Das menschliche Gehirn als Vorbild für künftige Computing-Generationen

Holger Boche, Pit Hofmann, Gitta Kutyniok (alphabetische Reihenfolge)

Smartphones, Laptops, Tablets – Alle unsere treuen technischen Alltagsbegleiter im digitalen Zeitalter nutzen Computing, basierend auf Nullen und Einsen. Was ist jedoch, wenn das menschliche Gehirn als Inspirationsquelle für zukünftige Ansätze genutzt wird? Moderne Computing-Systeme benötigen Unmengen an Energie, insbesondere für Aufgaben der künstlichen Intelligenz, etwa für Large-Language-Modelle wie ChatGPT. Zudem stoßen klassische Von-Neumann-Architekturen, wie sie in unseren Alltagsbegleitern zum Einsatz kommen, zunehmend an ihre Grenzen. Von-Neumann-Architekturen beschreiben dabei ein Rechensystem, bestehend aus Rechenwerk, Steuerwerk, Speicher sowie Ein- und Ausgabeeinheiten. Heutige Systeme werden damit gezwungen, Daten permanent zwischen dem Speicher- und dem Rechenwerk zu bewegen, was Zeit, aber natürlich auch Energie fordert.

Hardware-Konzepte, die sich am menschlichen Gehirn orientieren, bieten eine Möglichkeit, die aktuellen Limitationen und Herausforderungen zu überwinden. Millionen von Jahren der Evolution haben eines der energie- und recheneffizientesten Systeme hervorgebracht, das wir kennen: das menschliche Gehirn. Es kann hochkomplexe Aufgaben bewältigen, und zwar mit erstaunlicher Energieeffizienz. Der zum Zeitpunkt der Abfassung dieses Artikels leistungsstärkste Supercomputer der Welt, El Capitan, verfügt über eine geschätzte Rechenleistung von rund 1,7 Exaflops [1], während die des menschlichen Gehirns auf etwa 1 Exaflop [2] geschätzt wird. Diese Leistungssteigerung spiegelt sich jedoch nicht im Energieverbrauch wider, der bei etwa 29,6 Megawatt liegt – weit entfernt von den nur 20 Watt, die das menschliche Gehirn verbraucht [3]. Neuromorphe Hardware orientiert sich an eben jenem System und eröffnet damit neue Wege für effiziente Anwendungen der künstlichen Intelligenz.

Hinzu kommt, dass klassische digitale Systeme an die Grenzen der Berechenbarkeit stoßen, wie von Forschenden der TU München und der LMU München gezeigt wurde [4,5,6]. Neuromorphe Architekturen sollen diese Lücken schließen.

Neuromorphe Informationsverarbeitung als Basis

Neuromorphe Systeme orientieren sich bei der Informationsverarbeitung an biologischen Systemen. Informationen werden nicht als kontinuierliche Signale weitergegeben, sondern auf Basis von Ereignissen in Form diskreter elektrischer Impulse, sogenannter Aktionspotentiale (oder auch Spikes). Ein Modell, das derartiges Verhalten von Neuronen beschreibt, ist das Leaky-Integrate-and-Fire-Modell: Die empfangenen Signale werden über die Zeit integriert, und sobald ein Schwellenwert erreicht wird, erzeugt das Neuron einen Spike. Dabei wird der Unterschied zur klassischen Informationsverarbeitung deutlich, indem nur Rechenaufwand entsteht, wenn ein Ereignis auftritt. Informationen lassen sich damit in Zeit kodieren, beispielsweise durch den zeitlichen Abstand zwischen mehreren Spikes, ohne dabei Energie zu verbrauchen. In Bezug auf entsprechende Hardware-Implementierungen bedeutet dies einen minimalen Energieaufwand, solange keine Spikes (und damit Ereignisse) verarbeitet werden müssen.

