



More Moore or beyond Moore? Innovationen in Zeiten von DeepSeek

Bevor der Ingenieur Gordon Earle Moore 1968 zusammen mit Andy Grove und Robert Noyce den Halbleiterhersteller Intel gründete, veröffentlichte er 1965 – vor genau 60 Jahren – einen Artikel, der später in das sogenannte *Moore'sche Gesetz* mündete. Nach dieser Prognose verdoppelt sich die Anzahl der Transistoren integrierter Schaltkreise (ICs) in regelmäßigen Abständen (etwa alle 18 Monate).

Anwendungen künstlicher Intelligenz, wie z. B. große Sprachmodelle, haben einen immensen Bedarf an Rechenleistung und damit auch an Energie.

Die spannende Frage ist, wie sich dieser Bedarf zukünftig decken lässt: durch eine weiterhin zunehmende Anzahl von Transistoren auf einem Chip, was rein physikalisch herausfordernd bleiben wird, oder durch neue Technologieansätze, die über das Moore'sche Gesetz hinausgehen.

Der vorliegende Beitrag skizziert mögliche Antworten auf diese Frage, insbesondere vor dem Hintergrund der neusten Entwicklungen von Sprachmodellen, die nicht immer auf sogenannten „Transformern“ basieren müssen

More Moore – Halbleiterentwicklung von 1965 bis 2025¹

In den letzten 60 Jahren ist es stets gelungen, die Anzahl der Transistoren auf einem Silizium-Chip in bestimmten Zeitabständen zu verdoppeln. Diese, etwa 18-monatigen, Wegpunkte werden auch als Technologieknoten bezeichnet und beschreiben eine stetige Miniaturisierung bis in den atomaren Bereich hinein. Anfangs ließen sich nur etwa 50 Transistoren aus wirtschaftlicher Perspektive auf einem Chip integrieren. Heute, nach Investitionen von Hunderten von Milliarden Dollar, sind am sogenannten 7 nm Technologieknoten auf einem Quadratmillimeter Silizium etwa 100 Millionen Transistoren integriert. Derzeit sind digitale Prozessoren, die auf Basis der 5 nm Knoten-Technologie hergestellt werden, in Produktion und der 1 nm Knoten scheint in den nächsten Jahren in Sicht zu sein – 1 nm entspricht in etwa der Breite von fünf Siliziumatomen. Der Siegeszug dieser hohen Integration ist seinerzeit durch die Planartechnologie möglich geworden, die in Reinräumen zur Perfektion optimiert wurde. Kernelement dieser hochkomplexen und überaus präzisen Fertigung in den sogenannten „Foundries“, wie zum Beispiel der Taiwan Semiconductor Manufacturing Company (TSMC) ist die Complementary-Metal-Oxide-Semiconductor (CMOS) Prozesslinie. Doch nun stößt dieser Optimierungsprozess an physikalische Grenzen.

Ist dies dann das „Ende der Fahnenstange“? Wird es keine weiteren Sprünge in der Rechenleistung durch Fortschritte in der Halbleiterherstellung geben?

Die Antwort ist: „nein“ – der alleinige Fokus auf Technologieknoten verschleiert, dass es durchaus weitere Möglichkeiten gibt, wie die klassische Halbleitertechnologie zukünftiges Computing vorantreiben wird.

Die Metriken zur Messung der Integrationsdichte waren in erster Linie Dimensionen, die als *Metal Half-Pitch* und *Gate Length* bezeichnet werden. Metal Half-Pitch ist der halbe Abstand zwischen zwei Metallverbindungen auf einem Chip. Die Gate Length misst im zweidimensionalen, planaren Transistordesign den Raum zwischen den Source- und Drain-Elektroden eines Transistors. Historisch war dies die wichtigste Dimension für die Bestimmung der Transistorleistung, da eine kürzere Gate Length auf größere Dynamik beim Schalten zwischen zwei definierten binären Zuständen hindeutete.

Bis zur Mitte der 1990er Jahre waren Metal Half-Pitch und Gate Length die definierenden Merkmale der Chipherstellung und wurden zur „Knotennummer“. Diese Merkmale auf dem Chip wurden in der Regel mit jeder Generation um 30 Prozent verkleinert. Mitte der 1990er Jahre jedoch begannen sich die beiden Merkmale zu entkoppeln. In dem Bestreben, die Verbesserungen bei Geschwindigkeit und Geräteeffizienz fortzusetzen, haben die Chiphersteller die Gate Length stärker verkleinert als andere Funktionen. Ergebnis war zwar die Fortsetzung des Dichteverdopplungsweges nach Moore, jedoch mit einer unverhältnismäßig schrumpfenden Gate Length. Dennoch hielt man sich größtenteils noch an die hergebrachte Knotenbenennungskonvention.

Mittlerweile schlägt die IEEE International Roadmap for Devices and Systems (IRDS) vor, eine drei-stellige Metrik einzuführen, welche den *Contacted Gate-Pitch* (G), den *Metal-Pitch* (M) und – was für zukünftige Chips entscheidend ist – die Anzahl der Schichten auf dem Chip, die *Tiers* (T), kombiniert.

Die Gate-Pitch- und Metal-Pitch-Werte dieser *GMT-Metrik* werden bis etwa 2030 weiter abnehmen. Dann wird die Grenze dessen erreicht sein, was man mit der Photolithographie im *Extreme-Ultraviolet* (EUV) erreichen kann und das ist der Zeitpunkt, an dem die Anzahl der Schichten, Tiers (T), wichtig wird. Die Realisierung zweier Transistorschichten wird schließlich die Dichte der Bauelemente wieder fast verdoppeln – ein Grund für die Halbleiterindustrie weiter in Richtung monolithischer 3D-ICs zu forschen. Dies geschieht auch in Deutschland, beispielsweise in dem Sonderforschungsbereich zwischen der TU Dresden und der RWTH-Aachen, dem TRR 404 „*Active-3D*“.

