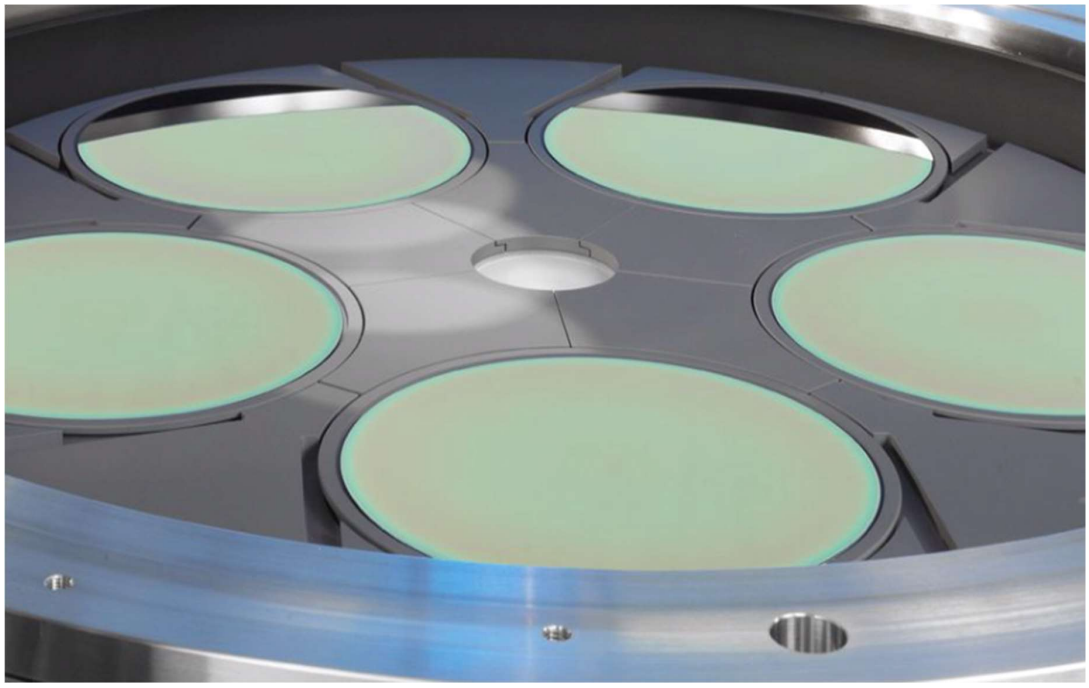
WerteFinder
VermögensVerwaltung

Aufkommende Schlüsseltechnologien der digitalen Zukunft



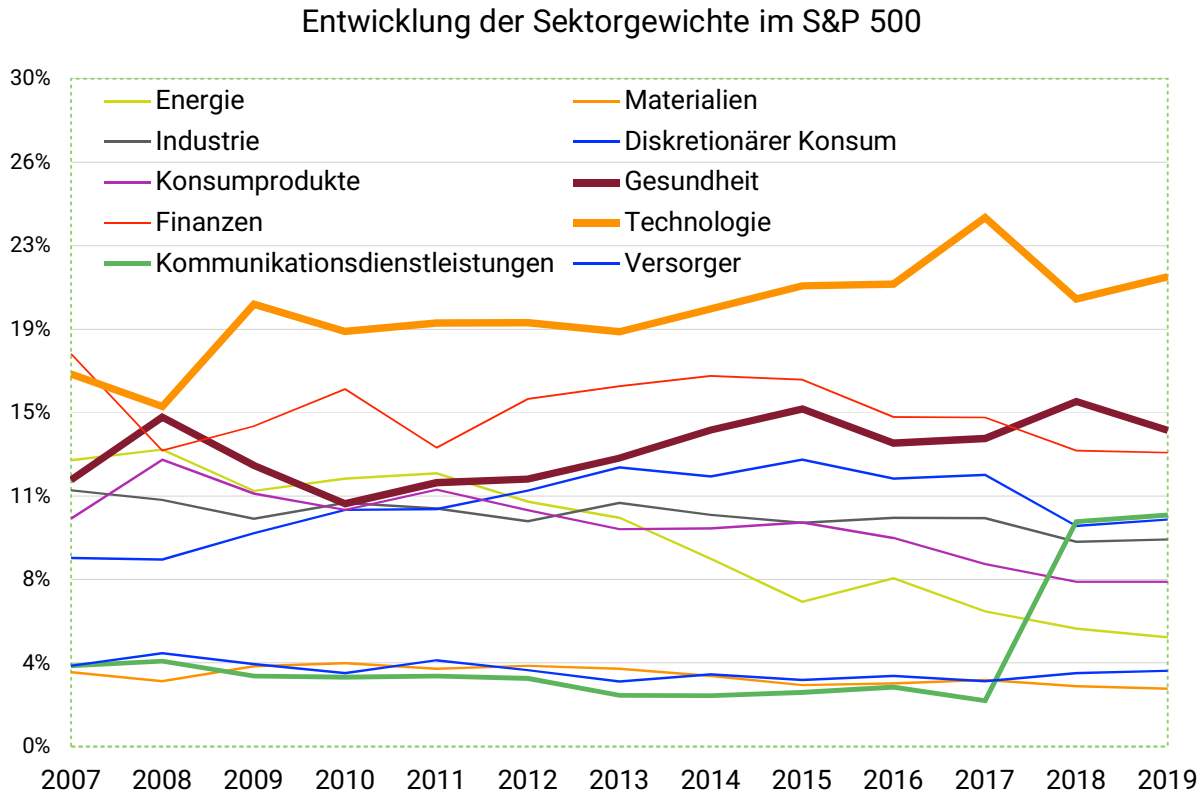
Dies ist ein internes Arbeitspapier der Wertefinder Vermögensverwaltung.
Dieses Arbeitspapier ist nicht für eine Veröffentlichung oder Verteilung vorgesehen.
Bitte vertraulich behandeln.

Inhalt

1. Vorbemerkung **3**
1.1. Motivation 4
1.2. Grundlagen 5
1.3. Anwendungen 5
2. Halbleiterindustrie **6**
2.1 Bauelemente 6
2.2 Memristor 7
2.3 Ziele 9
3. Technologische Lösungen **10**
3.1 Vorbemerkung 10
3.2 Wertschöpfungskette 11
3.2.1 Electronic Design Automation (EDA) 12
3.2.2 Packaging (Advanced) 13
3.2.3 Metrologie 14
3.3 Von-Neumann-Architektur 15
3.3.1 Speichertechnologie 15
3.3.1.1 High Bandwith Memory 15
3.3.1.2 Nichtflüchtige Speicher 17
3.3.2 Fazit 24
4. Neuromorphes Rechnen **25**
5. Materialien **29**
6. Schlussfolgerung **35**

1. Vorbemerkung

Die heutige Welt zeichnet sich nicht nur durch Umwälzungen in Gesellschaft, Politik und Wirtschaft aus, sondern vor allem durch den immer stärkeren Einfluss komplexer Technologien. Ein sichtbarer Beweis aus der Perspektive der Finanzmärkte ist das zunehmende Gewicht des Technologiesektors im richtungsweisenden amerikanischen Index S&P 500, der die 500 größten Unternehmen des Landes abbildet (s. nachstehenden Chart).



„Alte“ Branchen werden in immer schnellerem Maße verdrängt und es entstehen vollkommen neue Geschäftsmodelle, die wir uns heute noch nicht einmal vorstellen können. Die technologischen Errungenschaften verlaufen teilweise relativ planbar (z.B. Miniaturisierung in der Halbleiterindustrie), teilweise aber auch in Sprüngen (wie z.B. die Nutzung von GPUs bei neuronalen Netzwerken, die plötzlich ab dem Jahr 2012 eine signifikante Steigerung der Rechenleistung ermöglichte).

Als Vermögensverwalter sind wir gehalten, neue Entwicklungen frühzeitig zu erkennen, um nicht an Unternehmen festzuhalten, deren Bewertung zwar vermeintlich billig ist, die aber aufgrund des technologischen Wandels kein zukunftsfähiges Geschäftsmodell mehr besitzen. Vor dem Hintergrund einer vorausschauenden Sicht auf neue Entwicklungen hatten daher wir bereits vor fast drei Jahren in einem Bericht die damals wenig bekannte Entstehung des mobilen Internets der fünften Generation (5G) durchleuchtet. Inzwischen sind in vielen Ländern die ersten Frequenzen für 5G versteigert worden und der Aufbau der zugehörigen Ausrüstungen wird mit unterschiedlicher Schnelligkeit in den einzelnen Regionen und Ländern der Welt vorangetrieben. 5G ist eine

Schlüsseltechnologie, die viele der seit Jahren im Gespräch befindlichen neuen Anwendungen erst ermöglichen wird, aber selbst ebenfalls vollkommen neuer Schlüsseltechnologien bedarf.

Den neuen Entwicklungen in Forschung und Technik liegen überwiegend Fortschritte auf der Ebene der Halbleitertechnologie und der künstlichen Intelligenz (KI) zugrunde. Hier handelt es sich ebenfalls um Schlüsseltechnologien, die zusammen mit dem 5G die Grundlage für die neuen Anwendungen darstellen. Diese Schlüsseltechnologien bedingen sich jedoch auch wechselseitig, d.h. dass z.B. KI zur Effizienzsteigerung neuer Fortschritte bei der Halbleitertechnologie bedarf und umgekehrt bei der Halbleiterentwicklung auch Methoden der KI eingesetzt werden.

1.1. Motivation

An den Finanzmärkte gibt es eine Fülle von Investmentfonds, die nach Themen investieren. Vor einigen Jahren war es das Thema Umwelt, speziell Solarenergie, und in jüngster Zeit sind es Themen wie „Smart Mobility“, „Smart Grid“, „Robotics“, „Künstliche Intelligenz“, „Wasser“, „Infrastruktur“, etc., die auf der Suche nach Anlagegeldern sind. In Wertefinder verfolgen wir einen anderen Ansatz: Obwohl wir auch in Firmen investieren, die in der einen oder anderen Form den modernen Themen zuzuordnen sind, legen wir einen Schwerpunkt auf Unternehmen und Themen, die in der Öffentlichkeit weniger bekannt sind, die jedoch am Ursprung vieler technologischer Prozesse stehen. Es handelt sich hier um Unternehmen mit einer technologischen Vorherrschaft in Schlüsselthemen, die für die weitere Entwicklung ausschlaggebend sind und die oft am Anfangspunkt der gesamten Wertschöpfungskette der Mikroelektronik und damit der Anwendungen stehen.

Teilweise wird an den Produkten und Technologien schon seit Jahren geforscht und eine kommerzielle Vermarktung steht nicht unmittelbar bevor. Ein Beispiel für unseren Ansatz ist z.B. unser Einstieg vor knapp 3 Jahren in das Thema 5G. Hier ging es uns darum, in Firmen zu investieren, die die vollkommen neuen Schlüsseltechnologien für den Aufbau des Netzes herstellen.

Mit diesem Text möchten wir verdeutlichen, dass wir an der Schwelle zu bahnbrechenden neuen technologischen Entwicklungen stehen, die in Zukunft eine deutliche Steigerung von Übertragungsraten, Rechenleistung, Energieeffizienz und Zuverlässigkeit ermöglichen. Es handelt sich dabei nicht um den Quantencomputer, sondern um Technologien auf der ersten Stufe der Halbleiterindustrie und Mikroelektronik. In diesem Text richten wir das Augenmerk daher schwerpunktmäßig auf Firmen, die neue Schlüsseltechnologien entwickeln bzw. die Entwicklung dieser Schlüsseltechnologien ermöglichen. Da die bisherigen Schlüsseltechnologien an physikalische Grenzen stoßen, wenden wir uns den Unternehmen zu, die an der Vorfront neuer technologischer Entwicklungen bzw. Materialien stehen.

Zu einem besseren Verständnis dieser vollkommen neuen Herausforderungen an die Halbleitertechnik gehen wir in diesem Text auch kurz auf die technologischen Grundlagen der Halbleitertechnik ein. Dieser Text mag manchem Leser zu „technisch“ erscheinen, aber für ein tieferes Verständnis ist eine detaillierte Betrachtung unumgänglich. Gerade die letzten Monate haben gezeigt, dass Technologien, die zuvor als Nischenprodukt angesehen wurden, bald den Weg in die Serienproduktion finden. Wesentlich ist, dass die bisherige Architektur an physikalische Grenzen stösst und daher die Entwicklung und Durchdringung der neuen Technologien mit Hochdruck von Forschung und Herstellern vorangetrieben wird.

Aus dem Text sollte auch deutlich werden, dass es auf technologischer Ebene eine Fülle von vollkommen unterschiedlichen Ansätzen zur Lösung der hohen Anforderungen der neuen Anwendungen gibt, wobei der Grad der Kommerzialisierung jedoch sehr unterschiedlich ist.

1.2. Grundlagen

Ausgangspunkt vieler neuer technologischer Entwicklungen im Bereich KI, 5G und Halbleitern ist der Druck zu höherer Energieeffizienz um kleinere Bauteile und höhere Rechenleistungen zu ermöglichen. Gleichzeitig erfordert die außerordentlich hohe Zunahme von Daten eine Beschleunigung der Rechenleistung, die mit der herkömmlichen Halbleiterarchitektur kaum noch zu bewerkstelligen ist. Die Forschung auf diesen Ebenen stellt gleichsam die erste Stufe des technologischen Fortschritts dar. Obwohl die technologischen Entwicklungen auf dieser Stufe damit die Grundlagen für alle weiteren Technologieschritte darstellen, sind sie weniger bekannt und erreichen selten die Aufmerksamkeit der Medien. Wir möchten mit diesem Bericht diese Lücke füllen und konzentrieren uns auf eine Reihe neuerer Entwicklungen, die von hoher Bedeutung sind. Es handelt sich um Themen, die teilweise schon seit Jahrzehnten bekannt und deren Forschung sich teilweise noch auf Laborebene befindet. Trotzdem ist die Beschäftigung mit den Forschungsergebnissen hilfreich, da sie uns einen Blick in die Zukunft ermöglicht. Letztendlich geht es natürlich auch darum, die Geschäftsmodelle der beteiligten Unternehmen besser einzuschätzen und so von wichtigen Technologietrends frühzeitig zu profitieren. Wir glauben außerdem, dass es vorteilhafter ist in Unternehmen zu investieren, die führend bei der Herstellung von Schlüsseltechnologien sind als in Unternehmen zu investieren, die Endprodukte herstellen. Bei letzteren Unternehmen ist oftmals auch unklar, welches Produkt sich am Markt durchsetzt. Außerdem sind auf der Anwenderebene die Margen aufgrund hoher Konkurrenz bzw. niedriger Markteintrittsbarrieren in vielen Fällen geringer. Die Technologieführer bei den Schlüsseltechnologien oder Schlüsselkompetenzen profitieren jedoch von einer hohen Markteintrittsbarriere und hohen Margen.

1.3. Anwendungen

Den meisten der neuen Anwendungen (KI, Cloud, IoT, AR, VR, 5G, Streaming, etc.) ist gemeinsam, dass sie sehr schnell eine immense Fülle von Daten bewegen müssen, bei gleichzeitiger Reduzierung des Energieverbrauchs. Parallel zu den steigenden Anforderungen der Anwendungen schrumpfen jedoch die Möglichkeiten der Halbleiterindustrie den Anforderungen nach einer Energieersparnis gerecht zu werden, da physikalische Grenzen einer weiteren Skalierung der Bauelemente eine Ende setzen. Deutlich wird das an den drei Engpässen, die durch folgende Stichworte beschrieben werden:

- Moore’s Law
- Dennard Scaling
- von-Neumann-Flaschenhals

2. Halbleiterindustrie

Die Halbleiterindustrie ist ein Teilbereich der Elektroindustrie und die Grundlage für die gesamte Informations- und Kommunikationstechnik. Das Marktvolumen der Halbleiterindustrie wird auf über 400 Mrd. USD geschätzt. Die gesamte Informations- und Kommunikationstechnik hat ein Volumen, dass mit 4000 Mrd. USD jedoch zehnmal so hoch ist. Da die Halbleiterindustrie am Anfang der gesamten Wertschöpfungskette steht, wird damit die enorme Bedeutung der Halbleiterindustrie deutlich. Am Anfang der Wertschöpfungskette der Halbleiterindustrie steht übrigens Sand, nach Rohöl der zweitwichtigste Rohstoff der Welt, denn er dient als Grundstoff für die Fertigung von Halbleitern. Sand ist eine chemische Verbindung von Silizium und Sauerstoff (Siliziumdioxid) und über verschiedene industrielle Verfahren lässt sich daher aus dem Rohstoff (Quarz)Sand hochreines kristallines Silizium gewinnen in dem alle Atome entsprechend einer bestimmten Gitterstruktur angeordnet sind.

Das Besondere an dem Halbleiter Silizium ist, dass seine elektrischen Eigenschaften (insbesondere die Leitfähigkeit) durch die gezielte Dotierung von Fremdatomen verändert werden kann.

2.1 Bauelemente

Bei der Betrachtung der technologischen Grundlagen aller neuen Entwicklungen der Digitalisierung und ihrer verschiedenen Anwendungen müssen wir uns unvermeidlich mit den Grundbausteinen beschäftigen, die am Anfang der Wertschöpfungskette stehen: Halbleiter. Halbleiter sind die Grundbausteine von Bauelementen in fast allen Produkten der Elektronikindustrie und erhalten ihren Namen, weil sie im Gegensatz zu Metallen keine hohe Leitfähigkeit haben. Die Bedeutung der Halbleiter liegt in besonderen physikalischen Eigenschaften, so kann z.B. durch die gezielte Dotierung mit Fremdatomen im Halbleiter die Leitfähigkeit verändert werden und somit können unterschiedliche elektronische Schaltungen realisiert werden. Die besondere Bedeutung der Dotierung liegt darin, dass hierdurch elektronische Bauelemente mit einer richtungsabhängigen Leitfähigkeit (Gleichrichtereffekt) bzw. einer Schalterfunktion gebaut werden können. Unter dem Gleichrichtereffekt versteht man, dass der Strom bei Anlegen einer Spannung nur in eine Richtung fließen kann. Diesen Effekt macht man sich z.B. zunutze bei Diodenempfängern, einer einfachen Form eines Empfängers von Radiowellen (bei einem Diodenempfänger stellt die Diode einen der Grundbausteine für den Empfang dar, der ohne eine Stromversorgung auskommt).

