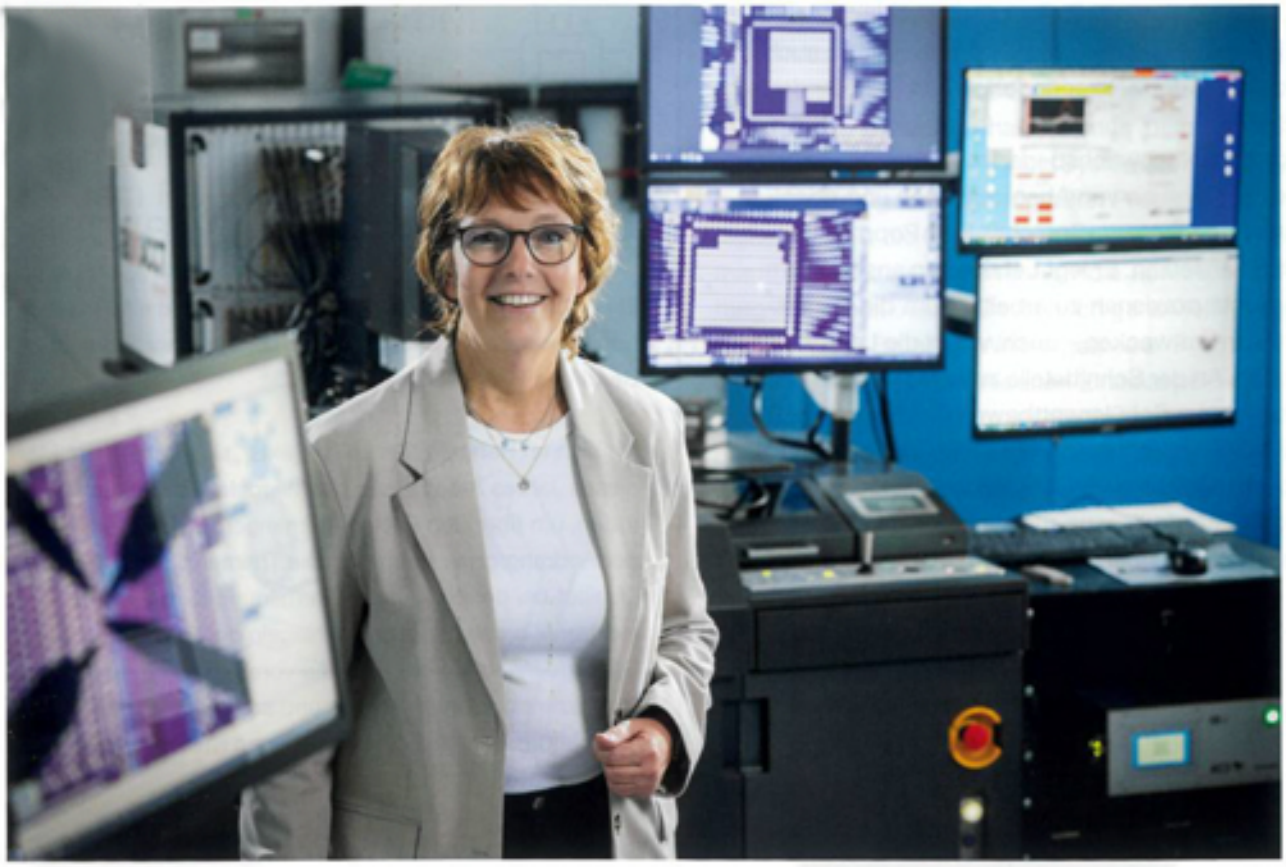


Künstliche Intelligenz

Chips nach Hirnprinzip

Selbstdenkende und -lernende Anwendungen brauchen sehr viel Strom. Forschende der Mikroelektronik arbeiten daher an neuartigen Chips, die Energie sparen, indem sie Daten dort verarbeiten, wo sie entstehen. Vorbild ist das menschliche Denkorgan.

