Halbleiterspeicher

Thomas Schumann

M.Sc.-Studiengang Elektrotechnik

Hochschule Darmstadt

Fachbereich Elektrotechnik und Informationstechnik

Einleitung

Diese Kurseinheit stellt Aufbau, Funktionsweise und Einsatzgebiete der wichtigsten Halbleiterspeicher dar.

In elektronischen Systemen müssen Informationen gespeichert werden. Je nach Komplexität der Aufgabe reicht der Speicherbedarf von nur einigen Tausend Bits (Kbits) bis zu vielen Milliarden Bits (Gbits). Beurteilungskriterien für Speicherbausteine sind Speicherkapazität, Kosten, schneller Zugriff auf Daten sowie Verlustleistung und Zuverlässigkeit.

Ein Halbleiterspeicher ist ein Datenspeicher, der als integrierte Schaltung realisiert ist. Die Daten werden dabei in kodiertem Zustand, meist binär kodiert, gespeichert. Eine Speicherzelle ist dabei die physikalische Realisierung der kleinsten Einheit eines Speichers. Die kleinstmögliche Einheit eines binär kodierten Speichers ist die 1-bit Speicherzelle.

Die Halbleiterspeicher basieren heutzutage hauptsächlich auf der CMOS-Halbleitertechnologie. Zur Realisierung von Speicherzellen werden MOS-Transistoren benötigt. Für das Verständnis der unterschiedlichen Realisierungsmöglichkeiten von Speicherzellen, sowie für die daraus resultierenden Eigenschaften, ist das Verständnis der Funktionsweise des MOS-Transistors notwendige Voraussetzung. Dabei ist es immer wieder wichtig, aus einfachen Modellbetrachtungen Auswirkungen auf das Verhalten des Bauelements bei Änderungen von Transistorparametern abschätzen zu können.

Durch die dramatische Kostensenkung eines integrierten Transistors konnten die Halbleiterspeicher ihren Siegeszug antreten. Mittlerweile gibt es kaum noch elektronische Geräte des täglichen Lebens, bei denen nicht Daten in digitaler Form gespeichert werden. Zunächst wird zwischen flüchtigen und nichtflüchtigen Speichern unterschieden. Flüchtige Speicher verlieren die gespeicherte Information nach Wegfall der Versorgungsspannung, nichtflüchtige Speicher behalten die Information auch danach.

Zu den flüchtigen Halbleiterspeichern gehören die beiden Vertreter DRAM und SRAM. Beide Typen sind RAMs, d.h. Random Access Memories. Bei dieser Klasse von Speichern kann man wahlfrei auf jede Zelle des Speichers zugreifen. Ein SRAM ist dabei ein statisches RAM, d.h. die Information in den Speicherzellen wird solange ohne Informationsverlust gespeichert, bis sie durch eine neue ersetzt wird. Ein DRAM ist ein dynamisches RAM, die physikalische Realisierung der Speicherzelle gestattet nur die Speicherung der Information über einen gewissen Zeitraum. Die Information

muss also, um nicht verloren zu gehen, immer wieder aufgefrischt werden. Der DRAM Speicher und der SRAM Speicher werden in den Kapiteln 2 und 3 behandelt. Unter den flüchtigen Halbleiterspeichern findet man auch die Klasse mit seriellem Zugriff, bei der man immer nur in einer bestimmten Reihenfolge Informationen schreiben bzw. lesen kann. Vertreter dieser Klasse sind die Schieberegister sowie FIFO-Speicher. Diese werden im Kapitel 4 vorgestellt.