Exkurs:

Kenngrößen der Halbleiter

Viele der späteren Ausführungen können in ihrer Bedeutung besser eingeordnet werden, wenn man einige der Grundlagen und Kenngrößen von Halbleitern kennt. Ein zentraler Begriff aller Halbleiter ist die sogenannte Bandlücke bzw. der Bandabstand. Hierbei handelt es sich um folgenden Tatbestand: Nach einem quantenmechanischen Modell zur Beschreibung von Energiezuständen (sogenanntes Bändermodell) können die Elektronen sich nur auf bestimmten Energieskalen aufhalten. Bei diesen Energieskalen wird unterschieden zwischen dem noch vollständig mit Elektronen besetzten Energieband (Valenzband) und dem nächst höheren Energieband (Leitungsband). Als Bandlücke wird der Abstand zwischen Valenzband und Leitungsband bezeichnet, der in Elektronenvolt (eV) gemessen wird. Ein eV bezeichnet damit die Energie, die benötigt wird, um von einem Band in das nächst höhere Band zu wechseln. Zur Leitfähigkeit tragen nur die

Elektronen im Leitungsband bei, so dass die Höhe der Bandlücke einen Anhaltspunkt dafür gibt, wie leicht Elektronen angeregt werden können um in das Leitungsband zu wechseln. Bei Silizium benötigt man etwa 1 eV, damit Elektronen angeregt werden in das Leitungsband zu springen. Dies geschieht bereits bei Raumtemperatur, so dass Halbleiter in begrenztem Maße eine Leitfähigkeit aufweisen.

Eng verbunden mit dem Begriff der Leitfähigkeit ist die sogenannte Elektronenbeweglichkeit. Sie ist ebenfalls eine der wichtigsten Kenngrößen jedes Halbleiters. Je höher die Elektronenbeweglichkeit, desto höhere Schaltgeschwindigkeiten lassen sich erzielen. Sie hängt u.a. ab von der Bandstruktur, der Dotierung sowie der Perfektion des Halbleiterkristalls. Besonders günstige Eigenschaften weisen die sogenannten Verbindungshalbleiter auf und je nach Zusammensetzung der Schichten verschiedener Verbindungshalbleiter lassen sich unterschiedlich gute Werte bei den entscheidenden Kennzahlen erzielen.

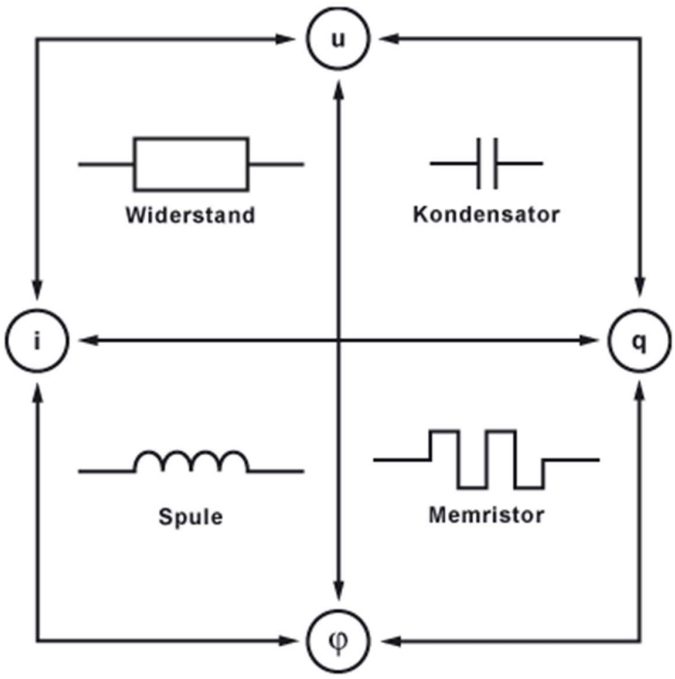
Ähnlich aufgebaut wie die Diode ist der Transistor, der in der Elektronik eine wesentliche Rolle einnimmt. Transistoren ersetzten in den 70er Jahren die bis dahin vorherrschenden Vakuumröhren, die in Radioempfängern zur Verstärkung der Eingangssignale genutzt wurden. Ein Transistor kann ähnlich wie die Röhre ebenfalls als Verstärker eingesetzt werden. Die Funktionsweise beruht darauf, dass eine kleinere Steuerspannung (aus der Antenne) einen größeren Laststrom beeinflusst und so das Ursprungssignal verstärken kann. Eine andere Eigenschaft des Transistors ist seine Funktion als Schaltelement und in dieser Funktion übernimmt er in Schaltungen bzw. Prozessoren eine der entscheidenden Aufgaben wahr. Während bis Anfang der 70er Jahre Prozessoren noch Transistoren, Widerstände und Kondensatoren als diskrete Bauteile zusammengefügt wurden, gelang es schließlich, alle Bauelemente auf Platinen in Form eines Integrierten Schaltkreises (engl. Integrated Circuit, IC) unterzubringen, wobei als Grundlage ein Siliziumkristall diente. Silizium ist auch heute noch das wichtigste Substrat.

2.2 Memristor

In der Elektronik werden drei verschiedene passive elektronische Bauteile unterschieden. Passiv bezeichnet hierbei, dass diese Bauelemente keine Verstärkerfunktion zeigen. Zu den drei fundamentalen Bauelementen eines Schaltkreises zählen Widerstände, Kondensatoren und Induktivitäten (Spule). Diese Elemente werden als fundamental bezeichnet, weil sie physisch nicht weiter unterteilt werden können ohne ihre Funktion zu verlieren.

1971 machte jedoch der amerikanische Wissenschaftler Leon Chua die Feststellung, dass es aus Symmetriegründen neben den genannten drei passiven Bauelementen noch ein viertes passives Bauelement geben müsse, welches die bisher fehlende Verbindung zwischen dem magnetischen Flux und der Ladung darstellen könne. Er kam zu dieser Überlegung, da sich die vier elektronischen fundamentalen Größen Ladung, Spannung, Strom und magnetischer Fluss wechselseitig durch mathematische Gleichungen beschreiben lassen (s. nachstehendes Schaubild) und diese Beziehungen durch Bauelemente physikalisch dargestellt werden können, mit Ausnahme der Beziehung zwischen magnetischem Flux und der Ladung. Dieses fehlende Bauelement bezeichnete Leon Chua als Memristor (abgeleitet von den Worten Memory und Resistor). Hewlett Packard hatte 2008 erstmals ein solches passives Bauelement konstruiert und Hewlett Packard Enterprise (HPE) forscht weiterhin seit mehr als zehn Jahren an der Entwicklung dieses Bauelementes. Das besondere Verhalten eines Memristors besteht darin, dass dieses Bauelement eine Gedächtnisfunktion hat. Diese Funktion äußert sich darin, dass der Widerstandswert des Memristors sich je nach der Menge

und Richtung der angelegten Spannung ändert und der neue Wert später beibehalten wird. Memristoren haben außerdem den großen Vorteil, dass sie im Gegensatz zu anderen Bauelemente sehr energiesparsam arbeiten. Aufgrund der Gedächtniseigenschaft eignen sie sich außerdem hervorragend für den Einsatz bei neuromorphem Rechnen. Die Vorteilhaftigkeit für das neuromorphe Rechnen gründet sich auch darin, dass sich im Gegensatz zu den traditionellen Bauelementen nicht nur diskrete Werte darstellen lassen, sondern auch kontinuierliche Werte. HPE zeigte in Las Vegas 2019 Memristoren, die 32 unterschiedliche Werte repräsentieren können. Offenbar ist es in der Forschung bereits gelungen, Memristoren mit mehr als 100 unterschiedlichen Widerstands-Leveln zu bauen.



Trotz der jahrelangen Forschungsanstrengungen von HPE und anderen Firmen bzw. Forschungsinstituten kann von einem Durchbruch und einer breiten kommerziellen Verbreitung des Memristors noch nicht gesprochen werden. Die amerikanische Firme Knowm (nicht börsennotiert) bietet zwar Memristoren zum Verkauf an, die Kosten sind jedoch relativ hoch. Im Unterschied zu den anderen passiven Bauelemente erfordern Memristoren einen wesentlich höheren Aufwand in der Produktion, da Materialien auf Nanoebene bearbeitet werden müssen. In Deutschland beschäftigt sich an der TU Dresden im Chua Memristor Center ein Forschungsteam mit Memristoren.

Auf Unternehmensebene scheint HPE neben Knwom die größten Anstrengungen für eine kommerzielle Produktion zu unternehmen. HPE informierte neulich, dass ein Produkt der Speichertechnologie für den kommerziellen Markt in drei Jahren erwartet werden kann. Offenbar sind jedoch noch viele Probleme bei der Stabilität und Zuverlässigkeit zu lösen.

HPE arbeitet im Zusammenhang mit Memristoren ebenfalls an einer völlig neuen Rechnerarchitektur (genannt „The Machine“). Hierbei handelt es sich um das historisch ehrgeizigste

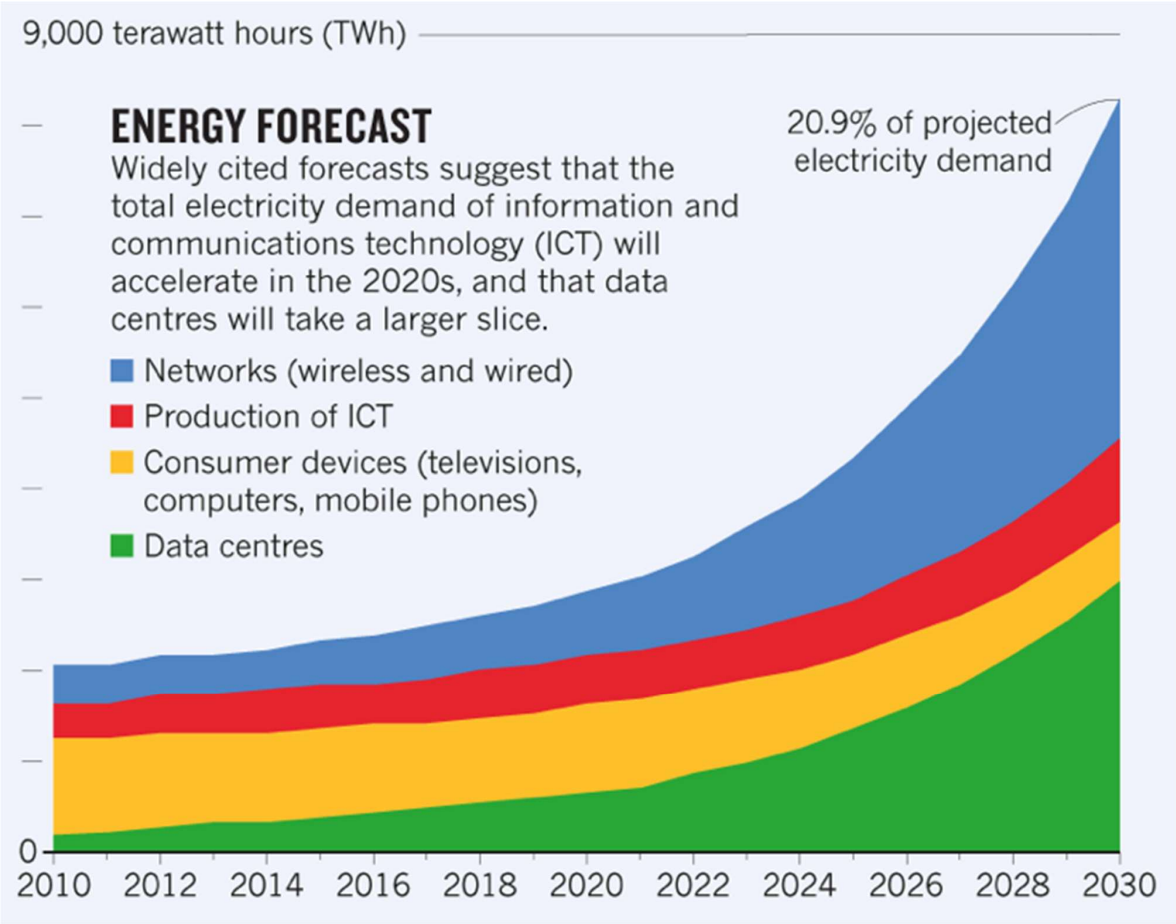
Forschungsprojekt von HPE. Der Anstoß erfolgte 2014 auf der Discover Las Vegas. Einen ersten Prototyp in Europa stellte HPE 2017 auf der CeBIT dar. Das Projekt hat u.a. zum Ziel die scharfe Trennung von Prozessor und Speicher aufzuheben, indem jedem Prozessor gleicher Zugang zu einem großen Pool eines gemeinsamen Speichers gegeben wird. Bei der Speichertechnik setzt HPE auf nichtflüchtige Speicher und strebt an, hierfür Memristoren einzusetzen, sobald die Technologie hierfür reif ist. Um die Verbindung zwischen Prozessor und Speicher schneller zu machen arbeitet HPE ebenfalls an einer Lösung mit optischen Komponenten.

2.3 Ziele

Allen Ansätzen zur Steigerung der Effizienz der technologischen Grundlagen der bisherigen Halbleitertechnik liegen Herausforderungen zugrunde, die im Zusammenhang mit den wachsenden Aufgaben der neuen Anwendungen stehen. Je nach Anwendung liegt der Schwerpunkt auf unterschiedlichen Anforderungen, aber im Wesentlichen sind die folgenden Aspekte wichtig:

- Energieeffizienz
- Geschwindigkeit
- Miniaturisierung
- Leistungsfähigkeit
- Zuverlässigkeit
- Lernfähigkeit (bei KI)
- Kosten

Die o.a. Aspekte einer Effizienzsteigerung auf verschiedenen Ebenen gelten nicht für alle Anwendungen in gleichem Masse: Während z.B. beim Autonomen Fahren eine geringe Latenz von enormer Bedeutung ist, kommt dem IoT eine stärkere Bedeutung bei der Energieersparnis zu. Fast allen Anwendungen scheint jedoch gemeinsam zu sein, dass insbesondere der steigende Energiebedarf eine Problematik ist, die es vordringlich zu lösen gilt. Der steigende Energiebedarf in den Bauelementen bereitet auf verschiedenen Ebenen Probleme (Kosten, Hitze, Zuverlässigkeit). Das Problem der Energieeffizienz betrifft nicht nur das einzelne Bauelement und die Anwendungen, sondern dieses Thema geht weit über die direkte Anwendung hinaus, da allein der Energiebedarf für den Informations- und Kommunikationssektor exponentiell schnell wächst und 2017 bereits 2 % des weltweiten CO2 Ausstoßes ausmachte, d.h. etwa ebenso hoch wie der durch den Flugverkehr ausgelöste CO2 Ausstoß. Wie die nachstehende Grafik von Anders Andrae zeigt, wächst bedingt durch Cloud Computing und Streaming Dienste insbesondere der Energiebedarf von Rechenzentren. Die EU misst daher dem Energieverbrauch in Rechenzentren besonders große Bedeutung zu und hat deshalb bereits vor Jahren einen sogenannten Code of Conduct erlassen. Der Öffentlichkeit ist allerdings die Energieproblematik der Rechenzentren kaum bewusst und die Medien konzentrieren sich auf die unmittelbar verständlichen Probleme des Fahrzeugsektors. Wir werden später sehen, dass die Reduzierung des Energiebedarfs von Rechenzentren eine der wichtigen Anwendungen der Leistungselektronik darstellt.



Von zentraler Bedeutung ist die Tatsache, dass es mehrere vollkommen unterschiedliche technologische Ansätze gibt den Energieverbrauch zu reduzieren: Im weiteren Verlauf dieses Textes werden wir sehen, dass Lösungsmöglichkeiten sich nicht nur auf Komponenten und Materialien erstrecken, sondern auch auf die Rechnerarchitektur, das Packaging und die Prozesstechnologie.

3. Technologische Lösungen

3.1 Vorbemerkung

Heutige Computer beruhen auf der sogenannten von-Neumann-Architektur, die sich dadurch auszeichnet, dass Speicher und Prozessor getrennt sind und über einen sogenannten BUS verbunden werden.

In einem Speicher ist die kleinste Einheit eine Speicherzelle. Eine DRAM-Speicherzelle besteht z.B. aus einem Kondensator und einem Transistor. Das Lesen und Schreiben in den Speicher erfolgt über den Transistor, der als Schalter arbeitet und einen Kondensator lädt. Je nach Ladezustand des Kondensators wird eine logische Eins oder Null gespeichert. Eine andere Variante einer Speicherzelle ist die sogenannte SRAM-Zelle, bei der die Datenspeicherung durch Schaltzustände

der Transistoren erreicht wird. SRAM-Zellen haben den Vorteil, dass eine Entladung des Kondensators durch Leck- bzw. Kriechströme vermieden werden kann. Nachteilig ist jedoch, dass die Darstellung auf einem Chip aufgrund der höheren Anzahl von Transistoren aufwändiger und teurer ist.

Der Prozessor (engl. Central Processing Unit - CPU) kann auf einem Quadratzentimeter heutzutage mehrere Millionen Transistoren enthalten und damit wird die Bedeutung des Transistors als zentrales Bauelement der Mikroelektronik deutlich.

Die Trennung von Speicher und Prozessor stellt jedoch zunehmend bei den vor allem durch KI-Anwendungen geforderten hohen Geschwindigkeiten einen Engpass dar, da Speicher und Prozessor unterschiedliche Datenbereitstellungszeiten aufweisen. Da bei dieser Architektur ein gemeinsamer BUS für Daten und Befehle genutzt wird muss die maximal übertragbare Datenmenge aufgeteilt werden. Während früher die CPU die langsamere Einheit des Rechners war, hat sich diese Situation in den letzten Jahren geändert und der Speicher ist die langsamere Einheit. Die CPU muss daher immer darauf warten, dass ihr vom Speicher Daten zur Verfügung gestellt werden.

