en termes de puissance des ordinateurs augmentent également. Le calcul devient plus intense, mot-clé Cognitive Computing : on souhaite extraire des informations utiles à partir d'énormes quantités de données déstructurées.

De nouvelles technologies informatiques sont déjà en cours de développement, mais il va falloir encore beaucoup de temps avant que des nanocommutateurs et autres systèmes soient disponibles à grande échelle. Comment combler cette lacune ? Bruno Michel, chercheur depuis déjà 14 ans chez IBM dans le domaine des technologies des puces des ordinateurs et leur refroidissement, ébauche une solution. Selon lui, la résistance thermique constitue le cœur du problème. Plus cette dernière est basse dans un ordinateur, plus celui-ci peut être construit de manière dense. Et dans la mesure où 99 % de l'énergie sont utilisés pour le transport des données et seulement 1 % pour le calcul, il est évident que les systèmes intégrés plus denses nécessitent nettement moins d'énergie en raison des voies de communication plus courtes. On peut donc énoncer l'affirmation suivante pour l'intégralité de la branche de l'informatique : plus les systèmes sont petits, plus ils sont

den gleichen Faktor effizienter als heutige Rechner ist. Zwei Gründe sind für diese Effizienz vorstellbar: Eine besondere Architektur und die Dichte. Da wir auch nach einem Jahrhundert Forschung die Funktionsweise des Gehirns nicht im Detail kennen, aber wissen, welche Rolle die Dichte bei der Effizienz spielt, schlägt er vor, dass man sich nun auf Letzteres konzentriert, um Rechner effizienter zu machen.

Den Wärmewiderstand verkleinern

Wenn der Wärmewiderstand reduziert wird, können Komponenten näher zusammengebracht werden. Dadurch sinkt der Energieverbrauch der Kommunikation und der Rechner wird effizienter. Erste Schritte in diese Richtung wurden bei den wassergekühlten Rechnern Aquasar (ETH Zürich) und Supermuc am Leibniz-Rechenzentrum in München realisiert. Letzterer leistet bei einem Volumen von 50 000 m^3 1 PFlops (10^{15} Flops). Die IBM-Forscher wollen dies mit einer 3D-Integration in Kombination mit einer Flüssigkühlung im Chipstapel, die gleichzeitig die Stromversorgung übernimmt, 50 Millionen Mal schrumpfen. Dadurch soll das System 5000 Mal effizienter werden – mit der aktuellen CMOS-Technologie. Zeithorizont: 10 bis 15 Jahre.

Ein weiterer Weg könnte mit dem Internet der Dinge beschritten werden. Am Körper getragene, sensorbestückte Mikrorechner, sogenannte Wearables, könnten eine neue Technologieplattform ermöglichen, die deutlich effizienter und preisgünstiger ist. Bruno Michel verweist hier auf das Gesetz, das durch Gordon Bell 1972 formuliert wurde. Bell beschrieb, wie neue Computer-Klassen entstehen, die neue Anwendungen ermöglichen und dadurch grössere Märkte und neue Industrien erobern – und dadurch viel preisgünstiger werden. Alle zehn Jahre findet ein solcher Technologiesprung statt. Begonnen hatte diese Entwicklung in den 1960er-Jahren mit den Mainframes, dann kamen die Minicomputer, gefolgt von den PCs, den Client-Server-Strukturen, dem Cloud Computing und heute den mobilen Geräten (Smart Phones, Tablets) bis zu drahtlosen Sensor-Netzwerken. Um möglichst sparsame Computer zu erhalten, muss man bei der neusten Technologie ansetzen, bei den preisgünstigen Wearables.

Heute sind Wearables oft noch zu gross und nicht zuverlässig und leistungsfähig genug. Aber wenn man diese Aspekte optimiert, kann ein Durchbruch im Computing erwartet werden. Dann wird das Prinzip, wonach man zum Computer gehen und seine Sprache lernen musste, umgekehrt. Man wird ein System haben, das immer präsent ist, sich intuitiv bedienen lässt und viel über die Person weiss – ein System, das in der Lage ist, unsere Emotionen zu erkennen und unsere Sprache zu sprechen.

Das Spektrum der Anwendungen solcher Wearables ist gross. Beispielsweise können sie chronisch Kranke an die Einnahme von Medikamenten oder durchzuführende Aktivitäten erinnern, oder den Zustand von Patienten, die früher

Um die Computer-Dichte zu erhöhen, muss der Wärmewiderstand reduziert werden.

Pour augmenter la densité des ordinateurs, la résistance thermique doit être réduite.