Der taiwanesischer Hersteller TSMC hatte im Rahmen der Halbleitermesse IEDM im Dezember 2024 Details zu geplanten Verbesserungen der kommenden Fertigung bereitstehender Technologieknoten bekanntgegeben. Neben den Transistoren, steht dabei auch die Verdrahtung und die Anbindung an andere Chips im Fokus, also das „*Packaging*“. Voraussichtlich im zweiten Halbjahr 2025 soll die neue N2-Fertigung (N2 \Rightarrow 2 nm) in die Massenproduktion starten. So spricht TSMC wahlweise von einer um 15 % höheren Performance oder einer um 30 % höheren Energieeffizienz. Bei niedrigen Spannungen wird die Energieeffizienz angeblich sogar vervierfacht. Die oft als Kenngröße verwendete Dichte von Speicherzellen „*SRAMs*“ soll stärker ansteigen. N2 schaffe angeblich 37,9 Mbit pro mm². Neben den Transistoren selbst, bei denen TSMC erstmals auf sogenannte *GAAFETs* (*Gate-All-Around Field-Effect-Transistors*) statt *FinFETs* (*Fin Field-Effect-Transistors*) setzt, werden Optimierungen an der Verschaltung genannt.

Einen besonderen Fokus legt TSMC mit der N2-Fertigung auch auf Chiplets. Einerseits werde eine neue Verdrahtungsebene eingeführt, um die Bonding-Anschlüsse für den Kontakt mit einem weiteren Chip passend zu positionieren. Andererseits wolle TSMC die sogenannten *Through Silicon Vias* (TSVs), mit denen der darüberliegende Chip angebunden wird, optimieren. Für Chiplets gebe es optimierte TSVs, welche die Chips miteinander verbinden und eine neue Kontaktschicht, die durch eine Passivierung nicht mit dem darüberliegenden Chip reagiert.

3D-ICs, welche die Transistorzahl pro Volumen und nicht nur pro Fläche erhöhen, weisen damit bereits den Weg von der klassischen Halbleitertechnik („*more Moore*“) hin zu Ansätzen einer Technologie „*beyond Moore*“. Weitere Technologieansätze in dieser Richtung sind Gegenstand des nächsten Abschnittes.

Beyond Moore – ein Blick in die Zukunft

Neben der Entwicklung von 3D-ICs (siehe oben) kommen weitere Technologien in Betracht, um den Bedarf des Maschinellen Lernens nach immer mehr Rechenpower und Energie zu stillen. Im Folgenden werden Entwicklungen in der Hardwarearchitektur sowie ein aktueller algorithmischer Ansatz skizziert.

Neuromorphic Computing (NMC)

NMC zielt als Ansatz für Algorithmen- und Hardware-Design darauf ab, die Funktionsweise von biologischen Signalverarbeitungsprozessen, wie beispielsweise von Neuronen, nachzuahmen. Im Vergleich zu herkömmlichen, sogenannten *von-Neumann-Rechnern* haben biologische Systeme eine geringe Energieaufnahme, da Informationen nicht unter großem Aufwand und Energiebedarf zwischen Recheneinheit (*Arithmetic Logic Unit - ALU*) und Speicher hin- und hergeschoben werden müssen. Ein prominenter Ansatz des NMC besteht darin, relativ einfache, abstrakte Modelle biologischer Neuronen und Synapsen zu erstellen. Der sogenannte *von-Neumann-Flaschenhals* wird umgangen, indem die Begrenzung zwischen Speicher und ALU in einem einzigen Bauelement genutzt wird (\Rightarrow *Computing in memory*). Energieverluste werden damit minimiert und eine hochgradig parallele Datenverarbeitung ermöglicht – genau die Plattform, die für Maschinelles Lernen notwendig ist. *Computing in memory* ist ein grundlegend neuer Ansatz in der Chiparchitektur. Wesentlich dabei ist die Durchführung sogenannter *Vektor-Matrix-Multiplikationen*, welche die Basis der meisten Machine- bzw. Deep-Learning-Algorithmen sind. Machine-Learning-Aufgaben sind rechenintensiv, jedoch nicht unbedingt komplex. In einem neuromorphen Computer wird der Algorithmus ganz wesentlich durch die Architektur des Systems definiert und nicht durch die sequentiell getaktete Ausführung von Anweisungen innerhalb einer ALU. Dies führt zu geringeren Verarbeitungszeiten bzw. Latenzen.

In der analogen Signalverarbeitung nutzt man unterschiedliche Eigenschaften der Bauelemente aus, um Rechenoperationen, wie die Superposition, das Integrieren sowie Differenzieren von Signalwerten, innerhalb eines Kontinuums von Leitwerten zwischen 0 und 1 durchzuführen. Auf dem Weg dorthin bieten memristive Bauelemente (*Memristor – engl. aus memory und resistor*) die Möglichkeit der Durchführung von Rechenoperation in klar definierten Verarbeitungsstufen zwischen den Schaltwerten 0 und 1. So speichern zum Beispiel Phasenwechselspeicher (PCMs) KI-Modellgewichte in den Leitwerten eines Materialphasenzustandes zwischen amorphen und kristallinen Phasen. Die veränderte Leitfähigkeit ändert den Wert von Matrixmultiplikationsoperationen, wenn sie durchlaufen werden. Nachdem ein KI-Modell in Software trainiert wurde, werden alle synaptischen Gewichte in diesen PCMs gespeichert – ähnlich wie Erinnerungen in biologischen Synapsen.