Bei den nichtflüchtigen Speichern spielt gegenwärtig der Flash-Speicher die wichtigste Rolle. Die beiden unterschiedlichen Architekturen des Flash-Speichers, NAND- und NOR-Flash, sind Gegenstand von Kapitel 5. Beim Flash-Speicher dient ein Transistor zur Speicherung der Information. Kann 1-Bit pro Transistor gespeichert werden, spricht man von der Single-Level-Cell Technologie (SLC). Ziel ist die Steigerung der Integrationsdichte und somit die Reduktion der Kosten pro Bit. So stellte die Firma Samsung im Jahre 2017 einen 512-Gbit NAND-Flash-Speicherchip mit einer Triple-Level-Cell Technologie (TLC) vor. Einzug gehalten hat der Flash-Speicher vor allem in elektronische Geräte die eine enorme Datenmenge für den Anwender speichern müssen, wie z.B. digitale Kameras, Smartphones etc. Bekannt geworden ist dieser Speichertyp als standarisierte Bauform des USB-Sticks, CompactFlash oder SD-Speicherkarte. Aber auch als Ersatz der Computer-Festplatte ist Form der Solid State Drive (SSD).

Für den nichtflüchtigen Speicher der Zukunft werden ein schnellerer Lese- und Schreibzugriff sowie beliebig viele Schreib-/Löschzyklen gewünscht. Deshalb wird an innovativen Speicherkonzepten intensiv geforscht. Als mögliche nächste Generation von nichtflüchtigen Speichern stehen heute drei Speichertypen im Fokus: der magnetoresistive RAM (MRAM), der resistive RAM (RRAM, ReRAM), und der Phasenwechselspeicher (PRAM, PCM). In Kapitel 6 wird die Funktionsweise dieser Speichertypen dargestellt. Auch werden die wesentlichen Eigenschaften der unterschiedlichen Technologien vergleichend gegenübergestellt.

Lernziele

Nach dem Studium dieser Kurseinheiten sollten Sie

- die wichtigsten Halbleiterspeicher klassifizieren können,
- den Lese-/Schreibvorgang der unterschiedlichen Speicherzellen detailliert darstellen können,
- Vor- und Nachteile der einzelnen Halbleiterspeicher beurteilen können,
- Eine Auswahl von Bitleitungs-/Wortleitungsarchitekturen an Hand von Zuverlässigkeitskriterien bewerten können,
- Tests von Speicherchips in den Produktionsprozess einordnen können.

Inhaltsverzeichnis

1 Grundkonzepte für Halbleiterspeicher

In diesem Kapitel werden zunächst die unterschiedlichen Kategorien von Halbleiterspeichern in einer Übersicht dargestellt. Für die wichtigste Kategorie, dem Speicher mit wahlfreiem Zugriff (RAM: Random Access Memory) wird dann die allgemeine Speicher-Architektur beschrieben. Abschließend werden Trends zur Entwicklung und Anwendung von flüchtigen sowie nichtflüchtigen Speichern aufgezeigt.

1.1 Klassifikation und Herausforderungen

Wie in Abb. 1.1 dargestellte, wird zunächst zwischen flüchtigen und nichtflüchtigen Speichern unterschieden. Flüchtige Speicher verlieren die gespeicherte Information nach Wegfall der Versorgungsspannung, nichtflüchtige Speicher behalten die Information auch danach. In einer zweiten Stufe der Klassifikation werden Speicher mit wahlfreiem Zugriff (random access) von denen mit seriellem Zugriff (serial access) unterschieden.

Flüchtige Speicher		**Nichtflüchtige Speicher**
Wahlfreier Zugriff	Serieller Zugriff	EEPROM Flash-Speicher MRAM ReRAM PRAM
DRAM SRAM	Schieberegister FIFO	

Abbildung 1.1: Klassifikation der Halbleiterspeicher

Zu den Vertretern der flüchtigen Speicher mit wahlfreiem Zugriff gehören der DRAM- und der SRAM-Speicher. Diese werden hauptsächlich eingesetzt, um Daten für die Verarbeitung im Prozessor bereitzustellen. Dabei kann man den Einsatzbereich für diesen Speichertyp in drei Hauptfelder unterteilen:

- integrierter SRAM-Speicher auf dem Prozessor (interner Cache),
- externer SRAM-Speicher (externer Cache),
- DRAM Hauptspeicher.