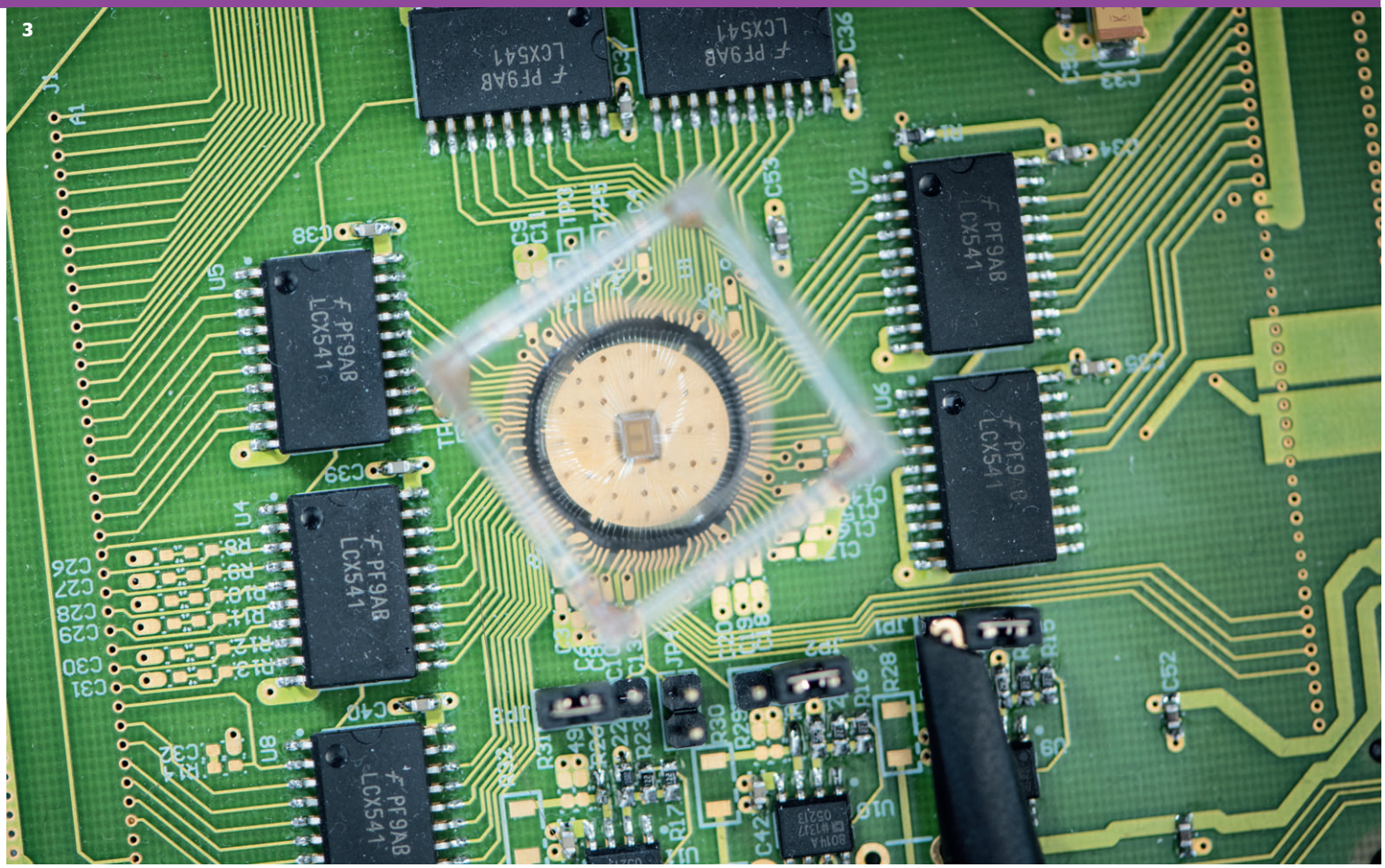
efficaces. Bruno Michel compare ce fait avec le cerveau humain qui est environ 10 000 fois plus dense et efficace qu'un ordinateur actuel. Deux raisons pourraient expliquer cette efficacité: une architecture particulière et la densité. Comme après un siècle de recherche, nous ne connaissons toujours pas le mode de fonctionnement du cerveau en détail mais nous savons quel rôle la densité joue sur l'efficacité, il propose de se concentrer sur ce point pour améliorer l'efficacité des ordinateurs.