Übergeordnetes Ziel ist es, Systeme zu entwickeln, die sich – wie biologische Zellen – dynamisch an das zu lösende Problem anpassen. Angestrebt wird dabei eine möglichst dichte Packung auf einem Chip, vergleichbar der hohen Integrationsdichte aktueller Technologieknoten der Digitaltechnik. Die Effizienz solcher Systeme basiert darauf, dass diese mit deutlich weniger Energieaufnahme auskommen und „lernende“ Strategien zur Signalverarbeitung nutzen, anstatt jeden Wert in 0 und 1 zu digitalisieren.

Ein Nachteil dieser analogen Ansätze besteht darin, dass sie derzeit nur auf Inferenz beschränkt sind. Für das Training können sie nur bedingt verwendet werden, weil die Genauigkeit beim Einstellen der Gewichte noch stark variiert. Gewichte können in PCM-Zellen fixiert werden, sobald ein KI-Modell auf einer digitalen Architektur trainiert wurde, aber die direkte Änderung der Gewichte durch das Training ist noch nicht präzise genug. Auch sind PCMs (noch) nicht langlebig genug, um ihren Leitwert mehrere Millionen Mal zu ändern, wie es beim Training notwendig wäre.

Der VDE hat daher im September 2024 eine Spezifikation (VDE SPEC) zum schichtenübergreifenden Entwurf solcher Ansätze mit Forschenden aus Industrie und Akademia veröffentlicht. Diese Spezifikation soll den Wettbewerb um die effizientesten Schaltungen messbar und validierbar machen.

NMC eignet sich bisher am besten für Edge-Anwendungen, zum Beispiel für den Einsatz in Smartphones, für autonomes Fahren, Verarbeitung von Sensordaten in Echtzeit, in Robotikanwendungen, Bilddatenerkennung sowie in audiovisuellen Anwendungsfällen. Der Vorteil entsprechender Chips in Edge-Anwendungen besteht darin, dass sie außergewöhnlich klein, leistungsstark und kostengünstig sind, aber eben auch sehr energieeffizient sein können.

Photonik im Machine- bzw. Deep-Learning²

Die klassische digitale Signalverarbeitung auf Basis von CMOS-Bauelementen ist überaus komfortabel zu skalieren, wie es beispielsweise derzeit die Firma NVIDIA mit den vielseitigen Parallelisierungen der CPU- (Central Processing Units) und GPU- (Graphics Processing Unit)- oder eben der XPU-Cluster wirtschaftlich überaus erfolgreich demonstriert. Dennoch stoßen auch diese Skalierungsprozesse an grundlegenden Einschränkungen, da jede noch so kleine Schalteinheit als Nebenprodukt auch ohmsche Verluste und damit Verlustabwärme erzeugt. Hinzu kommen weitere, physikalisch bedingte, inhärente Einschränkungen, wie Limitierungen in Bandbreite, Latenz und Energieeffizienz. Es wird erwartet, dass der prognostizierte weltweite Stromverbrauch von Rechenzentren, die mit CMOS-Chips betrieben werden, bis 2026 um einen Betrag steigen wird, der dem jährlichen Verbrauch eines weiteren europäischen Landes entspricht. Optische Bauelemente in Silizium aber auch auf anderen Substraten entwickeln sich daher zu einer vielversprechenden, energieeffizienten, CMOS-kompatiblen Alternative, die verstärkt in Deep-Learning-Beschleunigern zum Einsatz kommen wird. Photonen werden sowohl für schnelle als auch für energieeffiziente Kommunikationstechnik extensiv genutzt. Erste vielversprechende Ansätze der optischen Signalverarbeitung in Form von photonischen Prozessoren sind bereits auf dem Weg in die Anwendung. Die Entwicklungszyklen sind genauso dynamisch, wie in der Halbleitertechnologie, da vergleichbare Prozesslinien genutzt werden.

In jüngster Zeit haben sich GPUs und XPUs zum Standard bei der Skalierung von Rechenzentren entwickelt, um insbesondere die Anforderungen einer Parallelverarbeitung zu bedienen. NVIDIA hatte sich mit seinem 2020 vorgestellten A100-Beschleuniger zu einem Branchenführer entwickelt.

Doch während diese ICs eine hervorragende Leistung in Bezug auf Geschwindigkeit und Skalierbarkeit zeigen, ist ihr Energiebedarf extrem hoch. „Allein im Jahr 2023 hat NVIDIA 100.000 Systemeinheiten ausgeliefert, die jährlich durchschnittlich 7,3 TWh Strom aufnehmen. Der Energiebedarf von Rechenzentren teilt sich auf die benötigte Stromversorgung (40 %) für die Rechenknoten und den Kühlbedarf (40 %), der Rest entfällt auf die zugehörige Recheninfrastruktur.“

Auf der Suche nach Möglichkeiten, die Skalierbarkeit zu erhöhen und um den wachsenden Anwendungsanforderungen zu genügen, geraten photonische Beschleuniger zunehmend in den Blick. Sie bauen auf photonischen Technologien, wie Modulatoren, Photodetektoren und optischen Filtern auf, welche für die Implementierung von Rechenoperationen angepasst werden. Im Gegensatz zu herkömmlichen elektronischen Bauelementen, wie Transistoren nutzen photonische Beschleuniger Photonen, um Informationen zu verarbeiten. Sie nutzen die Eigenschaften des Lichts, um eine parallele Verarbeitung und eine schnelle Informationsübertragung zu ermöglichen, bei gleichzeitig geringerer Energieaufnahme und höherer Effizienz pro Fläche. Der Weg in die elektronisch-photonische Integration ist längst geöffnet.