Memristoren als Schlüsselkomponenten

Für die Entwicklung neuromorpher Hardware spielen insbesondere Memristoren eine zentrale Rolle. Memristoren bezeichnen elektronische Bauteile, deren Bezeichnung aus Memory (Speicher) und Resistor (elektrischer Widerstand) abgeleitet ist. Der elektrische Widerstand eines Memristors hängt von der Richtung und dem Betrag der angelegten Spannung ab. Damit ähneln Memristoren biologischen Synapsen, bei denen sich die Stärke einer Verbindung basierend auf einer gemeinsamen Aktivität gemäß dem Prinzip „Neurons that fire together wire together" [7,8] ändert. Memristoren bilden damit die Schlüsselkomponente für In-Memory-Computing-Architekturen, in denen Computing und Speicher zusammenfallen. Nueronale Netze können so ohne die ständigen Datenbewegungen klassischer Systeme trainiert werden.

Technologielandschaft neuromorpher Systeme

In der technischen Landschaft neuromorpher Systeme existieren weltweit mehrere Flaggschiffprojekte, die das Potential neuromorpher Systeme unterstreichen. Mit TrueNorth stellte IBM einen neuromorphen Prozessor vor, der nicht auf der Von-Neumann-Architektur basiert und energieeffizient arbeiten kann. Mit Loihi von Intel und dessen Nachfolger Loihi 2 wird sogar spiketime-abhängige Plastizität direkt in der Hardware unterstützt, was diese Systeme insbesondere für die Robotik interessant macht. Das Projekt SpiNNaker mit starker deutscher Beteiligung aus Dresden verfolgt hingegen einen parallelen Ansatz, um neuronale Systeme in Echtzeit zu simulieren.

Neuromorphe Systeme als Zukunft der KI?

Mit dem nahezu quadratischen Wachstum der weltweit erzeugten Datenmenge, wie in Abbildung 1 (links) dargestellt, steigt auch der dafür erforderliche Energiebedarf, beispielsweise für Speicherung, Verarbeitung und Transport, deutlich an. Selbst im theoretischen Grenzfall, in dem die Informationsverarbeitung ausschließlich durch das fundamentale physikalische Minimum beschrieben wird, wächst der Energieverbrauch proportional zur Menge der verarbeiteten Informationen. Dieses Minimum wird durch das Landauer-Limit definiert, das als Brücke zwischen Informationstheorie und Thermodynamik die minimale Energie festlegt, die bei einer irreversiblen Bit-Operation unvermeidlich verlorengeht. Damit führt die stetig zunehmende Datenproduktion selbst unter optimalen Bedingungen zwangsläufig zu einem steigenden globalen Energiebedarf. Abbildung 1 (rechts) zeigt diese Entwicklung im Verhältnis zur weltweiten Energieerzeugung. Während das Landauer-Limit die absolute physikalische Untergrenze markiert, liegen selbst hardwarebeschleunigte Architekturen deutlich darüber, wenn auch wesentlich effizienter als herkömmliche Prozessoren. Doch auch unter sehr günstigen Annahmen deuten Projektionen darauf hin, dass der Energiebedarf der Informationsverarbeitung langfristig in eine Größenordnung vorrückt, die mit der heutigen globalen Energieproduktion vergleichbar ist. Diese Aussicht macht klar, dass reine Effizienzsteigerungen nicht ausreichen, um den absoluten Bedarf an Energie auszugleichen. Stattdessen braucht es disruptive technologische Ansätze wie neuromorphes Computing, welches das Potential hat, den Energieverbrauch um ein Vielfaches zu senken.