Zu den Vertretern der flüchtigen Speicher mit seriellem Zugriff gehören das Schieberegister sowie das FIFO (first-in first-out). Ein typischer Einsatzbereich dieser Speicher ist die Videocodierung (z.B. MPEG): die Daten müssen dem Prozessor

seriell zur Verfügung gestellt werden, daher ist kein wahlfreier Zugriff auf den Speicher erforderlich.

Bei den nichtflüchtigen Speichern hat der Flash-Speicher gegenwärtig den höchsten Marktanteil. Der Flash-Speicher ist Mitte der 80er Jahre entwickelt worden und hat den Vorläufer, den EPROM-Speicher heute abgelöst. Beim EPROM (erasable programmable read-only memory) musste die Information mittels UV-Licht gelöscht werden, bevor dieser Speicher mit neuer Information beschrieben werden konnte. Dazu musste der Baustein von der Platine gelöst werden und in einen „EPROM-Programmierer“ platziert werden. Beim EEPROM-Speicher (electrically erasable programmable read-only memory) ist es zwar möglich die Information spannungsgesteuert zu löschen (ohne den Baustein von der Platine zu lösen), allerdings ist die Speicherzelle mit zwei Transistoren doppelt so groß wie beim Flash-Speicher und deshalb von den Kosten pro Bit nicht mehr konkurrenzfähig. Da der EEPROM aber noch zur Speicherung kleinerer Datenmengen eingesetzt wird (z.B. im Telefon als Rufnummernspeicher), wird er ebenfalls im Lehrbrief behandelt.
Innovative nichtflüchtige Speicher wie MRAM (Magnetoresistive RAM), ReRAM (Resistive RAM, auch RRAM genannt) und PRAM (Phase-change RAM, auch PCRAM oder PCM genannt), um nur die Wichtigsten zu nennen, basieren nicht mehr auf der CMOS-Halbleitertechnologie. Diese neuartigen Speichertechnologien sollen mittelfristig den Flash-Speicher ersetzen. Hauptproblem des Flash-Speichers ist die recht begrenzte Anzahl von Lösch-Schreibzyklen in der Größenordnung von 10^5 bis 10^6. Allerdings ist es bisher noch nicht gelungen die hohe Speicherkapazität des Flash-Speichers auch bei den innovativen Speichern zu erreichen.

In Abb. 1.2 ist die Speicherkapazität pro Chip für die unterschiedlichen nichtflüchtigen Speichertypen über der Zeitachse aufgetragen. Man erkennt, dass die Speicherkapazität der neuesten Generation des RRAMs immer noch um den Faktor 8-10 kleiner ist, als die des Flash-Speichers. Für MRAM und PRAM ergibt sich nochmals eine um den Faktor 4-8 geringere Speicherkapazität. Dies liegt an der deutlich größeren Chipfläche pro Speicherzelle bei den neuartigen nichtflüchtigen Speichern (vgl. auch Kapitel 6).

Der Flash-Speicher profitiert von der Reduktion der Transistorgeometrien in der CMOS-Technologie. Gordon Moore prophezeite bereits in den 80er Jahren, dass alle 2-3 Jahre eine neue Technologiegeneration in Produktion geht und dass sich dabei die minimalen Strukturgrößen um 30% reduzieren, was etwa einer Verkleinerung der Chipfläche pro Transistor um den Faktor 2 entspricht.