„Seit den frühen 1980er Jahren wurden viele wichtige Entwicklungsstufen bei optischen Bauelementen und integrierten Silizium-Photonik-Schaltkreisen erreicht, wie z. B. Wavelength Division Multiplexing (WDM-Filter), Mach-Zehnder-Interferometer und In-Phase/Quadratur-Modulatoren. Diese Entwicklung setzte sich mit dem Aufkommen kleinerer Mikroring-Resonatoren (MRRs) fort, die in vielen optischen Filterdesigns von entscheidender Bedeutung sind, so auch in Hochgeschwindigkeits- oder NRZ-Modulatoren (Non-Return-to-Zero) mit großer Bandbreite. Darüber hinaus wurden PAM4-Modulationsschemata (Pulse Amplitude Modulation with Four Levels) untersucht, bei denen Ringresonatoren verwendet werden, um den Durchsatz pro Bereich des Geräts zu erhöhen. Diese Ringresonatoren mit hohen Q-Faktoren wurden so konstruiert, dass sie sowohl bei optischen als auch bei Terahertz-Frequenzen als Basiselemente wie Schalter oder für Rechenoperationen wie Integration, Differenzierer und Speicherelemente fungieren.“

Optisches Rechnen wurde bisher bei Anwendungen, die einen großen Datenspeicher und eine effiziente Flusssteuerung erfordern, skeptisch betrachtet. „Aktuelle Forschungen zeigen jedoch die Fähigkeiten photonischer Beschleuniger für Anwendungen, die sich gut für den Einsatz in Rechenzentren eignen. Zu diesen Anwendungen gehören Prozesse, bei denen eine hohe Parallelität Voraussetzung ist, die durch nicht-kohärente Optik durch WDM, Polarisationsdiversität und Modenmultiplexing, erreicht wird.“

Die Integration von nichtflüchtigen PCMs als photonische Bauelemente ermöglicht die optische Datenspeicherung und das Computing in Memory.

In jüngerer Zeit wurden diese Geräte integriert, um energieeffiziente, kompakte und hochdurchsatzfähige Rechenbeschleuniger zu entwickeln. Eine vergleichende Analyse der theoretischen, maximalen Tera-Operationen pro Sekunde pro Quadratmillimeter (TOPs/mm²) sowohl für elektronische als auch für photonische Beschleuniger zeige einen klaren Vorteil im photonischen Bereich.

Zusammenfassend lässt sich sagen, dass viele Deep-Learning-Operationen durch photonische Bauelemente teilweise oder vollständig stark beschleunigt werden können. Sie können eine bemerkenswerte Performanz in der Informationsverarbeitung erreichen, bei gleichzeitig deutlich geringerer Energieaufnahme im Vergleich zu ihren elektronischen Pendanten.

Quantencomputing³

Quantencomputing wird zwar auf konventioneller Hardware ausgeführt, diese muss jedoch extremen physikalischen Bedingungen genügen, um die dafür zur Verfügung stehenden Algorithmen verwenden zu können. Das Ziel ist, komplexe Probleme zu lösen, die klassische Computer nicht oder nicht schnell genug oder nur unter deutlich größerem Rechenaufwand lösen können. „Anders als die heutigen High-Performance-Computing-Cluster beruhen Quantenalgorithmen auf neuen Ansätzen für die Lösung komplexer Problemstellungen und schaffen multidimensionale Rechenräume.“ Ein Quantencomputer verwendet Quantenbits (Qubits), um multidimensionale Quantenalgorithmen auszuführen.

Um beispielsweise die Zahlen 0 bis einschließlich 15 darzustellen, benötigt man vier klassische Bits. Ein Qubit kann, ebenso wie das klassische Bit, den Zustand „1“ oder „0“ repräsentieren. Um mit Qubits effektiv zu rechnen, werden quantenphysikalische Effekte der Überlagerung (Superposition) und der Verschränkung genutzt. Wie mit vier klassischen Bits können auch mit vier Qubits die einzelnen Zahlen von 0 bis 15 dargestellt werden. „Die Rechenleistungssteigerung beim Qubit besteht nun darin, dass es sich in einem Superpositionszustand befindet, in dem es mit einer gewissen Wahrscheinlichkeit den Zustand „1“ und „0“ zeigt. Aufgrund des Superpositionsprinzips können mit vier Qubits deshalb auch simultan alle Zahlen von 0 bis 15 auf einmal dargestellt werden. In jedem

Rechenvorgang wird dann das Ergebnis für all diese Zahlen gleichzeitig parallel berechnet. Während man also beim klassischen Computer die Berechnung für jede einzelne Zahl einzeln nacheinander durchführen muss, kann man beim Quantencomputer mit nur vier Qubits schon eine Berechnung für alle Zahlen von 0 bis 15 gleichzeitig durchführen. Dies ist durch die Kombination von Verschränkung und Superposition möglich. Dies ist nur einer der Vorteile des Quantencomputers.“

Ein Quanten-Prozessor an sich ist nicht viel größer, als ein digitaler Prozessor. „Das gesamte Quantenhardwaresystem dagegen hat etwa die Größe eines Autos und besteht hauptsächlich aus Kühlsystemen, um den supraleitenden Prozessor auf seiner extrem niedrigen Betriebstemperatur, nahe dem absoluten Nullpunkt, zu halten. Bei diesen Temperaturen lassen sich Quantenzustände beibehalten, die sogenannte Dekohärenz wird vermieden. Auch zeigen bei extrem niedrigen Temperaturen bestimmte Materialien einen wichtigen quantenphysikalischen Effekt, indem sich Elektronen ohne Widerstand durch sie hindurchbewegen können – dies macht sie zu Supraleitern. Wenn Elektronen Supraleiter passieren, treffen sie aufeinander und bilden sogenannte *Cooper-Paare*. Diese Paare können durch einen Prozess, der als *Quantentunneln* bekannt ist, eine Ladung über Barrieren oder Isolatoren transportieren.“