Von Patrick Torma

Microsoft erwägt, „Three Mile Island" wieder ans Netz zu bringen. Diese Nachricht ließ doppelt aufhorchen. Das stillgelegte US-Atomkraftwerk ist für seinen Beinahe-Super-GAU von 1979 bekannt. Das Vorhaben zeigt, welche Wege der Konzern bereit ist zu gehen, um den enormen Energiebedarf seiner Anwendungen zu decken. Das betrifft nicht nur Microsoft. Weltweit treiben KI-Anwendungen den Strombedarf von Rechenzentren in die Höhe. Die Energiemengen, die etwa das Training von GPT-4 verbraucht, würden eine Kleinstadt versorgen. Das Mooresche Gesetz, das besagt, dass sich

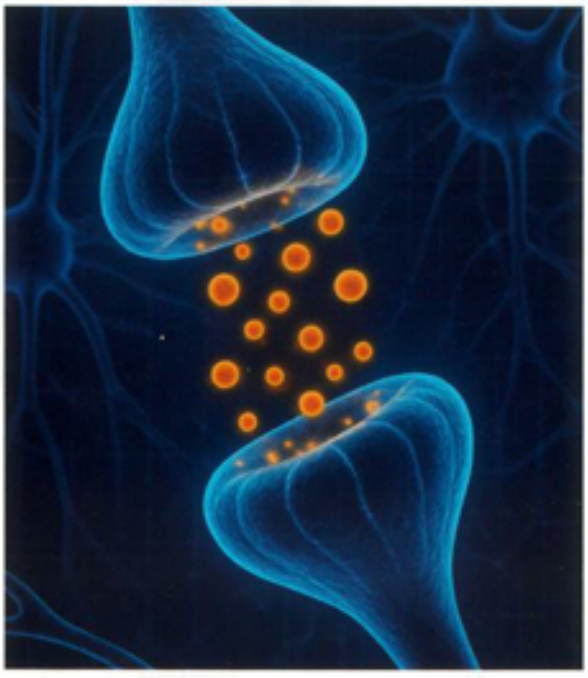
Prof. Dr. Regina Dittmann vom Forschungszentrum Jülich widmet sich unter anderem der Erforschung memristiver Mikroelektronik – einem Baustein auf dem Weg zu einer neuen Chiparchitektur.

die Anzahl der Transistoren auf integrierten Schaltkreisen etwa alle zwei Jahre verdoppelt, was zu einer Steigerung der Rechenleistung führt, schwächt sich zwar ab, doch künftige CPUs und GPUs werden weiterhin leistungsstärker. Ihre immer dichtere Chiparchitektur ist Teil des Energieproblems. Weil Prozessor und Speicher getrennt voneinander liegen, sind Daten ständig in Bewegung.

Gerade die Kerndisziplinen von KI – Mustererkennung, Analyse und Schlussfolgerung – erfordern einen enormen Datentransfer. Dieses Hin und Her geht zulasten der Latenz und kostet viel Energie. Um diesen klassischen Engpass aufzulösen, setzen aktuelle Forschungsansätze in der Mikroelektronik auf „In-Memory Computing". Die Idee: Rechenoperationen laufen in direkter Nähe zum Speicher. Die Blockbildung von Storage und Prozessoreinheiten wird aufgelöst, der energieintensive Datenverkehr minimiert. Wie Informationsverarbeitung im Energiesparmodus gelingt, macht uns die Natur vor. Unser Gehirn kommt mit einer Leistung von rund 20 Watt aus. Manche Glühlampe verbraucht mehr. Inspiriert vom menschlichen Denkapparat entwickeln Wissenschaftler Hardware-Strukturen, die an biologische Neuronen angelehnt sind. Ziel des neuromorphen Computings ist eine KI-Infrastruktur, die energieeffizient, schnell und autonom genug ist, um die Intelligenz an die Edge, sprich: so nah wie möglich an die Datenquelle bzw. in die Endgeräte, zu verlagern. Vom niedrigen Energieverbrauch könnten batteriebetriebene Sensoren oder Wearables profitieren. Statt den Akku in Stunden oder Tagen zu leeren, könnten Edge-AI-Anwendungen je nach Betriebsart monate- oder jahrelang laufen. Da der Umweg über die Cloud entfiele, blieben die Daten lokal. Die Latenz würde verringert, die Robustheit gestärkt. Selbst ohne Netz bliebe das System handlungsfähig.