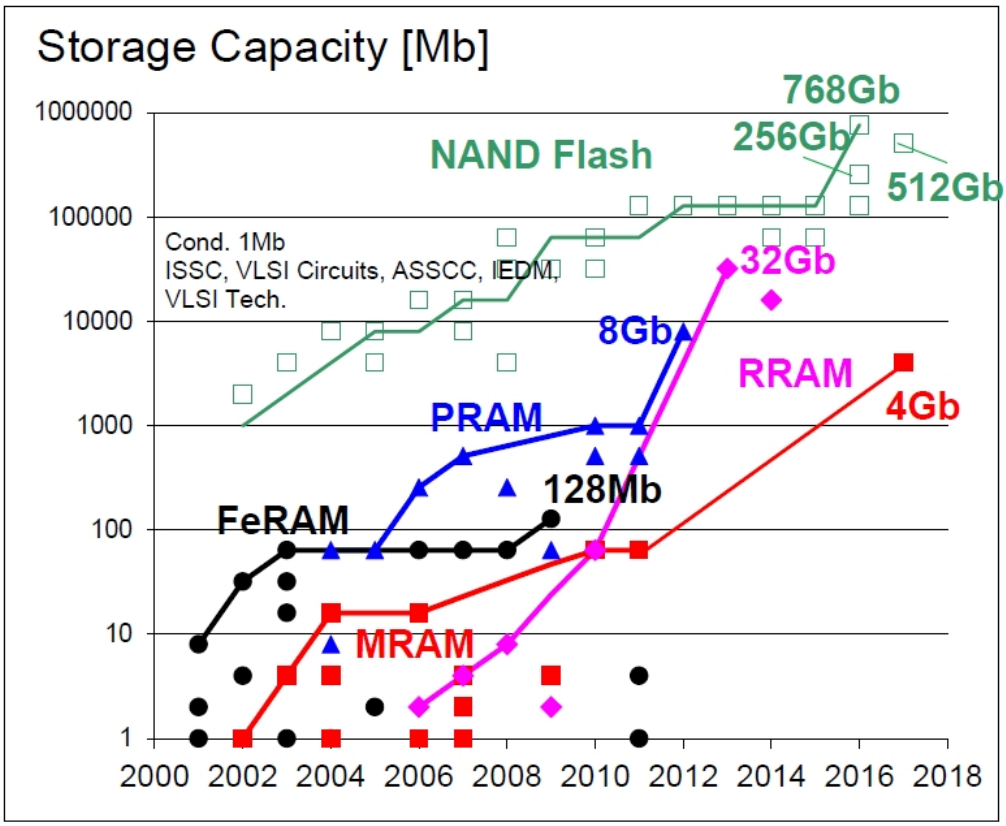

Abbildung 1.2: Entwicklung der Speicherkapazität [ISSCC17]

Die Herausforderung der Halbleiterhersteller eines speziellen Speichertypen besteht in der Reduktion der Kosten pro Bit durch die Einführung einer neuen Technologiegeneration. Ziel der Halbleiterhersteller ist natürlich diese Kostenreduktion in eine höhere Gewinnmarge umzusetzen, um die Investitionen zu decken und einen Nettogewinn ausweisen zu können. Das Produkt, also der Speicherbaustein, kann aber nur über den Preis an den Kunden gebracht werden, da es sich um ein so genanntes „Commodity Produkt“ (Eigenschaft des Produkts ist verbindlich vorgegeben) handelt. Hat also der Wettbewerber eine neue Technologiegeneration einige Monate früher in die Produktion eingeführt als das eigene Unternehmen, so kann der Wettbewerber aufgrund seiner geringeren Kosten pro Chip das Produkt zu einem geringeren Preis auf dem Markt anbieten und dabei selbst noch Gewinne einfahren. Diese Abwärtsspirale des Preises zeigt die Graphik in Abb. 1.3 am Beispiel eines 4GByte DRAM-Speichermoduls. Innerhalb eines Jahres, zwischen Oktober 2014 und Oktober 2015 hat sich der Marktpreis für dieses Speichermodul halbiert.