„Die Programmierung oder vielmehr die Manipulation von Quantenzuständen in Quantencomputern, zum Beispiel in Anwendungen der Firma IBM, wird über sogenannte *Josephson-Kontakte* als supraleitende Qubits durch Mikrowellenschaltungen vorgenommen. Dadurch ist es möglich Qubits zu beeinflussen und Operationen wie Speicherung oder das Auslesen zu ermöglichen.“

Ein Qubit selbst ist nur eingeschränkt nützlich. Es kann jedoch die Quanteninformation, die es enthält, in einen Überlagerungszustand versetzen, der eine Kombination aller möglichen Konfigurationen des Qubits darstellt. Mit Gruppen von überlagerten Qubits können komplexe, mehrdimensionale Rechenräume erzeugt werden. Komplexe Probleme können damit auf neue Art und Weise dargestellt werden.

Quantenverschränkung ist ein Effekt, bei dem sich Änderungen an einem Qubit direkt auf ein anderes Qubit auswirken, wenn diese zwei Qubits miteinander verschränkt sind. Diesen Effekt macht man sich insbesondere in der Quantenkommunikation zunutze.

„In einer Umgebung verschränkter Qubits, die in einen Überlagerungszustand gebracht werden, gibt es Wellen von Wahrscheinlichkeiten. Dies sind die Wahrscheinlichkeiten der Ergebnisse einer Messung des Systems. Die Wellen können sich durch Interferenz gegenseitig verstärken, oder sie können sich gegenseitig aufheben.“

Eine Berechnung auf einem Quantencomputer funktioniert, indem eine Überlagerung aller möglichen Rechenzustände vorbereitet wird. Ein vom Benutzer vorbereiteter Quantenschaltkreis nutzt nach einem Algorithmus selektiv Interferenzen an den Komponenten der Überlagerung aus. Viele mögliche Ergebnisse werden durch Interferenzen aufgehoben, während andere verstärkt werden. Diese verstärkten Ergebnisse stellen die Lösungen für die Berechnung dar.“

Ein Großteil der Arbeit auf dem Gebiet des Quantencomputings widmet sich der Realisierung von Fehlerkorrekturen, um rauschfreie Quantenberechnungen zu ermöglichen.

Im Folgenden werden vier weitere Computing-Ansätze skizziert, die sich der Thematik „beyond Moore“ zuordnen lassen:

Reversibles Computing⁴

Im Jahre 1961 entdeckte der Physiker und Informationswissenschaftler Rolf Landauer von IBM, dass das Löschen einer Information in einem Computer zwangsläufig Energie kostet, die als Wärme verloren geht. Er kam auf die Idee, dass wenn man Berechnungen durchführen würde, ohne jegliche Information zu löschen, zumindest theoretisch rechnen könnte, ohne überhaupt Energie aufwenden zu müssen. Landauer selbst hielt die Idee für unpraktikabel, denn wenn man alle Eingaben und Zwischenergebnisse speichern würde, würde man den Speicher schnell mit unnötigen Daten füllen.

Landauers Nachfolger, Charles Bennett von IBM (seines Zeichens auch Entdecker der Quantenteleportation), entwickelte einen Workaround für dieses Problem. Anstatt nur Zwischenergebnisse im Speicher zu speichern, könnte man die Berechnung umkehren (quasi „rück-berechnen“), sobald das Ergebnis nicht mehr benötigt wird. Auf diese Weise müssten nur die ursprünglichen Eingaben und das Endergebnis gespeichert werden.

Als einfaches Beispiel dient das Exclusive-OR-Gatter (XOR-Gatter). Normalerweise ist dieses Gate nicht reversibel – es gibt zwei Eingänge und nur einen Ausgang. Wenn man den Ausgang kennt, erhält man keine vollständige Information darüber, was die Eingänge waren. Die gleiche Berechnung kann reversibel durchgeführt werden, indem ein zusätzlicher Ausgang hinzugefügt wird, eine Kopie einer der ursprünglichen Eingaben. Dann können mit den beiden Ausgängen die ursprünglichen Eingaben in einem Rück-Berechnungsschritt wiederhergestellt werden.

In den 1990er Jahren begannen mehrere Studenten am MIT eine Reihe von Proof-of-Principle-Demonstrationen von reversiblen Computerchips. Diese Demonstrationen zeigten zwar, dass reversible Berechnungen möglich waren, aber der Energieaufwand wurde nicht unbedingt reduziert. Der elektrische Strom wurde zwar innerhalb der Schaltung zurückgeführt, ging jedoch anschließend in die Versorgung der externen Bauelemente ein.

Heutige CMOS-Implementierungen benötigen mehr als tausendmal so viel Energie, um ein Bit zu löschen, als theoretisch möglich ist. Obwohl diese hocheffizient arbeiten, sind sie weit entfernt vom sogenannten „Landauer-Limit“. Das liegt daran, dass Transistoren für ihre Zuverlässigkeit definierte Signalenergien aufrechterhalten müssen. Beim Schalten werden jedoch Bereiche der Strom-Spannungs-Kennlinie durchfahren, die zu ohmschen Verlusten führen. Die wichtigste Möglichkeit, unnötige Wärmeentwicklung beim Einsatz von FET-Transistoren zu reduzieren – sie adiabatisch zu betreiben – besteht darin, die Schaltzeiten gleich zu halten und die Wellenform zu ändern, die das Schalten ausführt. Adiabatisches Schalten erfordert jedoch – auch bei komplexeren Ramping-Wellenformen – immer noch Energie, um ein Bit von „0“ auf „1“ zu schalten und die Gate-Spannung eines Transistors von seinem niedrigen in den hohen Zustand zu ändern. Ein Ansatz bestehe darin, dass man, solange man elektrische Energie nicht in Wärme umwandelt, sondern den größten Teil davon im Transistor selbst speichert, einen Teil dieser Energie während des „Rück-Rechnungsschritts“ wiedergewinnt, wobei jede nicht mehr benötigte Berechnung umgekehrt wird. Der Weg, diese Energie zurückzugewinnen, könne darin bestehen, den gesamten Schaltkreis in einen Resonator einzubetten.