Um diese Schwierigkeiten zu meistern werden verschiedene Ansätze verfolgt, auf die wir im folgenden kurz eingehen. Ausser den grundsätzlichen Problemen der von-Neumann-Architektur stösst die Halbleiterindustrie auch bei der Miniaturisierung der Bauteile an technologische Grenzen. Nach dem sogenannten Moore'schen Gesetz ergibt sich etwa alle 18 Monate eine Verdoppelung der Anzahl der Transistoren auf einem integrierten Schaltkreis. Aus rein physikalischen Gründen (Quanteneffekte) kann diese Verdoppelung jedoch nicht unbegrenzt fortgesetzt werden. Allerdings ist eine weitere Verdoppelung alle etwa 18 Monate auch nicht erstrebenswert, solange der Engpass bei der Speicher-Prozessor-Verbindung (sogenannte von-Neumann-Flaschenhals) nicht überwunden wird. Im Zusammenhang mit Moore's Gesetz gibt es auch eine Wende bei dem sogenannten Dennard Scaling. Hierunter wird verstanden, dass bei Transistoren sich nach jeder neuen Generation die Dimensionen der Transistoren verringern und trotzdem die Energiedichte unverändert bleibt. Zwar kann man die Transsistorgrösse nach wie vor verringern, es gelingt jedoch seit einigen Jahren nicht mehr nach jeder Verringerung die Energiedichte konstant zu halten, da auch hier Quanteneffekte zu einer schnelleren Erhitzung der Chips führen.

Im Zusammenhang mit den durch die neuen Anwendungen bedingten Herausforderungen an eine deutlich höhere Effizienz der Halbleitertechnologie und die o.a. beschriebenen Engpässe werden zwei unterschiedliche Wege verfolgt:

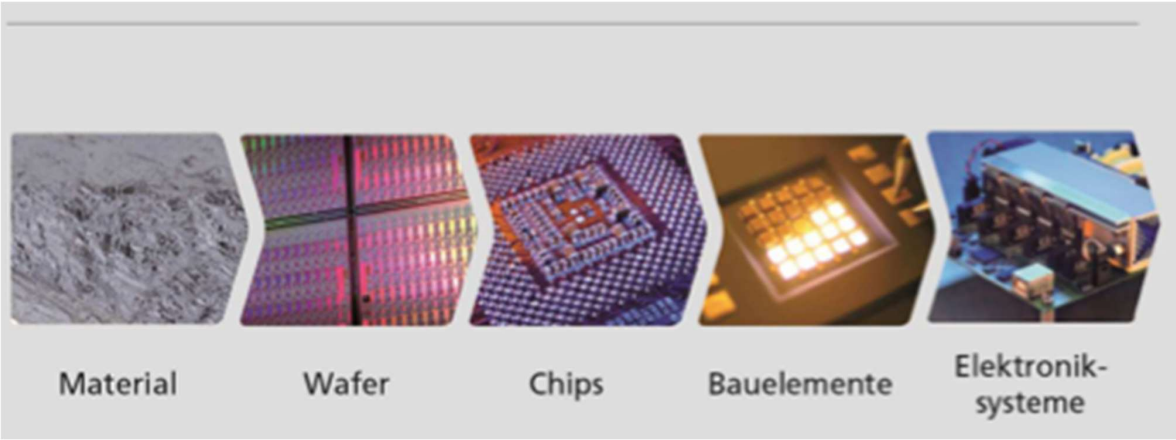
- Ansätze auf der Ebene der Verbesserung der Speichertechnik bei Beibehaltung der von-Neumann-Architektur
- Ansätze mit einer vollkommen neuen Architektur, die ohne die Trennung von Prozessor und Speicher auskommt

Für eine Einordnung und ein besseres Verständnis der technologischen Herausforderungen und Lösungsansätze in der Produktion von Halbleiter- und Mikroelektronikprodukten ist es nützlich, sich vor einer Beschreibung der Lösungsmöglichkeiten zunächst die Wertschöpfungskette anzuschauen.

3.2 Wertschöpfungskette

Am Anfang der Wertschöpfungskette steht das Ziehen eines Siliziumkristalls, aus dem dann später die Wafer gebaut werden, die wiederum als Grundlage für die Chips gelten. Das nachfolgende

Schaubild gibt nur einen groben Überblick über diese Wertschöpfungskette, da viele der interessanten Schritte nicht dargestellt sind. Auf vielen Ebenen der Massenproduktion gibt es zahlreiche Firmen, so dass wir auf diese Bereiche der Wertschöpfungskette nicht näher eingehen wollen, zumal aufgrund der grossen Konkurrenz die Margen tendenziell geringer sind. Interessant

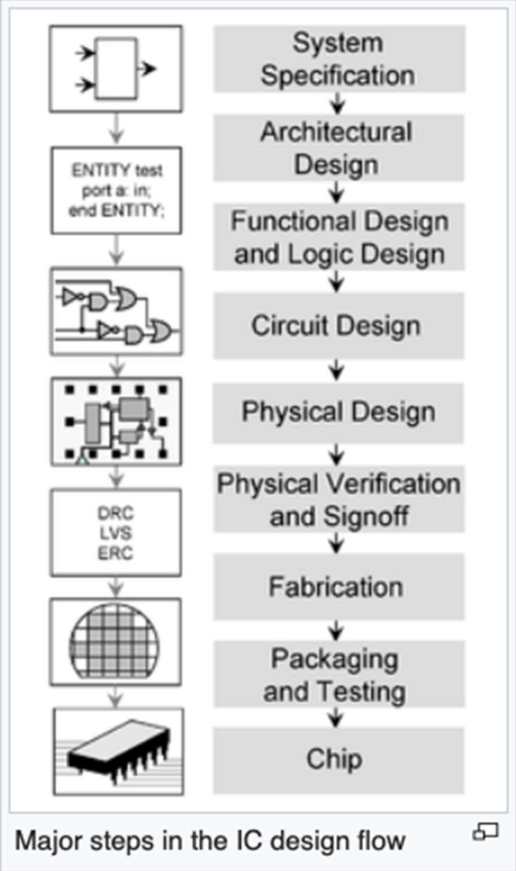


sind Unternehmen, die Maschinen herstellen zur Beschichtung der Halbleiter mit den verschiedenen Materialien (sogenanntes MOCVD-Verfahren). Weltweit einer der ersten Hersteller in diesem Markt ist die in Aachen ansässige Firma Aixtron, die ein Spin-Off der Universität Aachen ist. Vor 20 Jahren war Aixtron weltweit die einzige Firma mit Produktionsanlagen, inzwischen gibt es jedoch in den USA und China wichtige Konkurrenten.

Wir können an dieser Stelle nicht auf die gesamte Wertschöpfungskette eingehen, sondern wollen nur drei kritische, weniger bekannte Schritte der Wertschöpfungskette beleuchten, die sich als neue Schlüsselkompetenzen mit hohem Wachstumspotential etablieren. Es handelt sich um:

- Elektronik Design Automation
- (Advanced) Packaging
- Metrologie

Das folgende Schaubild zeigt, dass allein das Design in verschiedene Schritte unterteilt werden kann.



Major steps in the IC design flow

3.2.1 Electronic Design Automation (EDA)

Der Produktionsprozess von Mikroprozessoren ist zwar aufwändig und viele der grossen

Firmen in der Fertigung sind der breiten Öffentlichkeit gut bekannt, aber der Schwerpunkt der Innovation scheint sich von der Herstellung und Prozesstechnologie aufgrund der zunehmenden Komplexität immer mehr zum Design zu verschieben. Komplexe Chips werden heute mit Hilfe von rechnergestützten Werkzeuge entwickelt – damit ist das Design neben der Fertigung zu einer weiteren Schlüsselkompetenz geworden.

Die Unternehmen in diesem mit „Electronic Design Automation“ bezeichneten Sektor erfüllen daher in der Wertschöpfungskette der Mikroelektronik eine essentiell wichtige Rolle und wirken als Bindeglied zwischen den Halbleiterherstellern und Anwendern. Eine typische Fragestellung ist z.B. die Frage nach der Optimierung und Beschleunigung von Plattformen und Werkzeugen für den optimalen Einsatz von KI in der Fahrzeugindustrie. Ein anderes Beispiel ist die Prüfung der Ausnutzung aller Vorteile der 3D-Integration beim Packaging, die ebenfalls anspruchsvolle Design-Techniken erfordert (s.u.)

Führende Gesellschaften in diesem Wachstumsmarkt sind die amerikanischen Gesellschaften Synopsys, Cadence und Rambus.

3.2.2 Packaging (Advanced)

Im Zusammenhang mit den wachsenden Schwierigkeiten in der Lithographie schenken die Hersteller auch dem Thema „Packaging“ zunehmend Bedeutung. Dieses Geschäftsfeld tritt ebenfalls als neue wachsende Schlüsselkompetenz im Halbleitersektor auf. Als „Packaging“ wird der Prozess der Zusammenstellung einzelner Halbleiterkomponenten zu fertigen Halbleiterprodukten verstanden.

Die Industrie verfolgt seit einigen Jahren den Weg zu einer Fertigung, wo einzelne Chips gestapelt werden (2.5D bzw. 3D). Diese Technologie ist relativ neu und sehr aufwändig. Die Komplexität des Packaging wird aus folgendem Zitat eines Forschungsberichtes deutlich: „You may have three or four things in the package—an ASIC, HBM (high-bandwidth memory), interposer and package—but there are a lot of possible ways to put them together. There is a near infinite number of ways you can combine chips in a package“.

Ein intelligentes Packaging - angepasst an die jeweilige Anwendung - ermöglicht ebenfalls Energieeinsparungen, jedoch auf einem vollkommen anderen Weg als die traditionelleren Halbleitertechniken.

Die Unternehmen in diesem Sektor werden im englischen Sprachgebrauch als OSATs (Outsourced Semiconductor Assembly and Test) bezeichnet. Marktführer ist das in Taiwan ansässige Unternehmen ASE, gefolgt von der amerikanischen Firma Amkor, die einen Marktanteil von ca. 17 % hat. Allerdings erhalten diese Firmen Konkurrenz von den großen Foundrys TSMC, GlobalFoundries und Samsung, die diese Kompetenz eingesourced haben. Analysten glauben jedoch, dass die Unternehmen aufgrund der hohen Kosten das Packaging wieder an die OSATs outsourcen werden.

3.2.3 Metrologie

Bei der Metrologie handelt es sich um einen eigenständigen Wissenschaftsbereich, der in der Wertschöpfungskette der Halbleiter und Mikroelektronik ebenfalls zunehmend an Bedeutung gewinnt. Im Zusammenhang mit der steigenden Komplexität der neuen Technologien und neuen Anwendungen in Wissenschaft und Technik steigen auch die Anforderungen an zuverlässige Messtechniken, Methoden, Messstandards und Testverfahren. Der breiten Öffentlichkeit ist die Bedeutung dieses eigenständigen Wissenschaftsgebiets jedoch kaum bewusst, obwohl ein leistungsfähiges Messwesen einen wichtigen Bestandteil der Infrastruktur eines Industriestaates darstellt.

Aus Anlegersicht ist bedeutsam, dass dieser Wachstumsmarkt noch nicht durch die typischen Themenfonds abgebildet wird.

Einen kleinen Einblick in die Bedeutung der Metrologie erhält man, wenn man sich bewusst macht, dass selbst die Länge eines Meters keinesfalls eine Naturkonstante ist: Aufgrund einer ungenügenden Genauigkeit der Definition wurde 1983 die Definition neu angepasst, da die bisherige Konvention aus dem Jahre 1960 unzureichend war: Das Meter ist jetzt definiert als die Strecke, die Licht in einem Vakuum in einem Zeitintervall von 1/29979248 einer Sekunde zurücklegt. Wesentlich ist hierbei, dass eine Länge durch eine Geschwindigkeit definiert wird. Der Grund liegt darin, dass es sich bei der Lichtgeschwindigkeit um eine Naturkonstante handelt, die als unveränderbar gilt. Mit dieser neuen Definition wurde die Unsicherheit von 10^7 auf 10^{11} verbessert. Wem dieses Beispiel zur Verdeutlichung der Komplexität von Maßsystemen und damit der Bedeutung der Metrologie zu einfach erscheint, der mag sich verdeutlichen, dass am 20. Mai 2019 nach jahrelangen Vorarbeiten ein neues Maßsystem in Kraft getreten ist. Diese Änderung folgt auf einen Beschluss vom 16. November 2018 der 60 Mitgliedsstaaten der sogenannten Meterkonvention, die einstimmig für einen grundlegenden Wandel im Internationalen Einheitensystem (SI) stimmten. Jetzt bilden nicht mehr 7 Basiseinheiten das Fundament alles Messens, sondern 7 Naturkonstanten, d.h. alle Einheiten lassen sich jetzt als Kombination von unveränderlichen Naturkonstanten ausdrücken. Mit dieser Änderung erhofft man sich ein wesentlich präziseres Arbeiten, vor allem natürlich in Bereichen auf der Nanoebene wie sie u.a. in der Halbleiterindustrie vorkommen.

Die Messtechnik und Maßnahmen zur Qualitätssicherung nehmen auch in der Wertschöpfungskette der Halbleiter einen wachsenden Einfluss an. Die metrologisch abgesicherte Messung elektrischer und magnetischer Größen bildet dabei eine der Grundlagen für die Wertschöpfung in der Mikro- und Nanoelektronik. Dies gilt besonders für die neuen Technologien 3D NAND und EUV, die größere Anforderungen an die Chipqualität und Zuverlässigkeit stellen. Bisher kommen im Bereich der Meteorologie bei diesen Technologien noch traditionelle Geräte bzw. Verfahren wie z.B. CD-SEM (critical-dimension scanning electron microsoft), Röntgenprüfverfahren und Ellipsometer zum Einsatz. Marktexperten gehen jedoch davon aus, dass mit der Erweiterung der Schichten von mehr als 128 Schichten (möglicherweise bis zu 500 Schichten) der 3D NAND Technologie vollkommen neue Verfahren entwickelt werden müssen. Derzeit wird die sogenannte „small angle X-ray scatterometry“ (CD-SAXS) als eine Variante für 3D NAND Anwendungen diskutiert. Diese Technik wird bereits bei anderen Halbleitern (z.B. FinFETs) eingesetzt. KLA-Tencor beantragte hierfür bereits 2013 ein Patent. Ein Auszug aus der englischsprachigen Patentbeschreibung verdeutlicht am eindrucksvollsten die Anwendungen der Metrologie:

„Metrology processes are used at various steps during a semiconductor manufacturing process to detect defects on wafers to promote higher yield. Optical metrology techniques offer the potential for high throughput without the risk of sample destruction. A number of optical metrology based techniques including scatterometry and reflectometry implementations and associated analysis algorithms are commonly used to characterize critical dimensions, film thicknesses, composition and other parameters of nanoscale structures.“

Mit einem Marktanteil von mehr als 70 % ist die führende Gesellschaft in diesem Wissenschaftsfeld die amerikanische Firma KLA-Tencor, gefolgt von Hitachi High Tech, Nanometrics, Nova Measuring Instruments und Rudolph Technologies sowie einigen Start-Ups. KLA-Tencor sollte 2015 von LAM Research übernommen werden, LAM zog das Übernahmeangebot jedoch später wegen wettbewerbsrechtlicher Bedenken des amerikanischen Justizministeriums zurück.

3.3 Von-Neumann-Architektur

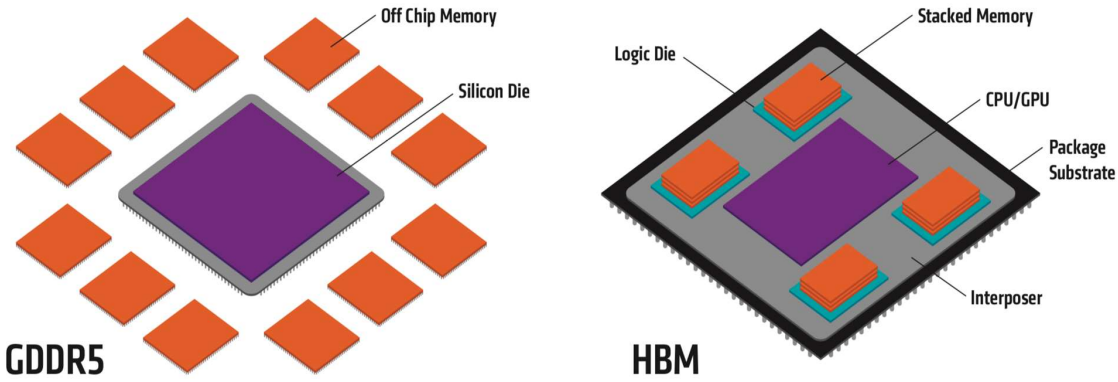
Im Zusammenhang mit den Unzulänglichkeiten der bisherigen von-Neumann-Architektur wird zunächst nach Wegen gesucht, auf Basis der bestehenden Architektur Verbesserungen vorzunehmen. Ein zentrales Forschungsgebiet für die kommerzielle Nutzung ist insbesondere die Speichertechnik. Die Anstrengungen richten sich einerseits darauf, die Übertragungsrate zu erhöhen, andererseits darauf, Speicher zu entwickeln mit sehr hohen Einschreibgeschwindigkeiten, hoher Endurance und langer Datenhaltbarkeit. Wesentlich ist, dass die neue Generation von Speichern nichtflüchtig ist, da nur auf diese Weise der Energieverbrauch reduziert werden kann.

Diese beiden Ansätze werden mit High Bandwith Memory und nichtflüchtiger Speicher (zur Erklärung s.u.) bezeichnet.