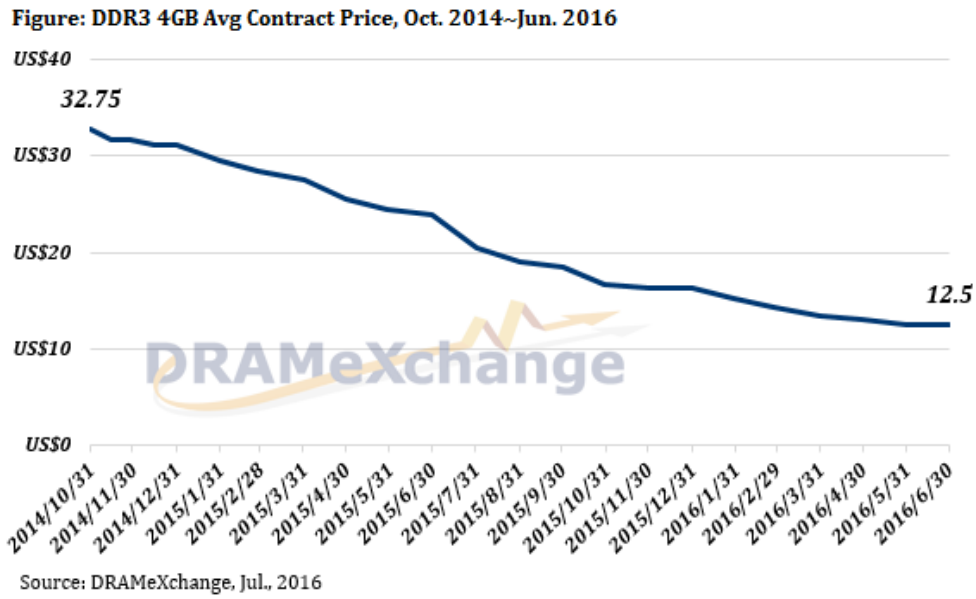

Abbildung 1.3: Entwicklung des Marktpreises eines DRAM-Speichers [DRAMeX16]

1.2 Architektur eines RAM-Speichers

Für Speicher mit wahlfreiem Zugriff (RAM-Speicher), dies betrifft insbesondere den DRAM-, den SRAM- sowie den Flash-Speicher, wird im nachfolgenden die allgemeine Organisationsform, auch Speicherchip-Architektur genannt, vorgestellt. Da der Zugriff auf jede Speicherzelle (storage cell) zum Lesen oder Speichern der Information ermöglicht werden soll, ist jede Speicherzelle über eine bestimmte Wortleitung (word line) und eine bestimmte Bitleitung (bit line) eindeutig auszuwählen. Dabei sind die Speicherzellen in einem so genannten Zellenfeld angeordnet. Die Wortleitung selektiert eine bestimmte Zeile innerhalb des Zellenfeldes, die Bitleitung eine bestimmte Spalte. In Abbildung 1.4 ist die Architektur eines RAM-Speichers als Blockschaltbild dargestellt. Eine Wortleitung wird durch eine bestimmte Adresse ausgewählt. Der Zeilendekoder (row decoder) reduziert die Anzahl der Adressbits. In dem dargestellten Beispiel von L-K Adressbits können mittels Binärkodierung 2^{L-K} Wortleitungen adressiert werden.