Ein Resonator ist vergleichbar einem schwingenden Pendel. Gäbe es keine Reibung, würde das Pendel ewig schwingen und bei jedem Schwung auf die gleiche Höhe steigen. Genau wie bei einem mechanischen Pendel ist dies jedoch auch in der elektronischen Schaltung nicht so und der Vorgang ist mit Verlusten behaftet. Hier ist das „Ausschwingen des Pendels“ ein Auf und Ab der Spannung, welche die Schaltung antreibt. Bei jedem Aufschwung wird ein Rechenschritt durchgeführt. Bei jedem Abschwung wird eine „Rück-Berechnung“ durchgeführt, bei der die elektrische Energie zurückgewonnen wird.

Durch die Einbettung der Schaltung in einen Resonator werden gleichzeitig die komplexeren Wellenformen erzeugt, die für das adiabatische Schalten der Transistoren erforderlich sind und es wird der Mechanismus zur Rückgewinnung der eingesparten Energie bereitgestellt.

Zur Zeit wird an einem Chip gearbeitet, der die Multiplikations-Akkumulationsoperation (MAC-Operation) ausführen kann – die grundlegende Berechnung in den meisten Anwendungen des maschinellen Lernens. Forschende erwarten in den nächsten 10 bis 15 Jahren eine 4.000-fache Leistungssteigerung. Doch ist dies noch Gegenstand der Grundlagenforschung.

Supraleitende (Computer-) Chips⁵

Ein supraleitender Chip würde ohne Leitungsverluste arbeiten und ließe sich dadurch mit deutlich höheren Frequenzen schalten als Siliziumhalbleiter. Das Grundproblem ist jedoch, dass ein Supraleiter Strom in jede Richtung gleich gut leitet und damit das Prinzip des Halbleiters nicht zur Anwendung kommt. Halbleiter dagegen, die Grundlage digitaler Computer, leiten nur nach Dotierung und können zur Steuerung des Stromflusses genutzt werden. Der Strom erfährt in Leitungsrichtung einen geringen Widerstand, in die Gegenrichtung (Sperrrichtung) hingegen einen großen. Durch die Kombination unterschiedlich dotierter Bereiche kann Strom nur in eine Richtung fließen. Zwar kann dieses, als nicht-reziproke Leitung bezeichnete, Verhalten auch bei Supraleitern erreicht werden, allerdings waren hierzu bislang extern erzeugte Magnetfelder erforderlich, die eine Miniaturisierung unmöglich machten. Forschende nutzten den sogenannten *Josephson-Effekt*, der das Tunneln von Elektronenpaaren zwischen zwei Supraleitern durch eine nicht-supraleitende Schicht beschreibt. Ein *Josephson-Kontakt* ermöglicht eine elektrische Steuerung des Tunnelns zwischen den beiden Supraleitern. Das klingt nach Transistor, jedoch ist der Effekt reziprok – die Elektronenpaare tunneln in beide Richtungen gleich. Durch das Applizieren von magnetischen Feldern können Elektronen nur noch in eine Richtung tunneln. Der Josephson-Effekt bleibt erhalten, das Tunneln kann durch eine angelegte Spannung gesteuert werden. Damit würden prinzipiell supraleitende Chips möglich, die sich in aktuelle Rechnerarchitekturen implementieren lassen.

Da die Supraleiter gekühlt werden müssen, ist ein Einsatz nur in Rechenzentren und Hochleistungsrechnern praktikabel. Ziel ist es, ein System zu entwickeln, das bei 77 Kelvin (-196° C) supraleitend ist. Dies würde eine Kühlung mit Flüssigstickstoff ermöglichen.

Spintronik Computing⁶

Spintronik ist vielversprechend für die Entwicklung energieeffizienter Schalter jenseits der CMOS-Technologie. Kern von spintronischen Bauelementen sind dünne magnetische Materialien, deren magnetische Momente kollektiv auf verschiedene Eingangsanregungen, wie elektrischen Strom, Spinstrom, Spannung, Dehnung oder Magnetfeld reagieren. Damit ist eine Modulation zur Informationsverarbeitung durchaus praktikabel umsetzbar. Aufgrund dieser kollektiven Reaktion kann ein Nanomagnet theoretisch seine Magnetisierung mit einer Energie schalten, die 10.000-mal niedriger ist als diejenige, welche bei FET-Transistoren benötigt wird.

Die Energieeffizienz bezieht sich bisher auf die Schaltungsebene, also die der Einzelbauelemente, muss sich aber letztendlich auf den gesamten Prozessor erstrecken, was eine Herausforderung für diese Technologie darstellt, um die Effizienz bewährter CMOS-Architekturen auf Systemebene zu erreichen. Obwohl der höhere Wirkungsgrad des magnetischen Schaltens physikalisch bestätigt ist, hat die Forschung in der Spintronik noch keinen zuverlässigen Schalter hervorgebracht, der bei niedrigen Energieniveaus arbeitet.