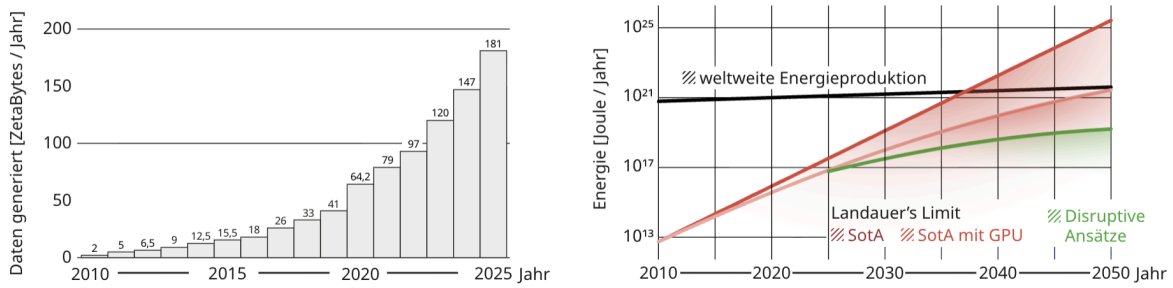

Abbildung 1: Weltweit generierte *Datenmenge von 2010 bis 2025 (links); Energieproduktion und -nutzung weltweit für den Stand der Technik und prognostiziert für disruptive Ansätze (rechts). Datensatz basierend auf [9].*

Ohne Kommunikation keine KI

Während neuromorphe Systeme in Bezug auf Energieeffizienz einen maßgeblichen Anteil liefern können, bleibt oft vergessen, dass ohne leistungsfähige Kommunikationsinfrastrukturen und -systeme nur ein Bruchteil des Potentials ausgeschöpft werden kann. Moderne KI-Anwendungen sind nicht mehr auf einzelne Knoten im Netzwerk beschränkt, sondern umfassen ein breites Set verteilter Pipelines, die Daten zwischen Sensorik, Edge-Geräten, Funknetzen, Rechenzentren und der Cloud austauschen. Für Anwendungsszenarien, wie beispielsweise das autonome Fahren oder die kollaborative Robotik, müssen neuromorphe Systeme in Echtzeit interagieren unter Zugrundelegung niedriger Latenz, hoher Resilienz und Zuverlässigkeit. Erst 5G- bzw. 6G-fähige Architekturen können diese Anforderungen zuverlässig erfüllen und sind damit unverzichtbar.

Technologische Souveränität als Standbein

Für Deutschland stellt neuromorphes Computing eine Chance dar, um mit Hilfe alternativer Ansätze die technologische Souveränität auf dem KI-Markt zu sichern. Im globalen Marktumfeld dominieren derzeit primär Akteure aus Asien und den USA. Durch Forschungsprogramme, starke universitäre oder universitätsnahe Partner und eine Beteiligung an europäischen Initiativen im Bereich der Mikroelektronik entsteht jedoch ein Ökosystem, das Deutschland in die Lage versetzt, (nicht nur) neuromorphe Systeme selbst zu entwickeln, um eine Unabhängigkeit von außereuropäischen Lieferketten zu erreichen und eine Vorreiterrolle im neuromorphen Computing zu übernehmen.

Herausforderungen und Forschungsfragen

Neuromorphe Systeme stehen trotz ambitionierter Visionen für die Zukunft des Computings vor mehreren Herausforderungen. Einerseits ist die Fertigung memristiver Bauteile technisch anspruchsvoll, insbesondere hinsichtlich der Stabilität, des Temperaturdrifts oder der langfristigen Zuverlässigkeit. Andererseits sind bestehende KI-Frameworks auf klassische neuronale Netze ausgelegt und nur eingeschränkt auf spikende neuronale Netze anwendbar. Eng damit verbunden ist auch die fehlende Orchestrierung in bestehende Infrastruktur, das heißt, die Zuweisung von Computing-Aufgaben an heterogene Systemkomponenten, beispielsweise neuromorphe Architekturen. In der Forschung werden deshalb derzeit großskalierte, spike-basierte Modelle entwickelt, welche mit klassischen Ansätzen koexistieren sollen, um eine industrielle Anwendung in der Zukunft zu ermöglichen.

Ausblick: What comes next?

Angesichts des steigendes Interesses der Öffentlichkeit an energieeffizienter KI und der wachsenden Nachfrage nach nachhaltigen Technologien, rücken die Bestrebungen um disruptive Ansätze, wie neuromoprhes Computing, zunehmend in den Fokus nationaler und internationaler Forschungsstrategien. Exemplarisch genannt sei an dieser Stelle das Projekt “Next Generation AI Computing (gAIn)”, ein Forschungsprojekt, gefördert sowohl vom Sächsischen Staatsministerium für Wissenschaft, Kultur und Tourismus als auch vom Bayerischen Staatsministerium für Wissenschaft und Kunst. Im Konsortium sitzt neben der LMU München und der TU München auch die TU Dresden. Zusammen erforschen die Konsortialpartner disruptive Computing-Ansätze, unter anderem neuromorphes Computing.