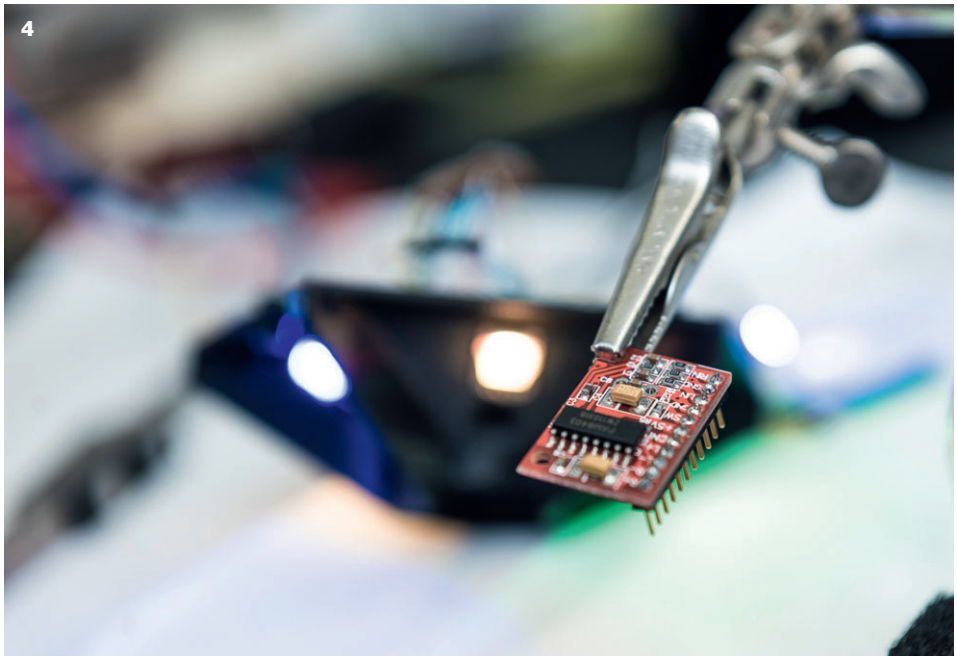
Réduire la résistance thermique

En réduisant la résistance thermique, les composants pourront être disposés de manière plus dense. Cela permettrait de réduire la consommation énergétique de la communication et de rendre le calculateur plus efficace. La première étape a été réalisée sur les supercalculateurs refroidis à l'eau Aquasar (ETH Zurich) et Supermuc du Leibniz Supercumputing Center, à Munich. Avec un volume de 50 000 m^3, ce dernier fournit une puissance de calcul de 1 PFlops (10^{15} Flops). Avec la technologie CMOS actuelle, les chercheurs d'IBM souhaitent, grâce à l'intégration 3D combinée avec un refroidissement de l'empilement de puces par un fluide qui lui procure du même coup son alimentation en courant, diviser ce volume par un facteur de 50 millions afin de rendre le système 5000 fois plus efficace. Horizon temporel: de 10 à 15 ans.

L'Internet des objets pourrait ouvrir une voie supplémentaire. Des microcalculateurs équipés de capteurs et portés sur le corps, lesdits Wearables, pourraient permettre une nouvelle plate-forme technologique nettement plus efficace et économique. Bruno Michel renvoie ici à la loi formulée en 1972 par Gordon Bell. Bell y décrit l'apparition de nouvelles catégories d'ordinateurs permettant de nouvelles applications et donc la conquête de plus grands marchés et de nouvelles industries, ce qui les rend bien moins onéreux. Un tel saut technologique a lieu tous les 10 ans. Ce développement a commencé dans les années 1960 avec les Mainframes, les mini-ordinateurs ont ensuite fait leur apparition, suivis des PC, des structures client-serveur, du Cloud Computing et aujourd'hui par les appareils mobiles (smartphones, tablettes) et les réseaux de capteurs sans fil. Pour obtenir un ordinateur aussi économique que possible, il faut miser sur les dernières technologies: les économiques Wearables.

Aujourd'hui, les Wearables sont encore souvent trop gros ainsi que pas suffisamment fiables et performants. Mais en optimisant ces aspects, on peut s'attendre à une percée dans le Computing. Le principe selon lequel nous devions aller vers notre ordinateur et apprendre son langage sera alors inversé. Nous disposerons d'un système toujours présent, utilisable de manière intuitive et qui en sait long sur la personne; un système en mesure d'identifier nos émotions et de parler notre langage.