Aufgrund dieser Herausforderungen hat sich die Spintronik in andere Bereiche des Computings ausgeweitet, die weniger direkt mit der klassischen Digitaltechnik konkurrieren, wie zum Beispiel High-Performance-Computing und unkonventionelles Computing.

„Die Zukunftsvision des Hochleistungsrechnens könnte eine heterogene Integration funktions-spezifischer Bausteine beinhalten. Darüber hinaus wurden Spin-Transfer-Torque Magnetic Random-Access Memory (STT-MRAM) und Spin-Orbit Torque Magnetic Random-Access Memory (SOT-MRAM) bereits erfolgreich für die Datenspeicherung kommerzialisiert oder befinden sich in der Entwicklung. Über den Speicher hinaus können alternative Rechenschemata auf Spintronik Vorteile bei der Energieeffizienz auf Systemebene bieten, indem sie die einzigartigen Eigenschaften magnetischer Materialien geschickt nutzen. Einer dieser Bereiche ist das unkonventionelle Computing, zu dem NMC, Computing in Memory, probabilistisches oder stochastisches Computing und das Computing in extremen Umgebungen, wie dem Weltraum gehören, wo die intrinsische Strahlungshärte magnetischer Materialien deutliche Vorteile bietet.“

Obwohl unkonventionelles Computing eine wichtige Rolle spielt, wird das Gesamtvolumen des Marktes für systemintegrierte Schaltkreise von digital programmierbaren Architekturen dominiert. Diese Architekturen zeichnen sich durch bekannte, befehlsprogrammierte Prozessoren, eine Speicherhierarchie und Kommunikationsnetzwerke aus, die insbesondere auf dem hochvolumigen Mobil- und Servermarkt weit verbreitet sind und derzeit hauptsächlich mit skaliertem Silizium-CMOS-Technologie implementiert werden.

DeepSeek – ein neuer, algorithmischer Modellansatz für mehr Leistung und Energieeffizienz bei Nutzung konventioneller Digitalhardware⁷

Das in China ansässige Startup DeepSeek konkurriert mit seinem Modell, DeepSeek-V3 nicht nur mit etablierten Tech-Giganten, wie OpenAI, Anthropic und Meta in der Leistung, sondern übertrifft sie auch in der Kosteneffizienz. Gleichzeitig macht DeepSeek trainierte Modelle öffentlich zugänglich.

Bestehende Large Language Models (LLMs) weisen folgende Einschränkungen auf:

- Ineffiziente Ressourcennutzung – Die meisten Modelle basieren auf dem Hinzufügen von Ebenen und Parametern, um die Leistung zu steigern – ein Ansatz, der immense Hardwareressourcen erfordert.
- Engpässe bei der Verarbeitung langer Sequenzen – Herkömmliche LLMs verwenden Transformer als grundlegendes Modelldesign. Transformer haben mit Speicheranforderungen zu kämpfen, die exponentiell wachsen, wenn sich die Eingabesequenzen verlängern.
- Trainingsengpässe aufgrund von Kommunikations-Overhead – Beim Modelltraining kommt es aufgrund des GPU-Kommunikationsaufwands häufig zu Ineffizienzen. So kann die Datenübertragung zwischen Knoten zu erheblichen Leerlaufzeiten und damit verbundenen Kosten führen.

DeepSeek-V3 überwindet diese Einschränkungen durch innovatives Design der verwendeten Modelle und realisiert Effizienz, Skalierbarkeit und hohe Leistung durch folgende Ansätze:

- Intelligente Ressourcenallokation durch sogenannte *Mixture-of-Experts (MoE)* – DeepSeek-V3 verwendet eine MoE-Architektur, die selektiv 37 Milliarden Parameter pro Token aktiviert. Rechenressourcen werden dort eingesetzt, wo sie benötigt werden.
- Effizientes Handling langer Sequenzen mit sogenanntem *Multi-Head Latent Attention (MHSA)* – Im Gegensatz zu LLMs, die auf Transformer-Architekturen basieren, die speicherintensive Caches zum Speichern von Rohdatenwerten erfordern, verwendet DeepSeek-V3 einen innovativen MHSA-Mechanismus. Dieser verändert die Art und Weise, wie Caches für

Rohdatenwerte verwaltet werden. Nur die wichtigsten Informationen werden weiterverarbeitet, während unnötige Details verworfen werden. Die Speicherauslastung wird reduziert – das System wird schneller und effizienter.

- Gemischtes Präzisionstraining mit dem 8-Bit-Gleitkommaformat, FP8 – hochpräzise Formate, wie FP16 oder FP32 erhöhen die Speicherauslastung und damit die Rechenkosten. DeepSeek-V3 verfolgt mit seinem FP8-Mixed-Precision-Framework einen innovativeren Ansatz. Durch die intelligente Anpassung der Präzision an die Anforderungen jeder Aufgabe reduziert DeepSeek-V3 die GPU-Speicherauslastung und beschleunigt das Training, ohne die numerische Stabilität und Leistung zu beeinträchtigen.
- Lösen von Kommunikations-Overhead mit *DualPipe* – DeepSeek-V3 verwendet ein innovatives DualPipe-Framework, um Berechnungen und Kommunikation zwischen GPUs zu überlappen. Dies ermöglicht es dem Modell, beide Aufgaben gleichzeitig auszuführen, was die Leerlaufzeiten reduziert, in denen GPUs auf Daten warten.

Effizienz und Wirtschaftlichkeit von DeepSeek-V3 beim Training

Das Modell wurde mit einem Datensatz von 14,8 Billionen hochwertigen Token über etwa 2,8 Millionen GPU-Stunden auf NVIDIA H800 GPUs trainiert. Dieser Trainingsprozess wurde mit Gesamtkosten von rund 5,57 Millionen USD durchgeführt, während GPT-4o von OpenAI Berichten zufolge über 100 Millionen USD für das Training benötigte.