Begriffserklärungen:

- **Von-Neumann-Architektur:** Die Von-Neumann-Architektur beschreibt ein Computersystem, welches typischerweise aus einem Rechenwerk, einem Steuerwerk, einem Speicher sowie Eingabe- und Ausgabe-Einheiten besteht. Programme und Daten werden im selben Speicher abgelegt.
- **Exaflop:** Ein Exaflop ist die Maßeinheit für die Rechenleistung eines Computers. Ein Exaflop pro Sekunde entspricht so einer Trillion Gleitkommaoperationen pro Sekunde.
- **spiketime-abhängige Plastizität**: Die spiketime-abhängige Plastizität beschreibt einen Mechanismus in der Neurobiologie, bei dem die Verbindungen zwischen Neuronen im Gehirn durch eine zeitliche Abfolge von Spikes verändert bzw. angepasst werden.

Referenzen

[1] Cooled Exascale Supercomputer, „El Capitan”, for Lawrence Livermore National Laboratory,” Nov. 2024 [Online]. Verfügbar: https://www.hpe.com/us/en/newsroom/press-release/2024/11/hewlett-packard-enterprise-delivers-worlds-fastest-direct-liquid-cooled-exascale-supercomputer-el-capitan-for-lawrence-livermore-national-laboratory.html

[2] Advait Madhavan, „Brain-inspired computing can help us create faster, more energy-efficient devices — If we win the race,” März 2023 [Online]. Verfügbar: https://www.nist.gov/blogs/taking-measure/brain-inspired-computing-can-help-us-create-faster-more-energy-efficient

[3] L. Smirnova, B. S. Caffo, D. H. Gracias, Q. Huang, I. E. Morales Pantoja, B. Tang, D. J. Zack, C. A. Berlinicke, J. L. Boyd, T. D. Harris, E. C. Johnson, B. J. Kagan, J. Kahn, A. R. Muotri, B. L. Paulhamus, J. C. Schwamborn, J. Plotkin, A. S. Szalay, J. T. Vogelstein, P. F. Worley, und T. Hartung, „Organoid Intelligence (OI): The new frontier in biocomputing and intelligence-in-a-dish,” Frontiers in Science, Bd. 1, Feb. 2023, doi: 10.3389/fsci.2023.1017235.

[4] H. Boche, A. Fono und G. Kutyniok, „Limitations of deep learning for inverse problems on digital hardware“, IEEE Transactions on Information Theory, Bd. 69, Nr. 12, S. 7887–7908, Okt. 2023, doi: 10.1109/TIT.2023.3326879.

[5] H. Boche, Y. N. Böck, C. Deppe und F. H. P. Fitzek, „Remote state estimation and Blum-Shub-Smale machines - A computability analysis with applications to virtual-twinning“, IEEE Transactions on Automatic Control, Bd. 70, Nr. 5, S. 3165–3180, Mai 2025, doi: 10.1109/TAC.2024.3502314.

[6] H. Boche, R. F. Schaefer, H. V. Poor und F. H. P. Fitzek, „On the need of neuromorphic twins to detect denial-of-service attacks on communication networks“, IEEE/ACM Transactions on Networking, Bd. 32, Nr. 4, S. 2875–2887, März 2024, doi: 10.1109/TNET.2024.3369018.
[7] S. Löwel und W. Singer, „Selection of intrinsic horizontal connections in the visual cortex by correlated neuronal activity,” Science, Bd. 255, Nr. 5041, S. 209–212, Jan. 1992, doi: 10.1126/science.1372754.
[8] D. O. Hebb, „The Organization of Behavior,” New York, NY, USA: Wiley, 1949.
[9] J. Ang, D. Apalkov, und F. Assaderaghi, „Decadal plan for semiconductors – Full report,” Jan. 2021 [Online]. Verfügbar: https://www.src.org/about/decadal-plan/