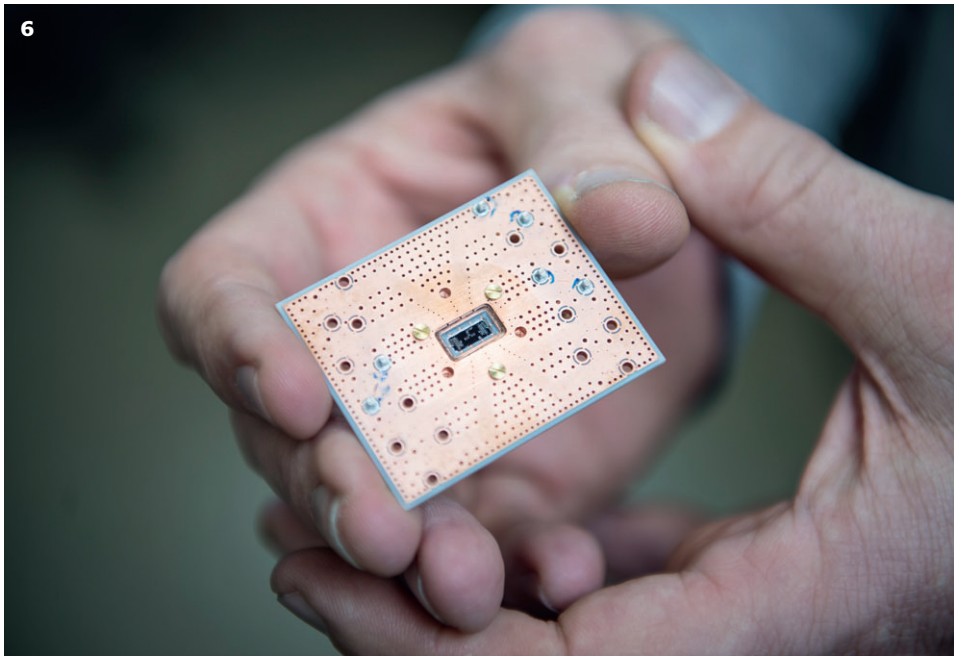
3 Mit PCM realisierte neuromorphe Computer ahmen bestimmte Funktionalitäten biologischer Neuronen und Synapsen nach.

Les ordinateurs neuromorphiques réalisés avec des PCM imitent certaines fonctionnalités des neurones et synapses biologiques.

4 Das Internet der Dinge soll künftig in miniaturisierter Form u.a. für medizinische Überwachung genutzt werden.

À l'avenir, l'Internet des objets devrait être utilisé sous une forme miniaturisée, entre autres, pour la surveillance médicale.

Bilder: Christoph Kaminski



5 Der Physiker Andreas Fuhrer forscht bei IBM in Rüschlikon an einem 2-Qubit-Quantenprozessor.

Le physicien Andreas Fuhrer effectue des recherches sur un processeur quantique de 2 qubits chez IBM à Rüschlikon.

6 Quantenrechner sind zwar klein, erfordern aber im Moment noch eine energieintensive Infrastruktur zur Kühlung und Ansteuerung.

Bien que petits, les calculateurs quantiques requièrent encore pour l'instant une infrastructure gourmande en énergie pour le refroidissement et la commande.

her vom Spital nach Hause zurückkehren, beobachten: Bei Verschlechterungen lösen sie einen Alarm aus – schon bevor eine kritische Situation eintritt. Heute ist man noch auf ein «Edge Device», den Cognitive Hypervisor, angewiesen, der zu Hause Sensordaten sammelt und dem Anwender oder den Ärzten Meldungen schickt. Künftig könnte all dies in einem «Hearable» im Ohr integriert sein. Auch im Fitness- und Wellnessbereich sind Anwendungen möglich.

Der potenzielle Markt ist beträchtlich. Neue Investoren könnten es ermöglichen, dass mehr Sensoren integriert, die Leistungsfähigkeit gesteigert und die Wearables noch kleiner und attraktiver gestaltet werden. Eine Chance für energieeffiziente, vernetzte Computer.

Man wird ein System haben, das in der Lage ist, unsere Emotionen zu erkennen und unsere Sprache zu sprechen.

Nous disposerons d'un système capable d'identifier nos émotions et de parler notre langage.