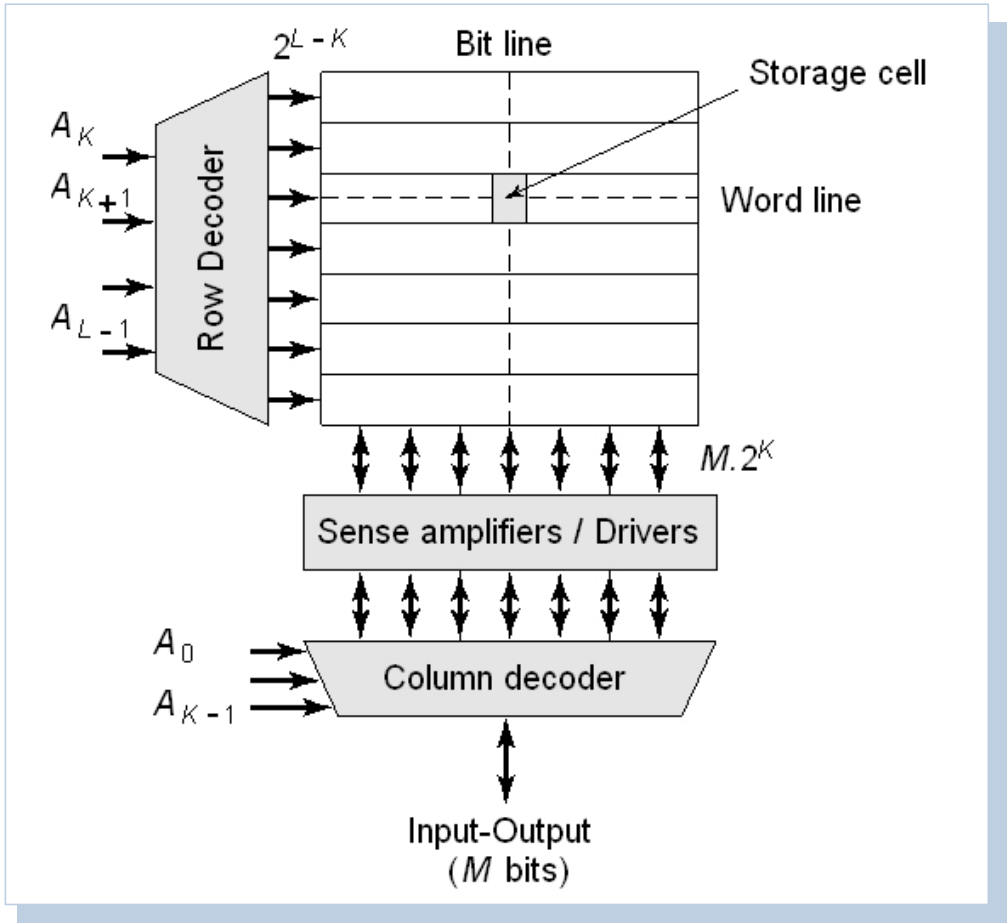

Abbildung 1.4: Blockschaltbild der Architektur eines RAM-Speichers [Rabaey03]

Neben den Adressleitungen, die mit den Pins des Chips verbunden sind, also von außen zugänglich sind, sind auch die Datenleitungen über die Pins mit der Außenwelt verbunden. Die Datenleitungen bilden einen bidirektionalen Datenbus (in der Abbildung als Input-Output Pfeil dargestellt), da die Information vom Speicher gelesen oder in den Speicher geschrieben werden kann. Die Wortbreite des Datenbus sei M-bit. Der Spaltendekoder (column decoder) wählt dann bei jedem Schreib-/Lesevorgang jeweils M benachbarte Zellen auf einer Wortleitung gleichzeitig aus. Mit K Adressleitungen für den Spaltendekoder können damit 2^K Worte der Wortbreite M-bit adressiert werden. Pro Zeile des Zellenfeldes befinden sich also $M \cdot 2^K$ Speicherzellen.

Außerdem ist der Abbildung 1.4 zu entnehmen, dass das Zellenfeld in etwa der Form eines Quadrates entspricht. Dies geschieht damit die Leitungslänge der Wortleitung etwa der der Bitleitung entspricht und damit etwa gleich große kapazitive Lasten von den Dekodern getrieben werden müssen. Geht man davon aus, dass die Signallaufzeit auf einer Leitung mindestens linear mit der Leitungslänge zunimmt, dann ist die quadratische Form des Zellenfeldes optimal. Werden Daten aus dem Zellenfeld gelesen, so müssen die Signale erst mit Hilfe von Verstärkerschaltungen