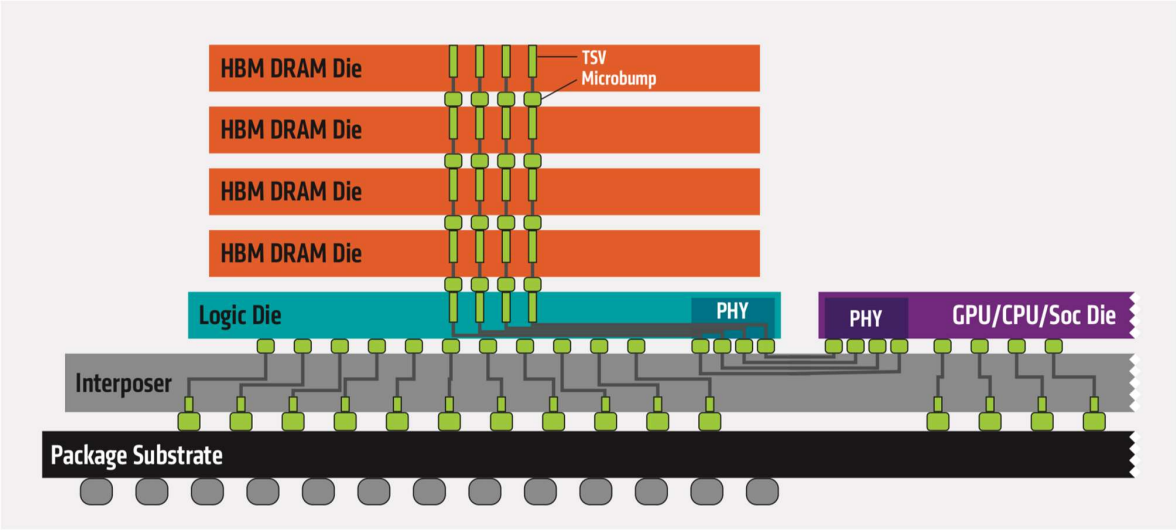
3.3.1 Speichertechnologie

3.3.1.1 High Bandwith Memory

Bei High Bandwith Memory (HBM) handelt es sich um eine Speicherschnittstelle, die den Arbeitsspeicher mit hoher Übertragungsrate an den Prozessor anbindet. Die Erhöhung der Übertragungsrate wird hierbei u.a. ermöglicht durch die vertikale Anordnung verschiedener Speicherzellen, wodurch u.a. eine engere Anbindung an den Prozessor ermöglicht wird. Die Entwicklung von HBM als neuer Industriestandard erfolgte 2008 durch AMD, zusammen mit verschiedenen Partnerunternehmen (SK Hynix, UMC, Amkor Technology). Inzwischen gibt es mehrere verschiedene Editionen von HBM (HBM3 unterscheidet sich z.B. von der derzeitigen HBM2 durch eine Verdoppelung der Stacks von 4 auf 8) und die Technik wird als wesentlich besser angesehen als der bisherige Standard GDDR6 (GDDR steht für Graphics Double Data Rate, wobei es sich um die Zusatz-Bezeichnung für die sogenannten DDR-Arbeitsspeicher handelt). In der Massenfertigung überwiegt aufgrund der Kostenvorteile aber noch der Standard GDDR6, obwohl Nvidia in seinen Grafikkarten auch die HBM-Technologie einsetzt.



HBM kommt insbesondere zum Einsatz als Grafikspeicher für leistungsstarke Grafikkarten, wobei die Vorteile in höherer Bandbreite, geringerem Stromverbrauch und kleinerem Platzbedarf bestehen. Der geringere Platzverbrauch ist unmittelbar einsichtig aus der vertikalen Stapelung der Speicherchips (s. nachstehendes Schaubild aus einer Präsentation von AMD). Ein Vorteil dieser Anordnung besteht auch in den geringeren Anforderungen an die Leiterplatinen. Für Chipdesigner

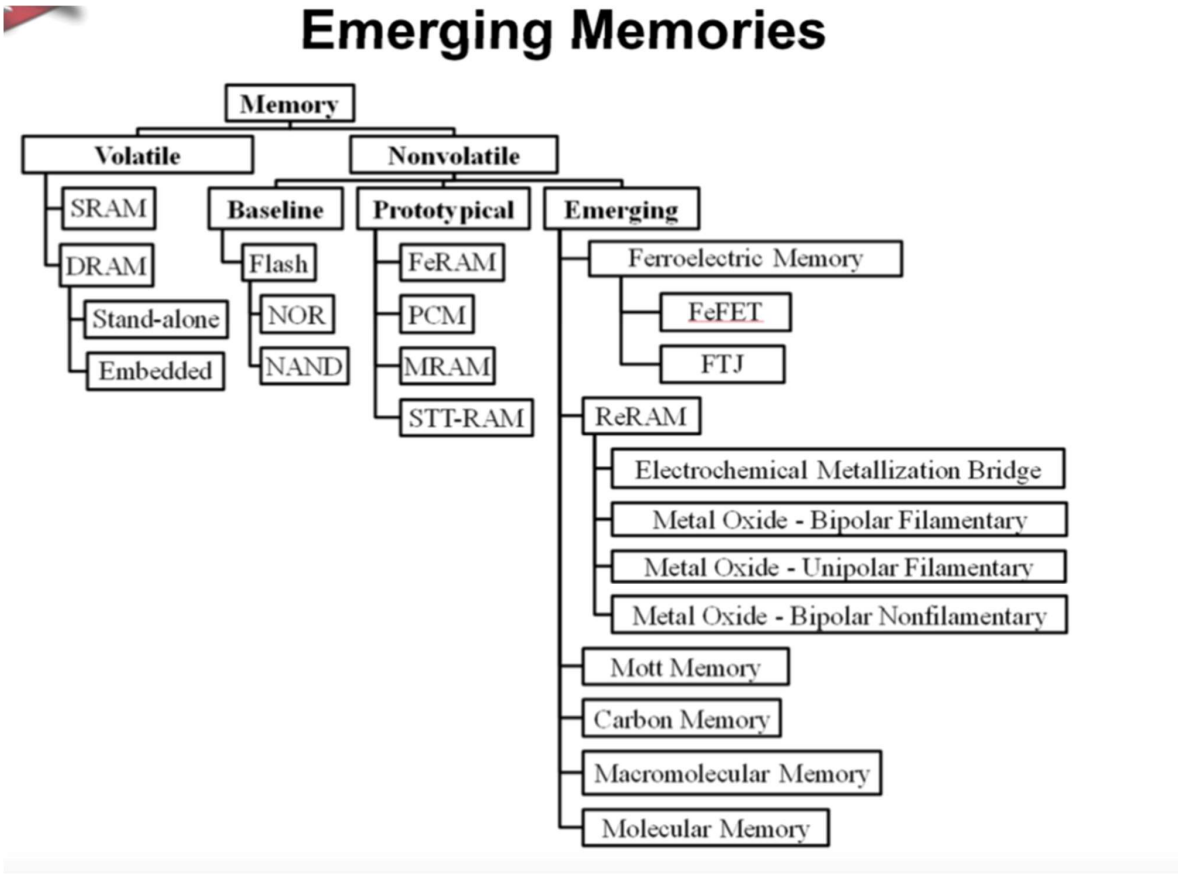


bietet HBM den Vorteil, dass man eine hohe Bandbreite erzielt bei einer geringeren Taktrate.

Beide Technologien sind komplementär, da sich der Einsatz nach der Anwendung richtet. Aufgrund der deutlich aufwändigeren HBM-Technologie und den damit höheren Kosten (Endprodukte sind etwa dreimal so teuer wie bei der GDDR5-Technologie) eignet sie sich insbesondere für KI-Anwendungen, Spiele oder Anwendungen im Bereich von Kryptowährungen.

3.3.1.2 Nichtflüchtige Speicher

Die derzeitigen Speichertechnologien der Massenproduktion lassen sich unterscheiden in flüchtige (engl. volatile) und nichtflüchtige (nonvolatile) Speicher. Unter nichtflüchtig wird ein Speicher verstanden, dessen Daten sich nach Abschalten der Energieversorgung nicht löschen. Mit dieser Eigenschaft ist automatisch eine höhere Energieeffizienz verbunden und der Trend in der Industrie geht daher natürlicherweise in die Richtung von nichtflüchtigen Speichern. Dies wird auch aus dem nachfolgenden Schaubild deutlich, welches eine Systematisierung der verschiedenen Speichertechniken zeigt.

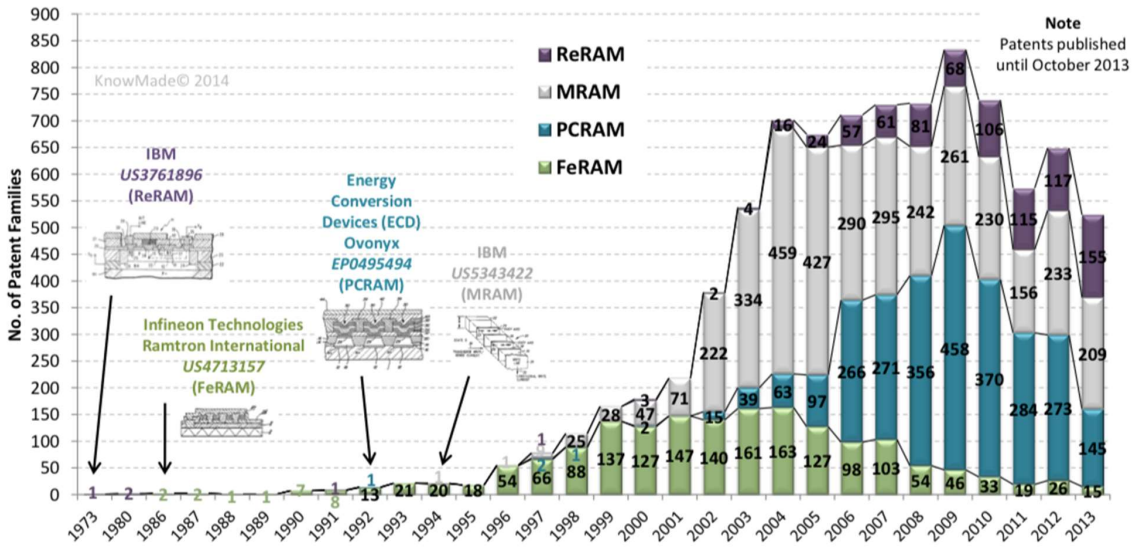


Zu den flüchtigen Speichern zählen DRAM und SRAM (s. auch oben). Derzeit überwiegt bei den nichtflüchtigen Speichern der Flash-Speicher, wobei zwischen den beiden Varianten NAND und NOR unterschieden wird. NAND ist die dominierende Technik, die vor allem bei der Speicherung von Massendaten verbreitet ist, während NOR vor allem für die Speicherung von Programmiercode eingesetzt wird. Eine weitere Unterteilung kann vorgenommen werden nach dem Einsatz der Speicher, als diskretes Element (stand-alone), oder eingebettet in ein System (System on Chip - SoC).

Aufgrund der zunehmenden unterschiedlichen Herausforderungen an die Speicher wird seit Jahren an einer Vielfalt von neuen nichtflüchtigen Speichertechniken gearbeitet. Das Ziel ist, letztendlich

die herkömmlichen Speichertechniken DRAM und SRAM zu ergänzen bzw. zu ersetzen. Diese Technologien sind unter den Namen MRAM, PRAM, RRAM und CBRAM bekannt (Details im nachfolgenden Text). Auch der zuvor beschriebene Memristor kann als nichtflüchtiger Speicher eingesetzt werden.

Das zunehmende Interesse an nichtflüchtigen Speichern in den letzten Jahren erkennt man auch an der Zunahme der Patente, insbesondere für MRAM (s. das nachfolgende Schaubild).



Für ein besseres Verständnis, warum in Forschung und Technik an einer neuen Generation von Speicherzellen gearbeitet wird, ist es hilfreich die Vor- und Nachteile der bisherigen Technologie zu vergleichen:

- DRAM

- Vorteil: Preiswert und für eine kompakte Herstellung geeignet
- Nachteil: Relativ langsam und hoher Energieverbrauch aufgrund von Leckströmen im Kondensator; Einsatz in SoC schwierig; flüchtig

- SRAM

- Vorteil: Schnell und leicht in SoC einzubetten
- Nachteil: Teuer und platzaufwändig, da mehrere Transistoren für die Speicherung benutzt werden; flüchtig

- Flash-NAND

- Vorteil: Preiswert und für die Massenproduktion geeignet
- Nachteil: Langsam und relativ hoher Energieverbrauch; Daten bleiben zwar erhalten (nichtflüchtig), aber nur für einen begrenzten Zeitraum von 5- 10 Jahren, Einsatz in SoC schwierig; begrenzte Anzahl von Schreib- und Löschvorgängen

Auch bei anderen Kennzahlen haben DRAM, SRAM und NAND Nachteile. Aufgrund der zunehmenden Anforderungen an einen geringeren Energieverbrauch der Speicherzellen sowie der technologischen Herausforderung (bei Flash-NAND) die Bauteile zu miniaturisieren, gibt es daher erhebliche Forschungs- und Entwicklungsaktivitäten einen nichtflüchtigen Speicher für die Massenproduktion zu entwickeln.

- MRAM (Magnetoresistive Random Access Memory)

Wie der Name sagt, wird bei MRAM die Information nicht mit elektrischen, sondern magnetischen Ladungselementen gespeichert. Physikalisch genutzt wird der Spin des Elektrons und in diesem Zusammenhang die Tatsache, dass ein einmal erreichter magnetischer Zustand erhalten bleibt. Wesentlich ist daher, dass der Speicher nichtflüchtig ist und daher die gespeicherten Daten auch bei einer Beendigung der Stromversorgung erhalten bleiben. Die MRAM Technologie wird aufgrund hoher Kosten bisher nur in Nischenmärkten der Industrie genutzt, wo eine hohe Haltbarkeit und Datenbeständigkeit auch bei hohen Temperaturen oder starker Strahlung (Raumfahrt) relevant sind. Aus physikalischer Sicht ist interessant, dass es sich um eine statische Technologie handelt, denn es müssen weder Atome bewegt werden, noch eine Phase verändert werden (wie bei PCM).

Traditionelle Speichertechnologien haben entweder hohe Einschreibgeschwindigkeiten oder sind nichtflüchtig, aber nicht beides. Der wesentliche Vorteil der MRAM Technologie liegt daher darin, dass sie sowohl hohe Einschreibgeschwindigkeiten haben als auch nichtflüchtig sind. Als nichtflüchtiger Speicher kann daher bei Nutzung von MRAM in Bauelementen auf Kondensatoren und Batterien oft verzichtet werden. MRAM zeichnet sich auch durch eine sehr hohe Endurance aus. Mit Endurance bezeichnet man die maximal zulässige Anzahl an Lösch- bzw. Speicherzyklen, die ein nichtflüchtiger Datenspeicher pro Speicherzelle verträgt, bis es zu spürbaren Fehlern bei Speicheroperationen kommt.

Mit der MRAM Technologie beschäftigen sich aufgrund der hohen Kosten und damit geringeren Einsatzmöglichkeiten bisher nur wenige Firmen. Neben Everspin Technologies und Avantage (nicht börsennotiert) ist noch die ebenfalls in Privatbesitz befindliche Spin Memory erwähnenswert, an der sich potente Investoren aus dem Halbleitersektor sowie ein Finanzinstitut beteiligt haben. Für den Herstellungsprozess von MRAM werden ähnlich wie bei anderen Halbleiterprodukten spezielle Maschinen benötigt. Neben LAM Research und Applied Materials stellt auch die deutsche Firma Singulus solche Maschinen her (es ist jedoch für Singulus ein Nischenprodukt).

Everspin und Avantage stellen bereits seit Jahren (Everspin seit 10 Jahren) MRAM Produkte her, so dass sie als Vorreiter angesehen werden können, die teilweise auch an die größeren Firmen ihre Lizenzen verkaufen (Everspin z.B. an GlobalFoundries). Gerade in den letzten Monaten haben sich jedoch neue Entwicklungen ergeben:

Auf der jährlich in San Francisco stattfindenden ISSCC (Solid-State Circuits Conference) hat Intel im Februar 2019 bekanntgegeben, dass die Firma bereit ist für die Serienproduktion von MRAM Chips in einem 22-nm-FinFET-Prozess (FinFET ist eine spezielle von Intel 2011 entwickelte

Transistortechnologie, die im Gegensatz zu früheren Generationen auf 3-dimensionale Transistoren setzt). Analysten vermuten, dass Intel dabei mit der amerikanischen Firma Everspin (ein spinn-off von Freescale) zusammenarbeiten könnte (dafür spricht auch, dass der Aktienkurs von Everspin nach der Ankündigung zunächst stark zulegte).

Samsung hat bereits im letzten Jahr angekündigt MRAMs als sogenannte embedded MRAMS in die Serienproduktion aufzunehmen und im Frühjahr 2019 berichtet, dass mit der Serienproduktion begonnen wird. In einem Video zur Erklärung der Technik (https://www.youtube.com/watch?v=EB14K8Gq5-w) weist Samsung darauf hin, dass eMRAM nur 1/400 der Energie der bisherigen eFlash Technik benötigt. Die Schreibgeschwindigkeit ist um den Faktor 1000 höher als bei der eFlash Technik. Experten sehen in dieser Technologie einen Ersatz für den DRAM Cache und später auch für SRAM.

Marktbeobachter weisen darauf hin, dass Samsung mit der Produktion von eMRAM in direkte Konkurrenz zu Everspin tritt. Everspin produziert zwar auch eMRAMs, aber die Hauptprodukte sind die diskreten Produkte Toggle MRAM und STT-MRAM. Der Hinweis von Samsung auf die einfache Integration in das Backende des Fertigungsprozesses und der Einsatz als „eingebettetes“ Produkt deutet darauf hin, dass Samsung (wie Intel) primär die Integration der eMRAMs – unter anderem in SoCs – plant und nicht die Fertigung von diskreten MRAM-ICs als Speicher (wie sie Everspin herstellt). Diese eMRAMS finden hauptsächlich im IoT und bei KI Anwendung.