Wenn Mähroboter künftig nicht mehr ihre Arbeit einstellen, sobald das Internet aussetzt, wäre das praktisch. Wichtiger sind jedoch Anwendungen jenseits reiner Komfortfunktionen. „Naheliegend ist der Einsatz in der Medizintechnik", erklärt Prof. Dr. Regina Dittmann, Leiterin des Instituts für Electronic Materials am Forschungszentrum Jülich. „Wir denken etwa an Warnsysteme, die Vitalfunktionen überwachen. Geräte, die wochenlang ohne Batteriewechsel laufen, wären ein großer Fortschritt." Dr. Markus Eppel, Gruppenleiter Advanced Analog Circuits am Fraunhofer-Institut für Integrierte Schaltungen (IIS), sieht ein weiteres dringendes Einsatzgebiet: das „Monitoring von kritischer Infrastruktur". Ein Thema, das durch den Teileinsturz der Dresdner Carolabrücke im September 2024 in den öffentlichen Fokus gerückt ist. Eppel nennt ein Beispiel aus eigener Entwicklung. 2022 präsentierte Fraunhofer eine smarte Schraube namens „Q-BO". Sie erfasst über Mikroelektronik sowohl den Anpressdruck als auch Vibrationen. Dadurch überwacht sie die Stabilität der Verbindungen. Auch Lagerschäden in Generatoren oder Motoren können erkannt werden. Noch läuft die Anomalie-Erkennung solcher Systeme meist auf klassischen Mikrocontrollern. „Deutlich effizienter, also mit längeren Wartungsintervallen oder schnelleren Reaktionszeiten, gelingt dies mit neuromorpher Hardware", ist Eppel überzeugt. Auf dieses Potenzial setzt nicht zuletzt die Automobilindustrie. Dort lockt der „heilige Gral" des vollautonomen Fahrens. Hersteller hoffen, dass neuromorphe Hardware das maschinelle Sehen voranbringt, ohne die Traktionsbatterie zu belasten.

Noch ist neuromorphe Mikroelektronik von einem breiten Einsatz entfernt. „Wir sehen erste kommerzielle Bausteine", ordnet Markus Eppel ein, doch eine „Validierung unter realen Einsatzbedingungen" stehe vielfach noch aus. In Jülich integriert das Team um Regina Dittmann neuartige Bausteine



Vorbild für ein künstliches Nervensystem: Im menschlichen Gehirn feuert ein Neuron nur, wenn wirklich etwas Relevantes geschieht – und hört nicht permanent mit, wie es aktuell noch Sprachsteuerungen tun.

FOTOS: SEITE 24: SASCHA KREKLAU / FZ JÜLICH; SEITE 25: STOCK.ADOBE.COM / BLURRED INK

in CMOS-Chips, „die Grundfunktionen neuromorphen Rechnens" demonstrieren. Systemtests laufen bislang in Simulationen. Die Suche nach wettbewerbsfähigen Netzarchitekturen führt zu unterschiedlichen technischen Konzepten. Das Fraunhofer IIS verfolgt mit seinem „Adelia"-Beschleuniger einen Mixed-Signal-Ansatz, der zur Ausführung neuronaler Netze analoge und digitale Schaltungen kombiniert. „In Tests verbrauchte unser Modell zur Sprachaktivitätserkennung bis zu 90 Prozent weniger Energie als rein digitale Ansätze – und das bei einem Genauigkeitsverlust von nicht mehr als 3 Prozent", erklärt Eppel.