(sense amplifier) hinsichtlich ihres logischen Pegels bewertet werden. Dies ist notwendig, da der Signalhub auf den Bitleitungen meist gering ist. Erst nach dieser Verstärkung werden die Signale auf den Datenbus geschaltet.
Die Architektur nach Abb. 1.4 wurde in der Praxis verwendet bis zu einer Speicherkapazität von 256Kbits. Zugriffszeiten wurden dann zu langsam, aufgrund der langen Wort- und Bitleitungslängen. Dann kam das Konzept der hierarchischen Speicherarchitektur zum Tragen, wie es in Abbildung 1.5 dargestellt ist. Hierbei wird das Zellenfeld in P kleinere Blöcke aufgeteilt (P=4 für einen 512 Mbit DRAM Baustein) und eine zusätzliche Blockadressleitung (2-bit für P=4) selektiert zunächst den Block, bevor Zeilen- und Spaltenadresse die Zelle auswählen.

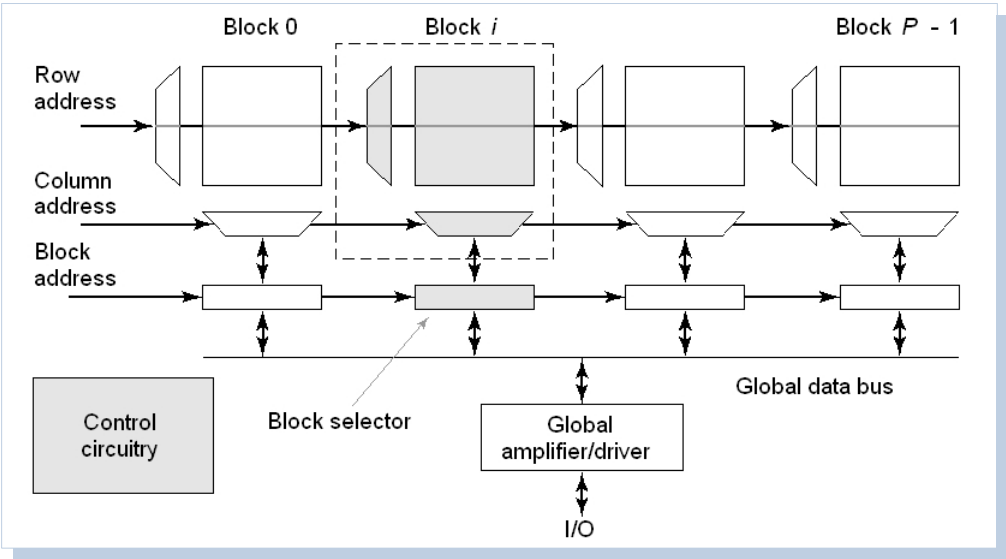

Abbildung 1.5: Hierarchische Architektur eines RAM-Speichers [Rabaey03]

Diese hierarchische Architektur eines RAM-Speichers bietet grundsätzlich zwei Vorteile:

- eine schnellere Zugriffszeit aufgrund kürzerer Wort- und Bitleitungen
- eine wesentlich geringere elektrische Leitungsaufnahme (Verlustleistung), da die Dekoder- und Verstärkerschaltungen in den nicht-adressierten Blöcken deaktiviert werden können

1.3 Trends

Moderne Mikroprozessoren nutzen einen hierarchischen on-chip Speicher, den sogenannten Cache-Speicher. Für diesen Cache-Speicher kommt der SRAM zum Einsatz, da dieser einen sehr schnellen Schreib-/Lesezugriff garantiert. Für Hochleistungsprozessoren wird ein Teil dieses Cache-Speichers durch eDRAM (Embedded DRAM, dt. eingebetteter DRAM) realisiert. Durch die kleinere

Speicherzelle des DRAM im Vergleich zum SRAM ergibt sich eine höhere Speicherdichte pro Chipfläche. Somit lassen sich höhere Speicherkapazitäten auf gleicher Chipfläche erreichen.