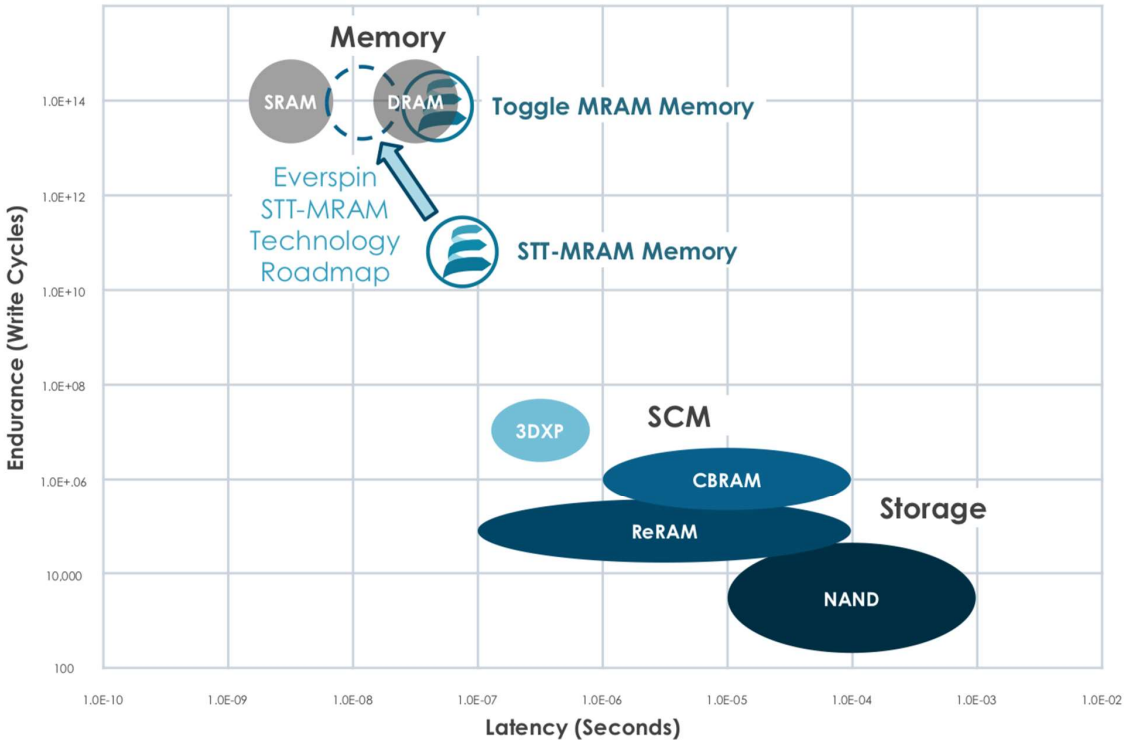
Experten erwarten, dass in wenigen Jahren die MRAM Technologie SRAM und DRAM ablösen wird. Die Idee besteht darin, dass MRAM als eine Art Universalspeicher entwickelt wird, der den herkömmlichen Arbeitsspeicher und die Festplatte zugleich ersetzt. Aus dem nachfolgenden Schaubild einer Präsentation von Everspin wird deutlich, dass die MRAM Technik im Hinblick auf einige Kenndaten der DRAM und SRAM bereits jetzt vergleichbar ist.

Bei der MRAM Speichertechnik gibt es verschiedene Varianten und derzeit ist am ausgereifteste die sogenannte STT-MRAM, wobei STT für Spin Transfer Torque steht. Die STT-MRAM Technik hat zwar gegenüber anderen Speichertechniken große Vorteile, der sogenannte „Einschreibstrom“ ist jedoch immer noch relativ hoch und aus Forschungsberichten geht hervor, dass auch die Zuverlässigkeit leiden kann. Forscher bemühen sich daher mit neuen Methoden Energieineffizienz und Zuverlässigkeit weiter zu erhöhen. Eine Variante besteht darin, den sogenannten Spin Hall Effekt zu nutzen bzw. nach Materialien zu suchen, die eine höhere Anisotropie (Richtungsabhängigkeit) haben.

Eine der physikalischen Grundlagen der MRAM Speichertechnik geht übrigens auf den sogenannten GMR-Effekt (Riesenmagnetowiderstand) zurück, der unabhängig voneinander von dem Deutschen Peter Grünberg und dem Franzosen Albert Fert entdeckt wurde. Beide Wissenschaftler erhielten im Jahre 2007 für die Entdeckung dieses Effekts den Nobelpreis für Physik.

Bei GMR handelt es sich nicht nur um einen bahnbrechenden technologischen Durchbruch, sondern diese neue Technik gab Auftrieb für eine völlig neue Elektronikform, Spintronik genannt. Das

besondere hieran ist wie bereits beschrieben, dass neben der elektrischen Ladung (wie sie in der Elektronik genutzt wird) auch der Spin des Elektrons ausgenutzt wird.



- RRAM bzw. ReRAM

Ein anderer Typ eines nichtflüchtigen Speichers ist der sogenannte RRAM (Resistive Random Access Memory) oder ReRAM. RRAM hat ggü. MRAM einige Vorteile. Eine Massenfertigung ist nur von Panasonic bekannt, die 2013 als erste Gesellschaft weltweit mit einer Produktion von ReRAM Produkten begann (ein 8-bit MCU). Auch Crossbar plant eine Produktion. Rambus und Adesto Technologies nehmen ebenfalls eine führende Rolle ein.

ReRAM war seit Jahren im Gespräch als Ablösung für Flash-Speicher, aber die Technologie ist (noch) zu kompliziert und zu teuer für die Massenproduktion. Außerdem konnten bei der NAND-Technologie Skaleneffekte noch länger als erwartet genutzt werden, so dass die Anstrengungen für die ReRAM Technologie anfänglich nachließen. Mit den neuen Anwendungen speziell bei IoT kommt jedoch neues Interesse auf, denn hierbei kommt es weniger auf schnelle Einlesegeschwindigkeiten an.

Bei ReRAM werden zwei unterschiedliche Ansätze unterschieden:

OxRAM (oxygen vacancy) und CBRAM

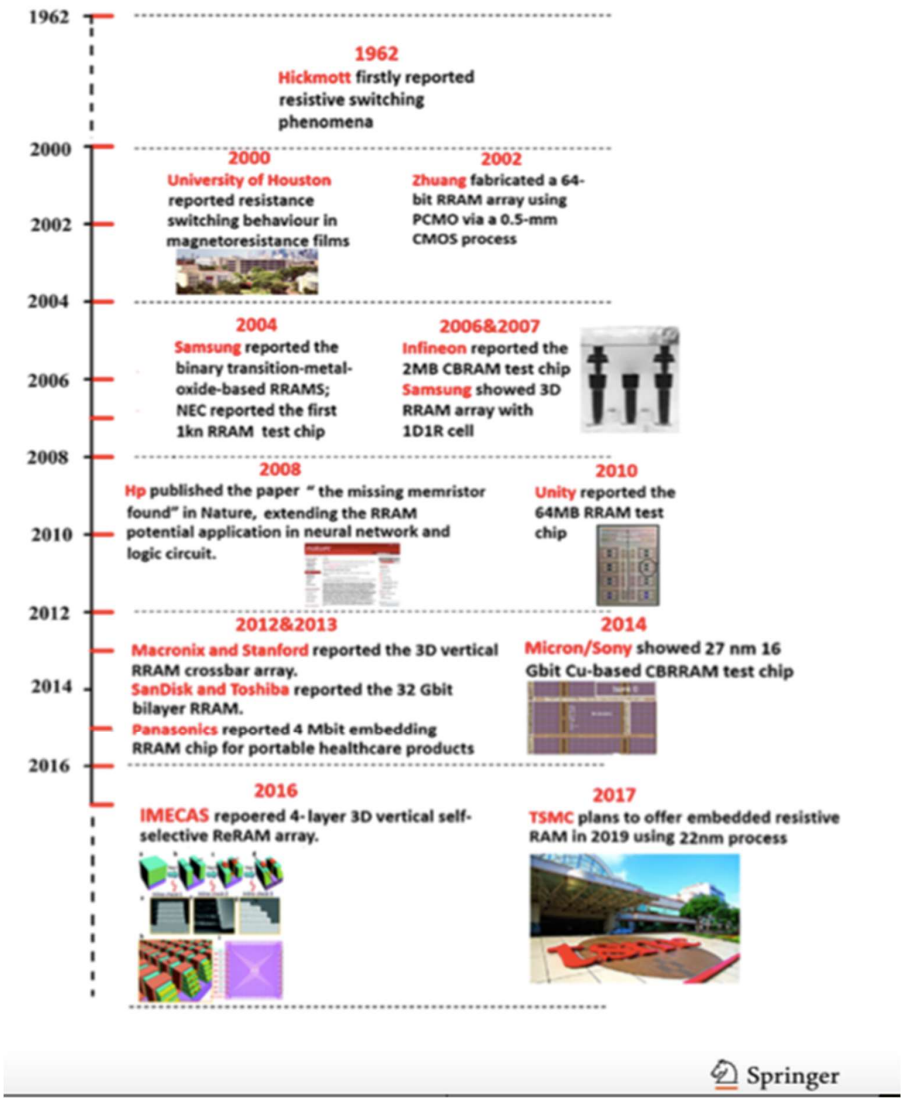
OxRAM

Diese Variante wird speziell von Panasonic unterstützt. Die kleine amerikanische Gesellschaft Crossbar weist daraufhin, dass OxRAM den Nachteil hat, dass sie schwierig zu skalieren ist.

- CBRAM (Conductive-bridging RAM)

CBRAM ist eine nicht-flüchtige Speichertechnologie, die elektrochemische Eigenschaften nutzt um Verbindungen zu bilden und aufzulösen. Diese Technologie wurde von der Arizona State University entwickelt und zeichnet sich besonders durch einen niedrigen Energieverbrauch aus. Infineon war einer der Erwerber von Lizenzen zur Herstellung der Technologie. Infineon lagerte dieses Geschäftsfeld jedoch später an ihre Tochtergesellschaft Qimonda aus, die anschließend an Adesto Technologies verkauft wurde. Adesto Technologies tritt daher heute als der marktführende Hersteller dieser Technologie auf, der aufgrund ihrer energiesparenden Eigenschaft insbesondere im IoT Bedeutung zukommt. Adesto Technologies wurde 2007 mit dem Ziel gegründet, die CBRAM Technologie zu erforschen und zu vermarkten. Adesto hat anderen Firmen Lizenzen verkauft mit dem Ziel die Technologie weiter zu fördern.

Das folgende Schaubild gibt einen Überblick über die Entwicklungsaktivitäten und „Player“ bei ReRAM. Es wird deutlich, dass seit der ersten Entdeckung inzwischen fast 60 Jahre vergangen sind.



- PRAM (Phase-change memory)

Ein weiterer Speichertyp ist die sogenannte PRAM Technik. Führende Hersteller sind Samsung, IBM und Intel zusammen mit Micron (3D XPoint). Es gibt in Expertenkreisen eine Diskussion, ob 3D XPoint eher der ReRAM zuzuordnen ist oder sogar Memristoren enthält und diese Diskussion zeigt, dass wenig über die zugrundeliegende Technologie bekannt ist. Bekannt ist jedoch, dass keine Transistoren verwendet werden und daher das Produkt platzsparend ist.

Nach einem Beschluss vom August 2018 werden Intel und Micron in der ersten Jahreshälfte 2019 gemeinsam eine zweite Generation des Produktes auflegen. Später wollen beide Firmen getrennt vorgehen und je nach Kunde spezielle Produkte auf der 3D XPoint Basis vermarkten.

Eine Analyse der Umsätze von Intel zeigt, dass das Unternehmen noch sehr PC-lastig ist, denn 53 % entfallen auf diesen Bereich. Intel unternimmt daher Anstrengungen in die neuen Zukunftsthemen vorzudringen. Die Entwicklung von Optane (dem Handelsnamen von 3D NAND bei Intel) fügt sich nahtlos in diese Strategie ein. Bisher macht Optane zusammen mit 3D NAND 6 % vom Umsatz aus. Optane findet Anwendungen insbesondere bei SSD und speziellen Chips für Rechenzentren. Nach Marktberichten verfügen die mit 3D XPoint gebauten SSD über die geringsten Einschreibzeiten, die derzeit im Markt darstellbar sind (gerade eine geringe Einschreibzeit gilt als schwierige Aufgabe in den SSD).

Die nachfolgende Tabelle entstammt aus einem Forschungsbericht aus dem Jahre 2011, gibt aber einen Eindruck über die besseren Kenndaten von PCM ggü. der Konkurrenztechnologie.

	DRAM	PCM	NAND Flash
Page size	64B	64B	4KB
Page read latency	20-50ns	~ 50ns	~ 25 µs
Page write latency	20-50ns	~ 1 µs	~ 500 µs
Write bandwidth	~GB/s per die	50-100 MB/s per die	5-40 MB/s per die
Erase latency	N/A	N/A	~ 2 ms
Endurance	∞	$10^6 - 10^8$	$10^4 - 10^5$
Read energy	0.8 J/GB	1 J/GB	1.5 J/GB [28]
Write energy	1.2 J/GB	6 J/GB	17.5 J/GB [28]
Idle power	~100 mW/GB	~1 mW/GB	1–10 mW/GB
Density	1×	2 – 4×	4×

Optane wurde anfangs vom Markt nur lauwarm aufgenommen, bekam aber von Analysten inzwischen Lob. Herausgestellt werden die für bestimmte Anwendungen äußerst geringe Latenzzeit, die von keinem anderen Produkt übertroffen wird (s.a. Schlussfolgerung zu den Markterwartungen).

- Flashspeicher (NAND und NOR)

Ebenso wie bei den elektronischen Bauelementen wurden im Zeitablauf auch die Speicherzellen einer SSD immer kleiner. Die Hersteller konnten während der letzten Jahre mit der bestehenden Technologie (sogenannte multiple pattern Technik) ständig eine höhere Speicherdichte und damit geringere Kosten erzielen. Die damit verbundene Verkleinerung stiess jedoch bei 15 nm an eine technologische Grenze, denn mit der Miniaturisierung stieg die Wahrscheinlichkeit, dass sich einzelne Zellen gegenseitig beeinflussen und so die Dateninhalte gefährden. Toshiba und Samsung begannen daher 2007 als erste Hersteller mit der Entwicklung von sogenannten 3D NAND Chips. Micron, Intel, SK Hynix folgten der Entwicklung, aber erst Samsung gelang 2013 die Massenproduktion. Inzwischen stellt auch Toshiba-Western Digital 3D NAND Chips her und arbeitet wie die anderen Hersteller an Chips mit 128 Schichten.

Analysten schätzen, dass bei Samsung Ende 2017 80 % der gesamten NAND Produktion auf 3D NAND entfiel. Wie der Name andeutet, werden bei der 3D NAND Technik Schichten vertikal aufgebaut. Aufgrund der vertikalen Stapelung bei der 3D NAND Technik müssen die Elektronen einen längeren Weg zurücklegen. Die Hersteller arbeiten daher an Lösungen um die Elektronenbeweglichkeit zu steigern. Dies kann z.B. durch eine sehr sorgfältige Dopierung mit dem Halbleiter Germanium geschehen.

Ein Nachteil der 3D NAND-Technik besteht darin, dass sie sich nicht monolithisch integrieren lässt und daher bei eingebetteten Speichern nicht eingesetzt werden kann. Durch die vertikale Anordnung konnten die Hersteller daher zwar die Skalierungsprobleme umgehen, da mehr bit pro Siziliziumfläche untergebracht werden können, für bestimmte Anwendungen reicht jedoch diese Technik nicht aus.

Die Entwicklung von 3D NAND führt mithin zu einer erheblichen Steigerung der Komplexität der Bauelemente und der Prozesstechnik. Damit kommt der Reinheit der Materialien, der Effizienz des Prozesses und der Kontrolle von Unreinheiten erhebliche Bedeutung zu.

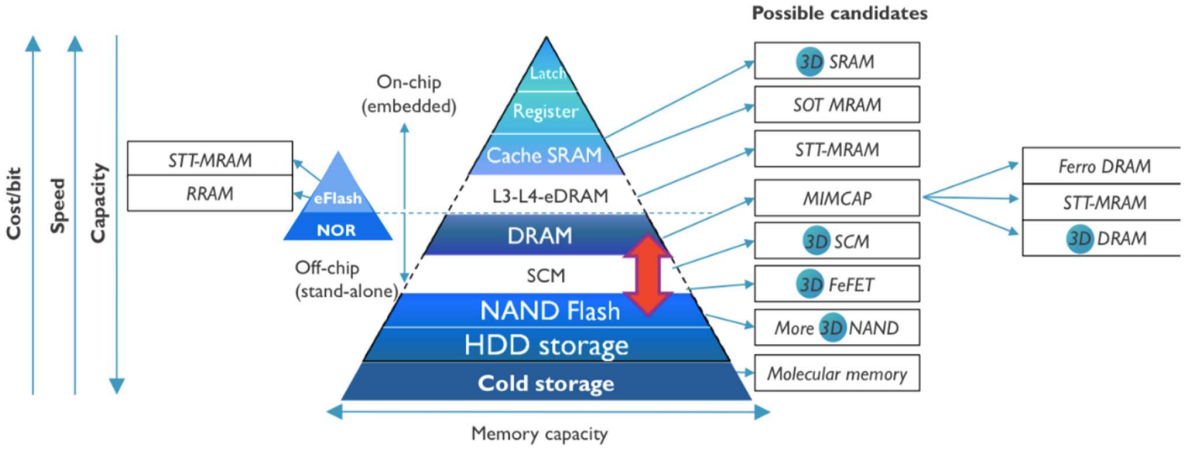
In dem Marktumfeld der Hersteller von NAND Speichern sind weniger die Hersteller von der Massenproduktion interessant, da sie neben NAND auch andere Produkte herstellen. Interessent ist jedoch die amerikanische Firma Entegris, da sie sich darauf spezialisiert hat, Methoden zu entwickeln zur Kontrolle von Unreinheiten sowie der Lieferung und Herstellung komplexer Materialien.

3.3.2 Fazit

Aus den unterschiedlichen Berichten ist zu entnehmen, dass die MRAM unter den aufkommenden neuen nichtflüchtigen Speichertechnologien das größte Potential für eine schnelle Verbreitung hat. Hierfür spricht auch, dass sich MRAM relativ leicht auf der bestehenden CMOS Architektur aufbauen lässt. MRAM zeichnet sich auch dadurch aus, dass die sogenannte Endurance um ein Vielfaches höher ist. Das gleiche gilt für die Dauer der Datenhaltbarkeit und die Schnelligkeit der Einschreibgeschwindigkeit.

Die Konkurrenztechnologie ReRAM ist der nächste Rivale, aber es ist zu beachten, dass diese Technologien unterschiedliche Märkte adressieren: Während MRAM Verbreitung findet bei

Anwendungen in denen es auf hohe Leistungsfähigkeit und Zuverlässigkeit ankommt (z.B. in Automobilen), findet ReRAM einen Einsatz bei Anwendungen bei denen es eher auf eine



kostengünstige energiesparende Lösung ankommt und bei der keine schnellen Einschreibzeiten erforderlich sind.

Insgesamt ist zu beachten, dass jeder Speichertyp spezielle Vor- und Nachteile hat, (Verlust durch Strahlung, Endurance, Kosten, Einschreibgeschwindigkeit, Lesegeschwindigkeit, etc.), so dass im Markt je nach Anwendung auch in Zukunft unterschiedliche nichtflüchtige Speichertypen verwandt werden.