Die übernächsten Generationen

In der Forschung befasst man sich auch mit neuartigen Konzepten. Am IBM-Forschungslabor in Rüschlikon sind dies u.a. das Rechnen mit Quanten und das Neuromorphic Computing. Beide Ansätze bergen viel Potenzial: Beim ersteren würden schon rund 60 Quantenbits genügen, um zu bisher unerreichten Rechenleistungen zu kommen, z.B. um chemische Reaktionen zu simulieren. Ein Grund für diese Leistungsfähigkeit ist die Menge der durch Qubits darstellbaren Zustände: Sie können nicht nur die Zustände 0 oder 1 einnehmen, sondern auch eine Überlagerung der beiden. Ein einzelnes Qubit kann so in jedem Zustand auf einer Kugeloberfläche sein, welche 0 und 1 berührt. Diese Überlagerung von mehreren Zuständen ermöglicht paralleles Rechnen. Bei günstigen Problemstellungen wächst demnach die Rechenleistung eines Quantenrechners exponentiell mit der Anzahl der Qubits.

Das Neuromorphic Computing, ein Rechenkonzept, das die Funktionsweise des Gehirns nachahmt, hat wiederum den Vorteil, dass die Logik und der Speicher am gleichen Ort existieren können. Dies macht es schnell und sparsam. Realisiert wird dies in Rüschlikon mit Phase Change Memory Devices, einem Material, das auch in beschreibbaren DVDs eingesetzt wird. Besonders geeignet ist diese Architektur für die Mustererkennung in riesigen Datenmengen und die Datenextraktion in «verrauschten» Umgebungen. Heute hat man funktionsfähige Chips mit einigen Millionen PCM-Zellen, die in Gruppen von künstlichen Neuronen und Synapsen miteinander verbunden werden können. Es wird jedoch noch einige Jahre dauern, bis diese Art des Rechnens praxistauglich ist.

Literatur | Littérature
→ Peter Bright, «Moore's law really is dead this time», Ars Technica, 11. 2. 2016.

Autor | Auteur
Radomír Novotný ist Chefredaktor Electrosuisse.
Radomír Novotný est rédacteur en chef Electrosuisse.
→ Electrosuisse, 8320 Fehraltorf
→ radomir.novotny@electrosuisse.ch

Le spectre des applications de tels Wearables est grand. Par exemple, ils peuvent rappeler à des malades chroniques de prendre leurs médicaments ou encore observer l'état de patients rentrés plus tôt de l'hôpital: en cas de dégradation, ils déclenchent une alarme, avant même qu'une situation critique ne se produise. Aujourd'hui, nous dépendons encore d'un « Edge Device », le Cognitive Hypervisor qui collecte les données des capteurs à la maison et qui envoie des messages à l'utilisateur ou aux médecins. À l'avenir, tous ces éléments pourraient être intégrés dans un « Hearable » porté dans l'oreille. Des applications sont également possibles dans le domaine du fitness et de la remise en forme.

Le marché potentiel est considérable. De nouveaux investisseurs pourraient permettre d'intégrer plus de capteurs, d'augmenter la performance et de rendre les Wearables encore plus petits et attractifs. Une chance pour des ordinateurs efficients mis en réseau.

Les générations suivantes

La recherche s'occupe également du développement de concepts entièrement nouveaux. Au laboratoire de recherche IBM à Rüschlikon, il s'agit, entre autres, du calcul quantique et du Neuromorphic Computing. Ces deux concepts renferment un grand potentiel: concernant les calculateurs quantiques, environ 60 qubits suffiraient pour atteindre des puissances de calcul inégalées, par exemple en vue de simuler des réactions chimiques complexes. Cette performance est notamment due à la quantité d'états que les qubits permettent: les qubits ne prennent pas uniquement les états 0 et 1 mais également une superposition des deux. Un seul qubit peut être dans n'importe quel état situé sur une surface sphérique comprenant les états 0 et 1. Cette superposition de plusieurs états permet le calcul parallèle. Si le problème est posé de manière appropriée, la puissance de calcul de l'ordinateur quantique augmente de façon exponentielle avec le nombre de qubits.

Le Neuromorphic Computing s'inspire du fonctionnement du cerveau. La logique et la mémoire coexistent au même endroit, ce qui permet d'augmenter considérablement l'efficacité énergétique. À Rüschlikon, ces systèmes sont réalisés avec des Phase Change Memory Devices (PCM) fabriqués avec le même matériau que celui utilisé dans les DVD réinscriptibles. Cette architecture est particulièrement adaptée à la reconnaissance de patterns dans de gigantesques quantités de données ainsi qu'à l'extraction de données dans des environnements « bruyants ». Nous avons aujourd'hui des puces exploitables dotées de quelques millions de cellules PCM pouvant être reliées en groupes de neurones et synapses artificiels. Il pourrait cependant s'écouler encore quelques années avant d'arriver à un produit final qui puisse être utilisé en pratique.