Ein weiteres Konzept findet sich im „Senna"-Beschleuniger. Er führt sogenannte Spiking-Neural-Networks mit Zeitdeterminismus und extrem kurzen Antwortzeiten unter 20 Mikrosekunden aus. Dadurch soll erreicht werden, dass KI-Systeme in jedem Fall eine Entscheidung treffen – ungeachtet von Speicherengpässen, die sonst zu Verzögerungen oder Fehlern führen. Dafür verarbeitet „Senna" keine fortlaufenden Zahlenwerte, sondern kurze Impulse, sogenannte Spikes. Somit orientieren sich Spiking-Neural-Networks am Effizienzwunder Gehirn: Ein Neuron feuert nur, wenn wirklich etwas geschieht. Eine Sprachsteuerung beispielsweise wäre dadurch nicht mehr gezwungen, permanent Umgebungsgeräusche zu analysieren. Sie würde erst reagieren, wenn das Schlüsselwort fiele. Damit dieses Reaktionsmuster dauerhaft funktioniert, müssen die gelernten Verknüpfungen möglichst energiearm konserviert werden. Von Gehirnzellen inspirierte Bauteile aus Jülich sollen dem Erinnerungsvermögen neuronaler Netze auf die Sprünge helfen. Memristoren – ein Kofferwort aus Memory und Resistor – sind quasi Widerstände mit Gedächtnis. Ihr Leitwert bleibt auch dann erhalten, wenn der Strom abgeschaltet ist. Gefertigt aus Hafnium- oder Tantaloxiden sowie 2D-Materialien sollen sie der Funktionsweise von Synapsen nahekommen.

Noch findet neuromorphes Computing vor allem in Laboren statt. Doch schon innerhalb der nächsten drei Jahre könnten konkrete Anwendungen auf dem Markt sein.

»In Tests verbrauchte unser Modell zur Sprachaktivitätserkennung bis zu 90 Prozent weniger Energie als rein digitale Ansätze.«

Dr. Markus Eppel, Gruppenleiter Advanced Analog Circuits am Fraunhofer-Institut für Integrierte Schaltungen (IIS)

Um den Kreislauf des klassischen Computings zu durchbrechen, kooperiert das Forschungszentrum Jülich mit der Rheinisch-Westfälischen Universität Aachen im Projekt NEUROTEC. „Unser Ziel ist es, die gesamte Wertschöpfung abzubilden – von der Materialentwicklung über die Chip- und Systemarchitektur bis zur Software", fasst Projektleiterin Dittmann zusammen. Mit an Bord sind industrielle Partner. Parallel dazu arbeitet das Zukunftscluster NeuroSys daran, ein „Ökosystem neuromorpher Technologien" zu erschaffen. Nach dem Motto „von der Kohle zur KI" soll die Region zu einem führenden Standort für neuromorphe Elektronik werden. NEUROTEC befindet sich bereits in seiner zweiten Projektphase. Bis Ende 2026 werden Fördergelder in Höhe von 36 Millionen Euro geflossen sein. Projektleiterin Dittmann verweist auf mehrere Start-ups als Indikator für den Erfolg. Vollständig autark werde die Region jedoch nicht arbeiten: Geplant seien engere Kooperationen mit dem „Silicon Saxony" in und um Dresden, Zentrum der deutschen Halbleiterindustrie. Bis Chips mit memristiven Bauteilen aus Jülich in Serie gehen, dürfte es bis Ende des Jahrzehnts dauern. Nach Einschätzung von Regina Dittmann könnten neuromorphe Chips auf Basis konventioneller Technologien aber zeitnah marktreif sein. Konkreter wird Markus Eppel: Er sieht „einen breiten Einsatz neuromorpher Systeme in führenden Technologieclustern bereits innerhalb der nächsten zwei bis drei Jahre".