Die Unterscheidung der verschiedenen alten und neuen aufkommenden Speichertechniken, die sie ersetzen werden ist nicht unmittelbar einsichtig. Deshalb diene als Zusammenfassung das nachfolgende Schaubild, welches die Zusammenhänge übersichtlich darstellt.

4. Neuromorphes Rechnen

Aufgrund der technologischen Schwierigkeiten bzw. hohen Kosten wird außerhalb der traditionellen Technologie auch nach einem vollständig anderen Ansatz zur Effizienzsteigerung gesucht: Bei allen Versuchen, die Rechenleistung des Computers zu erhöhen und den Energieverbrauch zu senken, stehen im Vordergrund oftmals Überlegungen, das menschliche Gehirn als Vorlage für die Rechnerstruktur zu wählen. Auch bei neuronalen Netzen steht die Anlehnung an die Funktionsweise des menschlichen Gehirns im Vordergrund, auch wenn eine wirkliche Abbildung bisher nicht gelungen ist.

Bei der von-Neumann-Architektur ist offensichtlich, dass diese Architektur die neurobiologischen Strukturen des Nervensystems nicht abbildet, denn im Gehirn sind Rechner und Speicher offensichtlich nicht getrennt. Forscher arbeiten daher seit Jahrzehnten daran, Chips zu entwerfen, die die Flexibilität des menschlichen Gehirns und die Fähigkeit von unstrukturierten Daten zu lernen nachahmen. Gerade der Erfolg der Halbleitertechnik brachte es mit sich, dass die Entwickler sich zunächst keine großen Gedanken um den Energieverbrauch machten. Gerade in einem Vergleich zwischen dem Energieverbrauch eines Computers und des menschlichen Gehirns werden aber die Unzulänglichkeiten der bisherigen Halbleitertechnik offenbar.

Die Unzulänglichkeiten der bisherigen Struktur treten vor allem in jüngster Zeit immer stärker zutage, da mit der schnellen Entwicklung von KI auch die Anforderungen wachsen. Neuronale Netzwerke benötigen einen immer größeren Teil des Speichers um z.B. die große Anzahl von Gewichten zu speichern, die im Rahmen des Trainings des neuronalen Netzwerkes benötigt wird.

Als Lösung bietet sich daher eine vollkommen neue Architektur an, die sich stärker an den Strukturen des menschlichen Gehirns inspiriert und die Trennung von Arbeitsspeicher und Rechner vermeidet. Es handelt sich um das sogenannte neuromorphe Rechnen.

Bei neuromorphem Rechnen handelt es sich um eine Entwicklung, die nicht nur von einem weltweitem großen Forschungsnetzwerk getrieben wird, sondern auch von einigen grossen Technologieunternehmen. Auf Forschungsebene ist in Europa führend das BrainScale System in Heidelberg, welches im Rahmen des Human

Brain Project der Europäischen Union zusammenarbeitet mit dem SpiNNaker System in Manchester. Die Idee bei neuromorphen Chips liegt darin Chips herzustellen, die logisch vergleichbar sind mit Neuronen. Der technologische Ansatz für solche Chips beruht auf sogenannten „spiking neuronal networks“, deren Idee darin besteht, dass zusammen mit der Aktivierung eines Neurons auch Signale übertragen werden. Wesentlich ist, dass neuromorphe Chips lernfähig sind und daher im Gegensatz zu neuronalen Netzwerken nicht mit einer großen Anzahl von Daten trainiert werden müssen.

Das BrainScale System in Heidelberg nutzt einen Ansatz mit 4 Millionen Neuronen und 1 Milliarde Synapsen. Dieses System bildet biologische Systeme in der Natur ab, jedoch 10.000 mal schneller. Ein Tag einer biologischen Entwicklung kann daher in 10 Sekunden entwickelt werden.

Neben den genannten universitären Forschungseinrichtungen beschäftigen sich auch einige große IT-Gesellschaften mit neuromorphem Rechnen. Hier sind in erster Linie IBM, Intel und Google zu nennen.

IBM veröffentlichte 2014 einen Bericht über die Entwicklung eines neuromorphen Chips, genannt True North. Intel berichtete im September 2017, dass die Gesellschaft den ersten neuromorphen Testchip entwickelt hat (genannt Loihi). Der Chip sollte in der ersten Hälfte 2018 Forschern zur Verfügung gestellt werden und die Verwendung befindet sich daher zunächst nur auf Laborebene. Der Chip hat 130.000 Neuronen und 130 Millionen Synapsen (zum Vergleich: das menschliche Gehirn hat 86 Milliarden Neuronen und 10.000 Synapsen pro Neuron). Intel behauptet, dass der Loihi aufgrund seiner Konstruktion im Gegensatz zu neuronalen Netzwerken wesentlich weniger Resourcen benötigt.

Trotz der enormen Resourcen, die IBM und Intel in die Entwicklung der neuromorphen Chips stecken, sind die Interessenten noch stärker an Forschungsinteressen orientiert als an kommerziellen Anwendungen. Es dürfte daher noch einige Jahre dauern, bis eine serienmäßige Produktion aufgebaut wird. Die Forschungsaktivitäten der Institute aber sind stärker auf die Modellierung von Neuronen ausgerichtet als darauf die Entwicklung von KI-Chips voranzutreiben.

Für IBM und Intel ist die Entwicklung von neuromorphen Chips noch ein Nischengeschäft, es gibt jedoch in Australien eine kleine Firma, die eine interessante Alternative zu den Chips von IBM und Intel entwickelt hat. Es handelt sich um die börsennotierte Firma Brainchip. Bei einer geringen

Börsenkapitalisierung von ca. 25 Mio. Euro eignet die Aktie sich nicht für die Aufnahme in das Anlageuniversum, oder allenfalls als spekulative Beimischung.

Aus den Berichten des Human Brain Project und den Firmennachrichten, kann man leicht den Eindruck gewinnen, dass ein kommerzieller Einsatz dieser vollkommen neuen Architektur fast unmittelbar bevorsteht. Hiervon kann jedoch noch keine Rede sein, wie man sieht, wenn man sich näher mit den Schwierigkeiten befasst, die auf Forschungsebene beschrieben werden.

Der Einsatz neuromorphem Rechnens dient insbesondere folgenden Zwecken:

- Lösung des von-Neumann-Flaschenhals
- Energieeinsparung
- Erhöhung der Geschwindigkeit
- online Learning

Die verschiedenen Systeme für neuromorphes Rechnen unterscheiden sich insbesondere bei den folgenden Aspekten:

- Modell
- Algorithmus
- Komponente
- Materialien

Bei der Unterscheidung der Modelle geht es insbesondere um die Einteilung in biologisch plausible Modelle und biologisch inspirierte Modelle (das biologisches Verhalten nachbilden, aber nicht notwendigerweise das biologische Verhalten und die Bestandteile eines Nervensystems explizit darstellen). Je nach Modell variiert z.B. der Mechanismus, wie ein Neuron Ladung aufbaut und auf welche Weise andere Neuronen aktiviert werden.

Die Wahl des Netzwerk Modells (Neuronen, Synapsen) beeinflusst auch die Wahl des Algorithmus für das Trainieren des neuronalen Netzes. Aus der Literatur ergibt sich, dass der STDP (Spike-timing Dependent Plasticity) Algorithmus den breitesten Raum einnimmt, so dass wir nur ihn kurz beschreiben möchten. Unter STDP wird ein biologischer Prozess verstanden, der die Stärke der Verbindungen zwischen den Neuronen adjustiert. Dabei wird das Gewicht der Synapsen je nach der Zeit, die zwischen dem Feuern der Neuronen (pre-synaptic Neuron und post-synaptic Neuron) vergeht, geändert. Dieser Algorithmus hat den Vorteil, dass er online auf dem Chip arbeiten kann, jedoch ist seine breite Anwendungsmöglichkeit noch nicht geprüft worden.

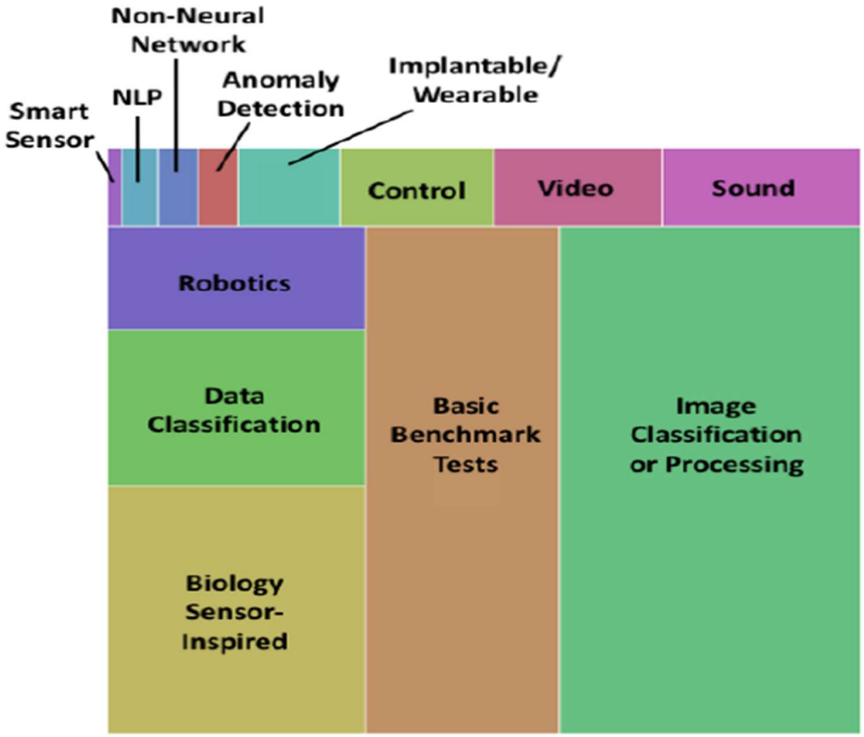
Bei den Komponenten besteht in der Forschung eine klare Präferenz für den Einsatz von Memristoren. Wie oben beschrieben ändert sich bei Memristoren der Widerstand in Abhängigkeit von der vergangenen Aktivität und sie eigenen sich daher besonders gut um das Verhalten von Synapsen abzubilden. Hiermit können z.B. auch Effekte des Vergessens dargestellt werden.

Ein wichtiger Aspekt bei der Entwicklung von neuromorphem Rechnen kommt den geeigneten Materialien zu. Je nach Typ des Materials weist z.B. ein Memristor verschiedene Werte für Endurance, Stabilität und Zuverlässigkeit auf. Zu den verschiedenen Materialien zählen neben Transition-Metall-Oxide (TMOs) Polymere, Ferroelectric Materialien, Graphen, Carbon nanotubes und organische Materialien. Der Einsatz der Materialien hängt ab vom Einsatzgebiet des neuromorphem Rechnens.

Fazit

Beim Einsatz in neuronalen Netzwerken sind bisher nach Forschungsberichten noch keine bahnbrechenden Erfolge zu verzeichnen. Einem Forschungsbericht aus dem August 2018 ist zu entnehmen, dass die meisten neuronalen Netzwerke mit weniger als 1000 Memristoren arbeiteten und daher für praktische Zwecke noch nicht eingesetzt werden können. Ein Problem bei der Integration einer größeren Anzahl von Memristoren liegt in dem als „parasitic wire resistance“ bezeichneten Effekt. Zur Überwindung dieses Effekts werden Methoden einer drei-dimensionalen Lagerung untersucht. Memristor-Speicherzellen passen aber gut in das Konzept neuronaler Netze, bei denen jeder Knoten abhängig von der Stärke der dort einlaufenden Verbindungen einen bestimmten Wert annimmt und diesen als Input an die von ihm ausgehenden Verbindungen weiterreicht. Diese Logik wurde bisher in Software abgebildet, was extrem umständlich ist. Aus diesem Grund plant HPE auch den Einsatz von Memristoren in seinem Projekt „The Machine“.

Trotz der großen Forschungsaktivitäten und eindeutigen Vorteile von neuromorphem Rechnersystemen ist daher nicht zu erkennen, dass diese Systeme anderen Systemen bereits überlegen sind, da die bisherigen Anwendungen noch mit den bisherigen Systemen ausgeführt werden können. Es fehlt bisher noch an einer klaren „killer application“, die ausschliesslich mit neuromorphem Rechnen ausgeführt werden kann. Bisher wurde neuromorphes Rechnen vor allem bei der Bilderkennung genutzt (vgl. nachstehende Schaubild). Hier leisten aber traditionelle neuronale Netze bisher (noch) gute Dienste. Allerdings erwarten Analysten mittelfristig eine zunehmende Durchdringung von neuromorphen Technologien, da die bisherigen Anwendungen der Bild- und Spracherkennung bei Smartphonen alle auf der Cloud basieren. Damit sind Zeitverluste verbunden, die vermieden werden können, wenn die Rechnung direkt auf dem Endgerät stattfindet auf der Basis energiesparender Techniken, wie sie bei den neuromorphen Technologien stattfinden.



- Neuromemristive Systeme

Von neuromorphen Computing sind zu unterscheiden neuromemristive Systeme (NMS). Während neuromorphe computing biologische Prozesse abbilden, wird in neuromemristiven Systemen versucht, die neuronale Plastizität des Gehirns zu implementieren. Hierunter wird verstanden, dass sich Synapsen und Nervenzellen nutzungsabhängig zwecks Optimierung verändern.

Der Unterschied zwischen neuromorphem Rechnen und neuromemristiven Systemen besteht darin, dass NMS die Halbleitertechnik CMOS und Memristoren nutzt, wodurch u.a. die Plastizität erreicht wird, die beim neuromorphem Rechnen nicht darstellbar ist. Außerdem abstrahieren NMS biologische Prozesse während beim neuromorphem Rechnen biologische Prozesse nachgebildet werden. Gemeinsam ist beiden Technologien, dass sie von Strukturen des Gehirns inspiriert werden.

5. Materialien

Neben den Effizienzsteigerungen der Rechnerarchitektur kommt den Materialien eine besondere Bedeutung zu. Bisher beruht die gesamte Halbleitertechnik überwiegend auf Silizium und für einige Anwendungen (Leuchtdioden und Solarzellen) auf Verbindungshalbleitern (überwiegend GaAs). Mit dem Erstarken der Leistungselektronik und Aufkommen vollkommen neuer Anwendungen gewinnen jedoch GaN und SiC stärker an Bedeutung. Noch handelt es sich bei diesen Verbindungshalbleitern um kleinere Produktionskapazitäten, Experten gehen jedoch davon aus, dass sie Silizium zunehmend verdrängen. Dieses Kapitel handelt von den Vorteilen und Einsatzmöglichkeiten von GaN- und SiC-Produkten. Die wichtigsten Player werden auch genannt.

Galliumnitrid (GaN)

Wer heute eine weiße LED einschaltet denkt kaum daran, dass dieses weiße Licht ohne GaN nicht möglich wäre. Worum handelt es sich bei GaN? GaN ist die Bezeichnung für Galliumnitrid, ein sogenannter Verbindungshalbleiter. Im Gegensatz zu dem weit verbreiteten Halbleiter Silizium werden bei einem Verbindungshalbleiter zwei Materialien zusammengefügt um auf diese Weise bessere physikalische Eigenschaften zu erzielen. Erst durch die Entwicklung der blauen Laserdiode auf Basis von GaN konnte übrigens weißes Licht in einer LED erzeugt werden.

Der entscheidende Vorteil von GaN gegenüber Silizium und GaAs (Galliumarsenid, einem anderen Verbindungshalbleiter) besteht in der sogenannten größeren Bandlücke bzw. Bandabstand (s. hierzu auch oben den Abschnitt über das Valenzband). Aufgrund der physikalischen Eigenschaften des größeren Bandabstands sind Bauelemente auf Basis von GaN wesentlich energiesparender und nehmen weniger Platz ein. In der Radiotechnik ist auch die höhere Schaltrate von Bedeutung. Zu den günstigen Eigenschaften gehören auch eine hohe Widerstandsfähigkeit gegen Wärme und Strahlung, so dass dem Material neben der kommerziellen Anwendung auch für militärische und Weltraumanwendungen Bedeutung zukommt.

Kennzahlen der Halbleiter

Eigenschaften	GaN	Si	SiC
Bandlücke (eV)	3,4	1,12	3,2
Kritisches elek. Feld (MV/cm)	3,3	0,3	3,5
Elektronenbeweglichkeit (cm2/Vs)	990-2000	1500	650

GaN ist kein neues Material und seit Jahren bekannt. Bis Mitte der 90er Jahre führte es bei Forschung und Herstellern ein Stiefmütterchendasein. Diese Vernachlässigung änderte sich schlagartig als bekannt wurde, dass es dem japanischen Wissenschaftler Nakamura gelungen war, auf der Basis von GaN eine blaue Laserdiode zu entwickeln.[1] Damit war der Weg für die Massenproduktion von weißem Licht freigegeben und Nakamura erhielt für seine Entdeckung 2014 den Nobelpreis für Physik.

Bisher erfolgte sein Einsatz überwiegend bei blauen und weißen Leuchtdioden, da die Produktion von GaN aufwändig und kostspielig ist. Inzwischen haben sich jedoch technologische Verbesserungen ergeben, so dass GaN zunehmend bei Leistungselektronik -Halbleitern eingesetzt wird. Hier geht es vor allem um Elemente im Bereich zwischen 400 V und 600 V. Die amerikanische Firma GaN Systems erklärte jedoch kürzlich, dass sie im Mai 2019 auf der Fachmesse PCIM für Leistungselektronik in Nürnberg einen Transistor auf Basis von 650 V vorstellen wird (Infineon erklärte bereits vor einiger Zeit, dass sie in diesen Bereich nicht vorstoßen wollen, da sie dieses Feld mit ihren SiC Produkten abdecken). Vor allem vor dem Hintergrund der steigenden Anforderungen im Hinblick auf Grösse und Energieeffizienz der Bauelemente im Zusammenhang mit den neuen Anwendungen kommt Bauelementen auf Basis von GaN daher eine wachsende Bedeutung zu. Ein wichtiger Einsatz erfolgt besonders bei Rechenzentren, denn gerade die energiesparenden Eigenschaften von GaN ermöglichen eine höhere Serverdichte und damit eine Energieersparnis.