Überlegene Inferenz-Fähigkeit

Der MHLA-Mechanismus stattet DeepSeek-V3 mit einer außergewöhnlichen Fähigkeit aus, lange Sequenzen zu verarbeiten, die es ihm ermöglicht, relevante Informationen dynamisch zu priorisieren. Diese Fähigkeit ist besonders wichtig für das Verständnis langer Kontexte, die für Aufgaben wie mehrstufiges Schlussfolgern nützlich sind. DeepSeek-V3 verwendet in einem modularen Ansatz Reinforcement-Learning, um MoE mit kleineren Modellen zu trainieren.

Energieeffizienz:

Mit dem 8-Bit-Gleitkommaformat FP8 und DualPipe-Parallelität minimiert DeepSeek-V3 die Energieaufnahme bei gleichbleibender Genauigkeit. Dies reduziert die GPU-Leerlaufzeit bei gleichzeitig steigender Energieeffizienz. Das aktuelle Beispiel von DeepSeek-V3 zeigt eindrücklich, dass Fortschritt nicht auf Kosten der Effizienz gehen muss.

Fazit

Mittlerweile gleichen die Energiekosten für die Kühlung von Rechenzentren denen des Serverbetriebs. Daher ist es dringend notwendig, Rechenprozesse auf Systemebene zu optimieren, um die unerwünschte Abwärme zu minimieren.

In modernen, digitalen Rechnersystemen entstehen Energieverluste an verschiedenen Stellen, unter anderem durch dynamisches Schalten, Standby-Energieaufnahme, Speicherzugriffe sowie durch parasitäre Verluste, wie Verbindungsaufwände und die Umwandlung zwischen Analog- und Digitaltechnik. Auch Kühlsysteme für zentrale Einheiten tragen zur Gesamtenergieaufnahme bei. Um die Energieeffizienz auf Systemebene zu steigern, müssen Ansätze, wie die hier skizzierten, zu „Beyond-Moore“ nicht nur auf Bauelementebene weiterentwickelt werden, sondern auch mit der hoch-optimierten Digitaltechnik, die auf dem etablierten CMOS-Prozess basiert, wettbewerbsfähig bleiben.

In naher Zukunft sind signifikante Innovationsschübe auf der Ebene der Algorithmen, wie gerade aktuell zu „DeepSeek“, zu erwarten. Doch um den enormen Bedarf an Rechenressourcen durch die intensive Nutzung generativer KI zu decken, bedarf es eines Perspektivwechsels in dem Entwurf der Hardware. Der Wettlauf um die effizienteste und am weitesten verbreitete Plattform hat gerade erst begonnen. Als Vertreter der Informationstechnischen Gesellschaft (ITG) im VDE möchten wir mit der hier vorliegenden, kurzen Zusammenstellung der aktuellen Technologien aufzeigen, welche Alternativen und Chancen bestehen, in den laufenden Wettbewerb einzutreten. Neben der Weiterentwicklung von „Transformern“ in großen Sprachmodellen spielt auch die Hardwareinfrastruktur eine entscheidende Rolle in diesem Wettbewerb, wobei diese nicht ausschließlich auf Digitaltechnik beschränkt ist.

¹ Moore, S.K. (2020). A Better Way to Measure Progress in Semiconductors. IEEE Spectrum. Verfügbar unter: <https://spectrum.ieee.org/a-better-way-to-measure-progress-in-semiconductors> (letzter Zugriff: 07.01.25)

² Atwany M, Pardo S, Serunjogi S and Rasras M (2024). A review of emerging trends in photonic deep learning accelerators. Front. Phys. 12:1369099. doi: 10.3389/fphy.2024.1369099

³ IBM. Was ist Quantencomputing. Verfügbar unter: <https://www.ibm.com/de-de/topics/quantum-computing#:~:text=Quantencomputing%20nutzt%20spezielle%20Technologien%20%E2%80%93%20einschlie%C3%9Flich,nicht%20schnell%20genug%20%C3%B6sen%20k%C3%B6nnen> (letzter Zugriff: 07.01.25)

⁴ Genkina, D. (2025). Reversible Computing Escapes the Lab in 2025. IEEE Spectrum. Verfügbar unter: <https://spectrum.ieee.org/reversible-computing> (letzter Zugriff: 07.01.25)

⁵ Hiltscher, J. (2022). Supraleiter: Elektronen-Einbahnstraße für schnelle, verlustfreie Computer. Verfügbar unter: <https://www.golem.de/news/supraleiter-elektronen-einbahnstrasse-fuer-schnelle-verlustfreie-computer-2205-165398.html> (letzter Zugriff: 07.01.25)

⁶ Incorvia, J.A.C., Xiao, T.P., Zogbi, N. et al. (2024). Spintronics for achieving system-level energy-efficient logic. Nat Rev Electr Eng 1, 700–713. Verfügbar unter: <https://doi.org/10.1038/s44287-024-00103-z> (letzter Zugriff: 07.01.25)

⁷ Zia, T. (2025). DeepSeek-V3: How a Chinese AI Startup Outpaces Tech Giants in Cost and Performance. UNITE.AI. Verfügbar unter: <https://www.unite.ai/deepseek-v3-how-a-chinese-ai-startup-outpaces-tech-giants-in-cost-and-performance/> (letzter Zugriff: 28.01.25)

Dr. Damian Dudek & Dr. Matthias Wirth

VDE Verband der Elektrotechnik
Elektronik Informationstechnik e.V.
Merianstraße 28
63069 Offenbach am Main
Tel. +49 69 6308-360
damian.dudek@vde.com
matthias.wirth@vde.com