Ein Mittel zur Reduzierung des Energieverlustes im Rechenzentrum besteht darin, eine ganze Phase der Leistungsumwandlung beim Trichterbetrieb vom Eingang des Rechenzentrums zum Endpunkt der Last zu eliminieren. Im Allgemeinen wird die Leistung in zwei Stufen umgewandelt - von 48 V auf der Backplane über 12 V zur Verteilung auf die Verarbeitungsplatinen bis hin zu etwa 1 V an der Stelle, an der die Leistung für die Versorgung der digitalen Chips eingesetzt wird. Mit GaN besteht jedoch jetzt die Möglichkeit, direkt von 48 V in den am Lastpunkt benötigten Bereich von 1 V zu konvertieren, ohne den Zwischenstopppunkt bei 12 V.

Bei den Anwendungen lassen sich zwei verschiedene Produktrichtungen unterscheiden:

- Hochfrequenztechnik
- Leistungselektronik

In der Hochfrequenztechnik ist das stärkste Wachstum bei Produkten für den Ausbau des 5G zu verzeichnen. Nach Schätzung der französischen Unternehmensberatung Yole Développement ist im Bereich der Hochfrequenztechnik bis 2023 ein Marktwachstum von 23 % jährlich zu erwarten.

In der Leistungselektronik können in Abhängigkeit von dem Spannungsfeld drei verschiedene Produktarten unterschieden werden:

- Niedervoltbereich (Leuchtdioden, kabelloses Laden, LiDAR, etc.) < 400 Volt
- Mittelvoltbereich (Audio) = 400 Volt)
- Hochvoltbereich (Rechenzentren, Automobil, Wechselrichter, Telekommunikationszentren) = 600 Volt

[1] In diesem Zusammenhang sei darauf hingewiesen, dass weisses Licht, welches auf der Grundlage von blauem Laserlicht hergestellt wird eine Gefahr für die Augen darstellt, da das blaue Licht sehr energiereich ist. Nakamura selbst wies auf diese Gefahr hin und erwähnte besonders Videospiele.

Für neue niedrigvoltige Anwendungen wie z.B. kabelloses Laden sind die Kosten oft noch zu hoch, jedoch ist hier ein neuer Trend zu beobachten: Bisher wird kabelloses Laden insbesondere bei Smartphones angeboten und erst seit kurzem auch bei Tablets und Notebooks. In Zukunft wird sich diese Technik möglicherweise auch auf Haushaltsgeräte ausdehnen. GaN Produkte kommen auch beim autonomen Fahren zum Einsatz, so z.B. benötigt LiDAR (eine Technik ähnlich dem Radar, nur werden anstatt von Radiowellen Laserstrahlen angewendet) schnelle GaN Feldeffekttransistoren und Dioden.

Bei den Firmen mit Bezug zu GaN kann man unterscheiden zwischen Firmen, die sich mit der Produktion von GaN-Halbleiterelementen beschäftigen (Transistoren, etc.) und solchen, die überwiegend solche Produkte vertreiben. Bei den Produzenten wiederum kann unterschieden werden zwischen denen, die sich vorwiegend auf Produkte im niedrigvoltigen Bereich konzentrieren und denen, die sich stärker auf hochvoltige Produkte beschränken bzw. einen Mix aus diesen Produkten. Die amerikanische Firma Cree scheint sich mit ihrer Tochtergesellschaft Wolfspeed auf den niedrigvoltigen Bereich zu konzentrieren (Infineon wollte diese Gesellschaft 2017 übernehmen, die Übernahme wurde jedoch von der US-Regierung nicht genehmigt). Infineon hat sich in diesem Bereich rechtzeitig 2015 durch die Übernahme des amerikanischen Pioneers International Rectifier positioniert, jedoch eher spät mit der Fokussierung auf Leistungselektronik begonnen. STMicroelectronics hingegen scheint erst relativ spät erkannt zu haben, dass der Trend bei GaN immer stärker zu Produkten der Leistungselektronik geht. Gerade im Automobilbereich sind jedoch die höchsten Margen zu erzielen, aber dort es auch im Rahmen der Produktgenehmigung eine lange Vorlaufzeit notwendig.

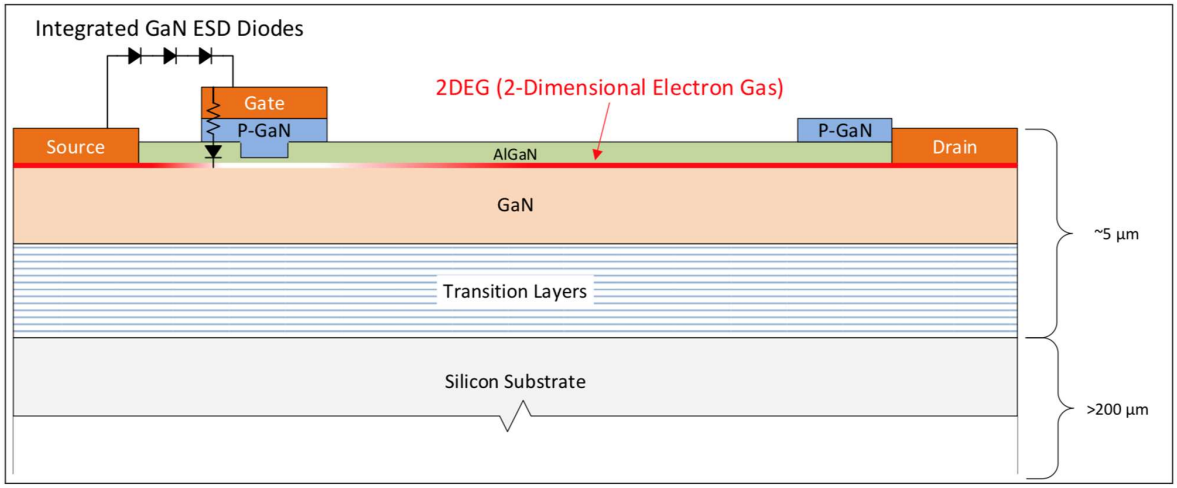
Amerikanische Unternehmen wie z.B. EPC (nicht börsennotiert), Qorvo und Macom stellen GaN Verstärker her, die sich gut zum Einsatz in mobilen Netzwerken wie dem 5G eignen. Die Unternehmen gehen davon aus, dass mit GaN die bisherigen Verstärker auf der jahrzehntealten LDMOS Technologie (laterally diffused metal oxide semiconductor) in Zukunft abgelöst werden, da mit LDMOS nur das Frequenzband bis 3,6 GHz abgedeckt wird (Grosse Teile der 5G Technologie beruhen aber gerade auf Frequenzbändern weit über 3,6 GHz).

Die französische Unternehmensberatung Knowmade, spezialisiert auf die Analyse von Patenten in der Mikroelektronik, hat mehr als 30 Unternehmen identifiziert, die für GaN-Produkte Patente angemeldet haben für die Anwendung in der Hochfrequenztechnik (überwiegend 5G). Nach einer Analyse von 3.700 Patenten kommt die Firma zu folgendem Schluss:

Bei Produkten der Hochfrequenztechnik (Verstärker, Filter, Switches, MMICs, etc). nimmt Cree nach dem Bericht bei den meisten Produkten die stärkste Stellung ein. Toshiba hat nach Cree eine starke Position bei MMICs während Sumitomo Electric nach Cree eine starke Position bei HEMTs hat. Bei Verstärkern, Switches und Filtern nimmt Cree ebenfalls eine führende Position ein, gefolgt von japanischen Firmen und in den USA Qorvo sowie MACOM. MACOM baute erst später ein Produktportfolio auf, aber zeichnete sich jetzt durch eine erhöhte Patentaktivität aus. Europäische Firmen sind weniger stark vertreten (Infineon). Bemerkenswert ist die zunehmende Aktivität chinesischer Firmen. Herausgehoben wird die starke Stellung (bezogen auf neue Patente) von Intel und MACOM bei GaN Produkten, die nur auf Silizium als Substrat beruhen (im Gegensatz zu SiC).

In der Leistungselektronik ist Marktführer die private Gesellschaft Efficient Power Conversion (EPC), aber auch Infineon wird als starker Marktführer angesehen.

GaN wird oft als Material angesehen, welches langfristig in der Elektronik Silizium ablösen wird. Den mit der Halbleitertechnik wenig vertrauten Leser mag daher überraschen, warum trotz der deutlich besseren Eigenschaften die Entwicklung und Einführung besonders der Leistungshalbleiter einen derart langen Zeitraum einnimmt (denn die ersten GaN Leistungstransistoren für 600 V wurden bereits 2012 entwickelt). Zu einem besseren Verständnis ist es hilfreich wenigstens ansatzweise zu verstehen, welche technologischen Schwierigkeiten bestehen. Hierzu betrachten wir ein Bauelement auf GaN Basis. GaN Halbleiter werden nicht im freien Raum produziert, sondern werden auf ein Substrat aufgebaut, normalerweise auf einem Siliziumwafer mit einer perfekten Kristallstruktur. Zwischen GaN und reinem Silizium besteht jedoch ein 17 % Strukturbruch bei den sogenannten Lattice Parametern. Für das epitaxische Wachsen von GaN auf das Siliziumsubstrat ist dieser Strukturbruch hinderlich. Besonders störend ist jedoch der Unterschied im thermischen Ausdehnungskoeffizienten, der bei Silicon 2,59 beträgt und bei GaN 5,59 (jeweils x10-6/K). Um die beim Wachstum bei Temperaturen von mehr als 1000 Grad auftretenden Spannungen innerhalb der Schichten zu reduzieren wird daher eine zusätzliche Isolationsschicht benutzt. Hierbei handelt es sich typischerweise um Alluminiumnitrid (AlN), jedoch scheint dieses Material nicht optimal für Anwendungen in der Leistungselektronik geeignet zu sein. Bei der Herstellung von Leistungselektronik-Halbleitern kommt es daher besonders darauf an, die Schicht zwischen GaN und Silizium optimal zu bestimmen. Es müssen nicht nur die optimalen Parameter erzielt werden, sondern es muss auch ein reibungsloser Produktionsprozess für die Serienproduktion gewährleistet werden. Das nachstehende Schaubild zeigt, dass die Schicht zwischen GaN und Silizium aus mehreren Schichten bestehen kann (von Infineon im Schaubild als „Transition Layers“ bezeichnet) und Infineon weist daraufhin, dass die Wahl der Schichten das Ergebnis einer Vielzahl von Patenten ist.



Bei der mit 2DEG bezeichneten Schicht handelt es sich um das Zweidimensionale Elektronengas, welches erzeugt wird indem zwei Halbleiter-Schichten mit unterschiedlicher Bandlücke aufeinander gebracht werden. Es handelt sich hierbei um einen Quanteneffekt, der dazu führt, dass sich die Elektronen nur in der Ebene anstatt im drei-dimensionalen Raum bewegen. Forscher hatten erstmals 1975 beschrieben und später 1994 detailliert, dass durch Einführung von 2DEG die Elektronenbeweglichkeit in einer GaN/Si Struktur erheblich gesteigert werden kann. Mit dieser Entdeckung konnte daher die Effizienz von GaN-Produkten deutlich verbessert werden und 2DEG wird heute in praktisch allen Halbleiterprodukten angewandt (z.B. den HEMT Transistoren - hohe

Elektronenbeweglichkeit Transistoren). Der erste Transistor auf dieser Basis wurde 2005 von Nitronex eingeführt (die Firma wurde 2014 von Macom Technology Solutions übernommen).

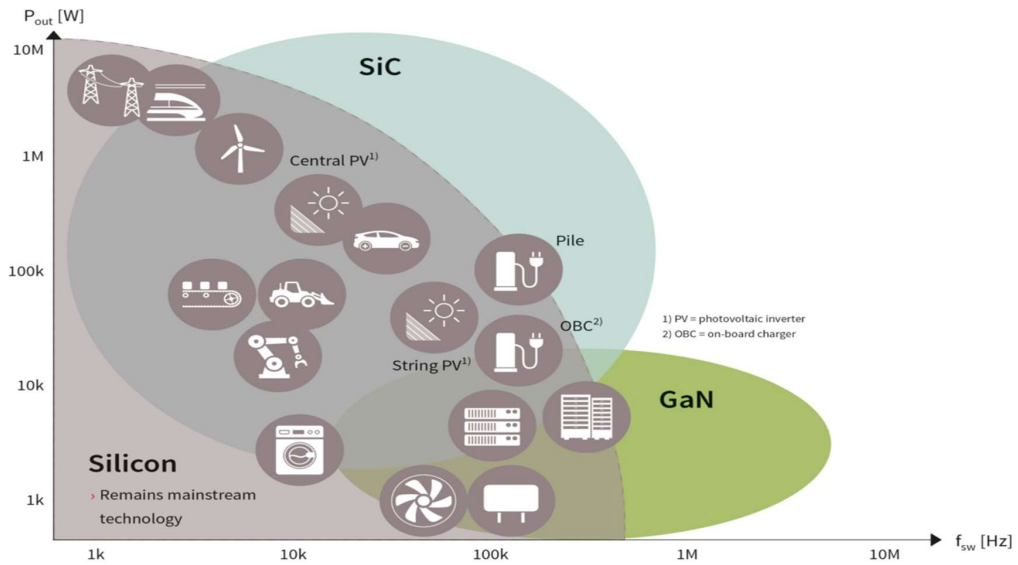
Bei Produkten der Hochfrequenztechnik scheint sich als Substrat statt Silizium SiC durchgesetzt zu haben. Qorvo und GaN Systems setzen auf diese Produktionsart. Macom (in Partnerschaft mit STMicrosystems) hingegen setzt auf Si als Substrat für die GaN-Produkte im Hochfrequenzbereich und weist auf die deutlich geringeren Kosten hin. Ein Auszug aus einer Unternehmensbrochüre von Macom macht dies deutlich: *„GaN on Si stands to benefit from silicon cost structures that are 100X lower cost than today's GaN on SiC technology, owing to the 200X to 300X slower material growth of SiC boules compared to silicon, and the attendant equipment depreciation and energy consumption absorbed by its fabs. GaN on SiC will therefore remain perpetually higher cost and thus prohibitive for mainstream use in commercial basestations."*

Siliziumcarbid (SiC)

SiC zeichnet sich dadurch aus, dass es ebenso wie GaN über einen sehr hohen Bandabstand verfügt. Es eignet sich daher besonders für Anwendungen, wo hohe Schaltgeschwindigkeiten gefragt sind, wie bei bestimmten Arten von Transistoren. Im Zusammenhang mit den Veränderungen in der Automobilindustrie (autonomes Fahren, Elektromobilität) wird erwartetet, dass die Nachfrage nach Leistungshalbleitern stark zunehmen wird, da sie sich für bestimmte Anwendungen besonders eignen. SiC steht hierbei in Konkurrenz mit GaN. Insbesondere bei hochspannigen Anwendungen (> 600 Volt) wird jedoch SiC eingesetzt, obwohl aus Presseberichten hervorgeht, dass einige Produzenten schon über GaN Anwendungen von 900 und 1200 Volt nachdenken. Bisher sind beide Produkte jedoch komplementär, da für den Bereich über 600 V (noch) Komponenten auf SiC Basis eingesetzt werden (vgl. jedoch die Ankündigung von GaN Systems).

Leistungshalbleiter auf SiC Basis eignen sich z.B. besonders für den Einsatz als Schaltregler in Elektrofahrzeugen und Bahn- und Stromversorgungsnetzen.

Das nachstehende Schaubild einer Produktspezifikation von Infineon verdeutlicht, dass SiC und GaN verschiedene, teilweise überlappende Anwendungsmöglichkeiten haben.



Energy Harvesting

Die Bedeutung der Energieeffizienz der elektronischen Bauelemente hat neben der Verwendung von energiesparenden Verbindungshalbleitern zu einem vollständig neuen Ansatz geführt. Dieser mit „Energy Harvesting“ bezeichnete relativ neue Trend zielt darauf ab, elektronische Bauelemente zu entwickeln, die im Idealfall keine eigene Energieversorgung benötigen, sondern denen Energie über eine Änderung der Umgebungstemperatur, durch Druck, eine Änderung des Lichtes bzw. durch Bewegung zugeführt wird. Diese Technologie kommt insbesondere bei Sensoren im Rahmen des IoT zum Einsatz, da dort eine möglichst batterieunabhängige wartungsfreie Stromversorgung oftmals von essentieller Bedeutung ist. Unternehmen mit Bezug zu Lösungen bei dem Thema „Energy Harvesting“ treten auf verschiedenen Ebenen in Erscheinung, da bei den Bauelementen unterschiedliche Herausforderungen auftreten (Energiespeicherung, MCU, Energiemanagement, Energiekonvertierung). In einem Forschungsbericht werden mehr als 30 Firmen genannt, die sich in unterschiedlicher Form mit Methoden der Energieeinsparung auf Chip-Ebene beschäftigen. Bei den grossen Gesellschaften sind GeneSiC (nicht börsennotiert), ON Semiconductor, Renesas, Infineon, Rohm und STMicroelectronics zu nennen.

Prozesstechnologie

Silicon-on-Insulator (SoI)

Im Gegensatz zu den Methoden der Verbesserung der Energieeffizienz bei den Materialien der Bauelemente, des Packaging und der Speichertechnologie wird beim SoI versucht mittels einer besonderen Prozesstechnik in Verbindung mit dem Siliziumsubstrat die Energieeffizienz zu erhöhen. Der Begriff bezeichnet einen speziellen Isolierschicht-Transistor bei dem eine dünne Siliziumschicht durch eine isolierende Schicht vom Silizium-Substrat getrennt ist.

SoI-Transistoren zeichnen sich daher dadurch aus, dass aufgrund der Schichtenzusammensetzung im Transistor die sogenannten Leckströme sinken und damit integrierte Transistoren mit geringeren Arbeitsströmen betrieben werden können (und gleichzeitig auch die Schaltzeiten verkürzt werden).

Die SoI-Technologie soll zunehmend die bislang vorherrschende CMOS-Technologie ersetzen. IBM forscht an dieser Technologie schon seit den frühen 70er Jahren. Auch andere Unternehmen beschäftigen sich mit der Entwicklung der SoI-Technologie, aber nach Angaben von IBM war die Firma das erste Unternehmen, welches in der Lage war, SoI beim Bau eines kompletten Mikroprozessors einzusetzen. Die langsame Entwicklung der SoI-Technologie erklärt sich aus den technischen Schwierigkeiten sowie mit der schnellen Weiterentwicklung der CMOS-Technologie. Nachdem jedoch die Fortschritte der CMOS-Technologie zunehmend geringer ausfielen, nahm das Interesse an SoI stark zu. Die technologischen Schwierigkeiten bei der Herstellung von SoI beruhen übrigens darauf, dass der Siliziumfilm, der für die Herstellung der Transistoren benutzt wird aus perfektem kristallinem Silizium besteht, während die Oxid-Isolationsschicht eine andere Kristallstruktur als reines Silizium hat. Bei einer schlechten Anpassung der Kristallstrukturen ergeben sich jedoch Defekte, die zu höheren Leckströmen führen und die Geschwindigkeit reduzieren.

Neben IBM, AMD und Samsung sind bei den wichtigen Herstellern noch Globalfoundries und Freescale zu nennen. STMicroeletronics scheint auf diesem Gebiet ebenfalls eine führende Rolle einzunehmen. Diese Gesellschaften verfügen über verschiedene Produktlinien von denen die SoI-Technologie nur ein Produkt von vielen darstellt. Als „pure play“ ist jedoch die französische

Gesellschaft Soitec interessant, die sich auf die Herstellung von speziellen SoI-Wafern spezialisiert hat (die Namensgebung SoI deutet bereits auf die Technologie hin).

- Silicon-On-Thin-Buried-Oxid (SOTB)

Die SOTB-Prozesstechnologie ist eine Abwandlung von SoI und gilt ebenfalls als eine Technik zur Herstellung von energieeffizienten Bauelementen. Kürzlich wurde von der japanischen Firma Renesas Electronics (Renesas) ein Ansatz vorgestellt, der eine derart hohe Energieeinsparung ermöglicht, dass mit SOTB gebaute IoT-Anwendungen allein auf Basis der Umgebungsenergie in der Lage sind Embedded-KI-Berechnungen durchzuführen. Renesas weist darauf hin, dass bisher aufgrund der Prozessgeometrie eine Verkleinerung der Bauteile zwangsläufig zu höheren Leckströmen und damit höherem Stromverbrauch führte. Renesas beabsichtigt im Oktober 2019 mit der Serienproduktion eines energiesparenden Chips zu beginnen, dessen Leck- und Schaltströme um 90 Prozent gegenüber herkömmlichen Methoden verringert werden. Der Chip soll ebenfalls die Versorgung aus der Umgebungsenergie über Solarenergie, Vibration oder Piezoelektrizität ermöglichen, so dass zusammen mit der deutlich erhöhten Energieeffizienz eine batteriefreie Funktionsweise von IoT-Produkten möglich sein soll. Wenn der Chip von Renesas die Erwartungen erfüllt, so würde dies einen entscheidenden Fortschritt bedeuten beim Einsatz von KI auf den Endgeräten (sogenannte embedded KI) um auf diese Weise die Verbindung mit der Cloud zu vermeiden. Renesas arbeitet übrigens bei der Entwicklung der SOTB-Technologie zusammen mit der französischen Firma Soitec.

6. Schlussfolgerung

Ziel dieses Berichtes ist, dem Leser deutlich zu machen, dass die neuen Anwendungen eine deutliche Effizienzsteigerung von Komponenten und Systemen der Schlüsseltechnologie erforderlich machen. Besonders deutlich wird dies bei dem starken Anstieg des Energieverbrauchs der Rechenzentren. Diese Effizienzsteigerung ist jedoch nur möglich, wenn sich in vielen Bereichen der Schlüsselindustrie grundlegende Verbesserungen ergeben, denn die bisherige Praxis einer Verkleinerung der Bauelemente stösst an physikalische Grenzen. Mit aller Berechtigung wird daher vom Ende des Moore'schen Gesetzes gesprochen.

Eine vollständig neue Architektur, die die bisherige Trennung von Speicher und Prozessor in der Mikroelektronik auflöst wird nicht so schnell in der Massenproduktion verfügbar sein, obwohl im Labor bereits gute Erfolge erzielt wurden. Möglicherweise gelingt HPE mit seinem Projekt „The Machine" hier ein deutlicher Fortschritt und auch ein kommerzieller Erfolg.

Abseits von den Versuchen einer neuen Architektur sind die größten kommerziellen Erfolge zu erwarten bei der Durchdringung der nichtflüchtigen Speichertechnik und des zunehmenden Einsatzes von GaN und SiC als Ersatz für Silizium. Die Entwicklung von Bauelementen auf Basis von GaN und SiC mit hoher Zuverlässigkeit ist ein sehr komplizierter Prozess, so dass die Markteintrittsbarriere hoch ist. Andererseits ist das Marktpotential so hoch, dass für alle der bisherigen „Player" ein hohes Umsatzpotential zu erwarten ist.

Als Treiber für das Marktwachstum sowohl von GaN-Produkten als auch neuartiger Speichertechnologien gelten derzeit vor allem Cloud Rechenzentren. Viele andere Anwendungen wie Elektromobilität und Autonomes Fahren werden erst später für Umsatzwachstum sorgen.

Aus kommerzieller Sicht werden im Bereich der Speichertechnologie 3D XPoint von Intel gute Marktchancen eingeräumt und Marktexperten erwarten, dass bei dieser nichtflüchtigen Speichertechnik ggü. allen anderen nichtflüchtigen Speichern das größte Umsatzpotential besteht. MKW Ventures, eine Unternehmensberatung, erwartet bereits 2020 Umsätze von USD 1,5 Mrd., also eine Verdoppelung ggü 2018. Für die gute Marktentwicklung spricht der geringere Preis ggü. DRAM und die steigende Nachfrage für SSDs bei Rechenzentren (zum Vergleich: das gesamte Volumen von DRAM wird auf USD 100 Mrd geschätzt).

Das Volumen von MRAM ist mit weniger als USD 100 Mio ggü. DRAM und NAND (ca. USD 60 Mrd.) noch verschwindend gering, dürfte aber nach Ansicht von Marktexperten ebenfalls exponentiell schnell wachsen. Dafür sprechen die überlegenen technischen Daten und die leichte Kombination mit der CMOS Technologie für den Einsatz von eingebetteten Produkten, die besonders beim IoT und bei KI gefragt sind. Für einen Markterfolg spricht auch das Bekenntnis von inzwischen vier Foundries für diese Technologie (u.a. Globalfoundries zusammen mit Everspin), da durch ihren Einsatz auch das Vertrauen der Anwender in die neue Technologie wächst. Gleichzeitig ergeben sich mit den stärkeren Umsätzen Kostenvorteile und Experten erwarten ein signifikantes Marktwachstum, wenn sich die Kosten für MRAM den Kosten für SRAM angleichen. MRAM würde dann eine Lücke zwischen dem schnellen DRAM und Flash SSD füllen.

In der NAND Technologie wird sich der Prozess zu einer vertikalen Stapelung (3D NAND) fortsetzen, aber aus Anlegersicht profitiert hiervon möglicherweise am stärksten Entegris, da sie die Schlüsselkompetenz für die Prüfung und die Lieferung von ultra-reinen Materialien mitbringen.

Auch neue Methoden beim Packaging ermöglichen eine Effizienzsteigerung und in diesem Zusammenhang kommt Unternehmen, die sich mit dem Design, dem Prüfen und Testen von Chips, Materialien und Systemen beschäftigen gerade im Zusammenhang mit dem Aufkommen der neuen Technologien eine wachsende Bedeutung zu. Von zunehmender Bedeutung könnten sich ebenfalls neue Prozesstechnologien erweisen, die die Herstellung von extrem energiesparenden Chips ermöglichen.

Im Zusammenhang mit KI, Videos und Spracherkennung gibt es einen generellen Trend, die Rechenleistung und Datenspeicherung immer stärker am Endgerät durchzuführen (sogenanntes Edge-Computing). Gleichzeitig dienen die neuen Methoden auch dazu, die Datenbewegung innerhalb des Chips durch eine Annäherung von Speicher und Prozessor zu reduzieren. Idealerweise erfolgt diese Verknüpfung von Prozessor und Speicher in Memristoren als einem vollständig neuen Bauelement in der Mikroelektronik. Technologische Schwierigkeiten und hohe Kosten verhindern bisher eine Massenproduktion und es bleibt abzuwarten, ob es HPE im Rahmen des Projektes „The Machine“ gelingt, Memristoren in großer Stückzahl zu vertretbaren Kosten zu produzieren.

Wesentlich ist, dass mit dem Auftreten von Spintronics als neuer eigenständiger Technologie vollkommen neue Produkte entstehen. Das stärkere Aufkommen von GaN ermöglichst u.a. neue Produkte wie z.B. das kalbellose Laden. Von der zunehmenden Komplexität der Technologien profitiert der eigenständige Wissenschaftsbereich Metrologie sowie die Unternehmen der Electronic Design Automation.

Die vollständig neuen aufkommenden Technologien auf der Basis von neuromorphem Rechnen lassen zwar eine erhebliche Leistungssteigerung erwarten, bis zu einer kommerziellen Nutzung mit einer Serienproduktion dürften aber noch Jahre vergehen. Auf kommerzieller Ebene liefern sich

Intel mit Truenorth und IBM mit Loihi jedoch schon jetzt einen Wettkampf und in einer mittelfristigen Zukunft kann sich hier eine interessante neue Chipgeneration ergeben.

Alle bisherigen Technologien beruhen auf der sogenannten CMOS-Technik, die technologisch bei einer weiteren Skalierung jedoch z.B. aufgrund von Quanteneffekten an physikalische Grenzen stößt. Intel entwickelt daher seit acht Jahren eine vollkommen neue Technologie. Es handelt sich um ein skalierbares logisches Spintronic Bauelement, dessen Technik als MESO (magnetoelectric spin orbit) bezeichnet wird. Diese Technik macht sich ebenso wie MRAM den Spin des Elektrons zunutze und nutzt ausserdem das magnetische Moment um die Rechenleistung auszuführen. Marktexperten halten diese Technik für sehr erfolgversprechend, erwarten eine kommerzielle Nutzung aber erst in einigen Jahren.

Exkurs Künstliche Intelligenz

Das Thema „Künstliche Intelligenz“ ist keineswegs ein neues Thema, aber erst seit wenigen Jahren beobachten wir eine starke Beschleunigung der Durchdringung von KI-Methoden. Worauf ist diese Beschleunigung zurückzuführen und warum kam es erst in den letzten Jahren zu einem Durchbruch bei vielen Anwendungen (z.B. Spracherkennung)?

Zur Erklärung sei zunächst eine Abgrenzung der grundlegenden Begriffe vorgenommen. Künstliche Intelligenz ist ein Oberbegriff aller Methoden, die eine Automatisierung intelligentes Verhaltens ermöglichen. Schränkt man den Begriff KI auf Machine Learning ein, so spricht man von Algorithmen, die die Optimierung bestimmter Rechnungen ermöglichen. Hierzu zählen z.B. die bekannten Algorithmen der linearen Regression, logistischen Regression, Support Vector Machines, Naive Bayes, Cluster-Methoden (n-nearest neighbor und k-means clustering) und die Hauptkomponentenanalyse (Principal Components Analyse). Bereits mit diesen Methoden konnten hochkomplexe nicht-lineare Muster erkannt werden. Diese Methoden sind jedoch seit Jahrzehnten bekannt und stellen keine neuen Erkenntnisse dar.

Neben den Algorithmen des Machine Learning wird seit geraumer Zeit an der Entwicklung von neuronalen Netzwerken geforscht. Auf diesem Gebiet gab es in den letzten Jahrzehnten verschiedene bahnsprechende Neuentwicklungen, wobei entscheidend die Hinwendung zu einem wahrscheinlichkeitstheoretischen Ansatz war. Im Vordergrund dieses Ansatzes stand die Entwicklung sogenannter Bayesscher Netzwerke, da sie eine deutliche Verringerung der möglichen Verbindungen zwischen den Variablen in einem Netzwerk ermöglichte. Obwohl neuronale Netzwerke (inzwischen als deep learning bezeichnet) daher von den besseren methodischen Ansätzen profitierten war die Rechengeschwindigkeit für die Anwendung von Netzwerken mit einer hohen Anzahl von Neuronen und Schichten bis 2012 noch unzureichend. Ein Durchbruch erfolgte jedoch 2012, als bei einem internationalen Wettbewerb Wissenschaftler der Universität von Toronto mithilfe von Chips der GPU (Graphik Processing Unit) wesentlich bessere Ergebnisse erzielten (s. auch das nachstehende Kreislaufdiagramm zur historischen Entwicklung von KI). Deep Learning hat sich heute insbesondere bei der Sprach- und Bilderkennung als eine sehr zuverlässige Technik etabliert, wie jeder Anwender eines Smartphones tagtäglich erfahren kann. Das Autonome Fahren nutzt ebenfalls Algorithmen des Deep Learning.

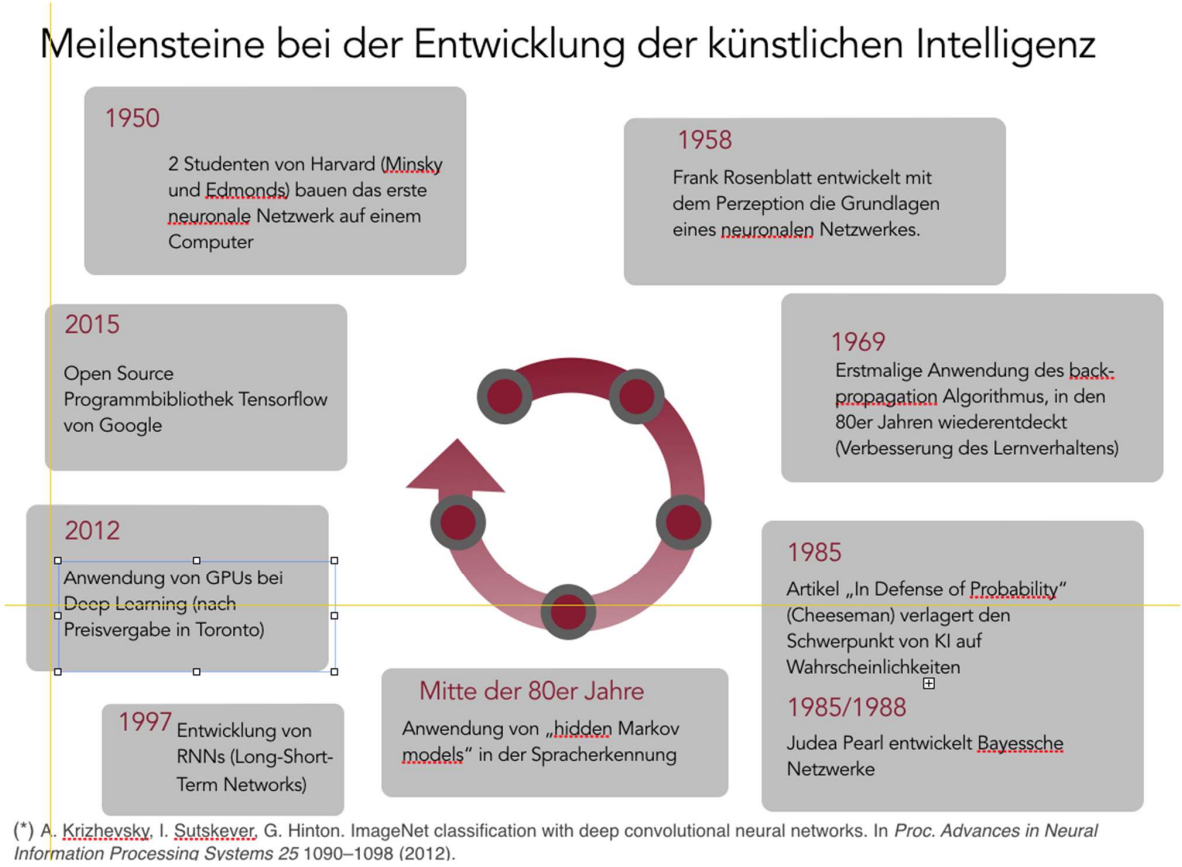
Eine Besonderheit des Deep Learning besteht darin, dass das Netzwerk selbstständig „lernt“ und die Parameter automatisch so angepasst werden, dass zwischen Eingangs- und Ausgangsdaten die höchstmögliche Übereinstimmung hergestellt wird. Vereinfacht ausgedrückt handelt es sich bei

diesem Lernprozess lediglich um die optimale Anpassung der Gewichte (z.B. der Synapsen) des Netzwerkes.

Neben dem relativ neuen Begriff des Deep Learning wird auch zwischen dem sogenannten Supervised learning und Unsupervised learning unterschieden. Beide Methoden dienen dem Erkennen von Mustern. Supervised learning unterscheidet sich dabei von Unsupervised learning dadurch, dass beim Supervised learning die zu erkennenden Muster vorgegeben werden, während beim Unsupervised learning das System selbst versucht relevante Muster zu erkennen. Beispiel: Wenn ein soziales Netzwerk verschiedene Kundengruppen bilden möchte, so werden beim Supervised learning die Kundengruppen vorgegeben und das System muss die Kunden den entsprechenden Kundengruppen zuordnen. Beim Unsupervised learning werden die Kundengruppen nicht explizit vorgegeben, sondern das System erkennt selbstständig, welche Kundengruppen sich sinnvoll abgrenzen lassen.

Allen bisherigen Ansätzen der KI ist gemeinsam, dass sie eine ungeheure Fülle von Daten benötigen um zu guten Ergebnissen zu kommen. Es gibt jedoch neue Ansätze, die einen ganz anderen Weg gehen und stärker an die Funktionsweise des Gehirns angelehnt sind. Hierzu zählt das neuromorphe Rechnen, aber auch Ansätze, die stärker an das menschliche Lernen angelehnt sind, wie z.B. Baysian Programming Learning. Der neueste Ansatz, der ebenfalls ohne eine hohe Anzahl von Daten auskommt ist das sogenannte Free Energy Prinzip, welches jedoch auch auf der Bayesschen Formel aufbaut.

Das folgende Schaubild verdeutlicht einige der wichtigen Stufen der Entwicklung von KI.



(*) A. Krizhevsky, I. Sutskever, G. Hinton. ImageNet classification with deep convolutional neural networks. In *Proc. Advances in Neural Information Processing Systems 25* 1090–1098 (2012).