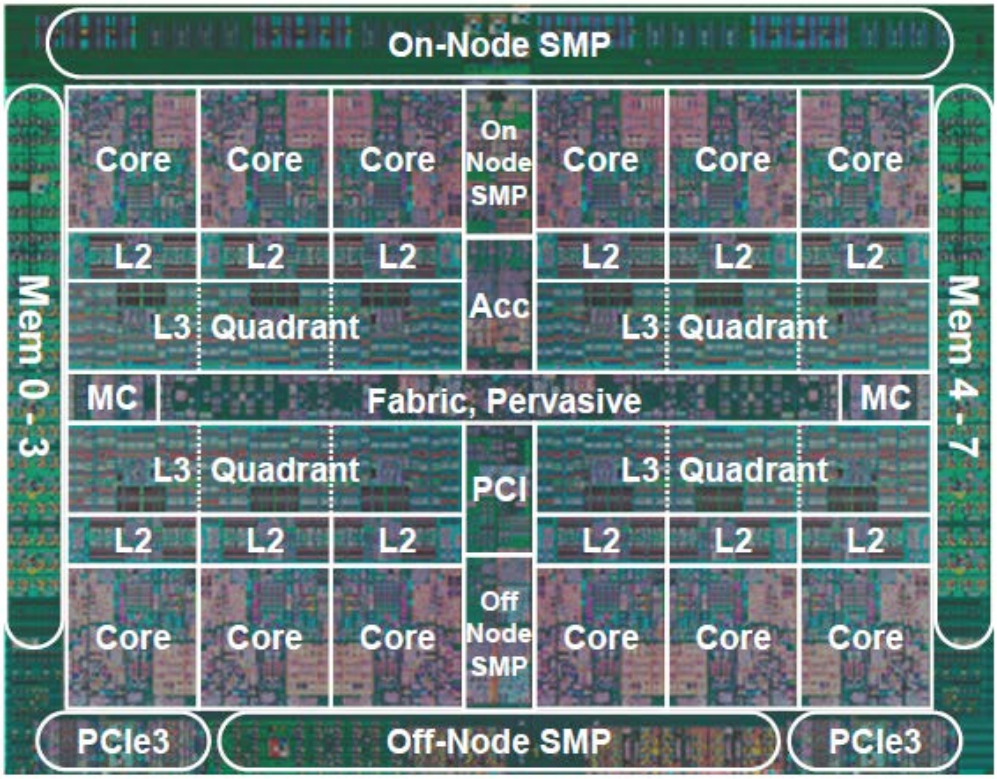

Abbildung 1.6: Architektur des IBM Power8 Prozessors [Fluhr14]

Abbildung 1.6 zeigt die Architektur des IBM Power8 Prozessors, hier mit dreistufigem Cache-Speicher (sog. L1, L2 und L3 Cache). Der L3 Cache mit jeweils 8 MB pro Kern (Core) wurde mittels eDRAM realisiert, L1 und L2 Cache mittels SRAM. Da der L3 Cache am weitesten vom Rechenkern entfernt sitzt, sind hier die Anforderungen an die Schreib-/Lesegeschwindigkeit nicht so hoch wie für den L1 und L2 Cache. Für den L3 Cache ist eine hohe Speicherkapazität wichtig, damit möglichst selten auf den wesentlich langsameren externen DRAM-Hauptspeicher zugegriffen werden muss.

Den gegenteiligen Ansatz verfolgt Microsoft/AMD mit dem SoC (System on Chip) Scorpio für die Spielekonsole Xbox: Hier wird eine sehr breite Busanbindung (384 Bit Interface) an einen externen GDRR5-SDRAM (Graphics Double Data Rate Synchronous DRAM) Speicher genutzt, um bei einer Datenrate von 6,8 GHz eine Speicherbandbreite von 326 GByte/s zu erzielen. Scorpio vereint dabei einen 8-Kern-Hauptprozessor (CPU) und einen Graphikprozessor (GPU) auf einem Chip. Abbildung 1.7 zeigt die Architektur mit 12 Speichercontrollern (DCT), die jeweils über einen 32 Bit Datenbus ein GDDR5 Speichermodul ansteuern.

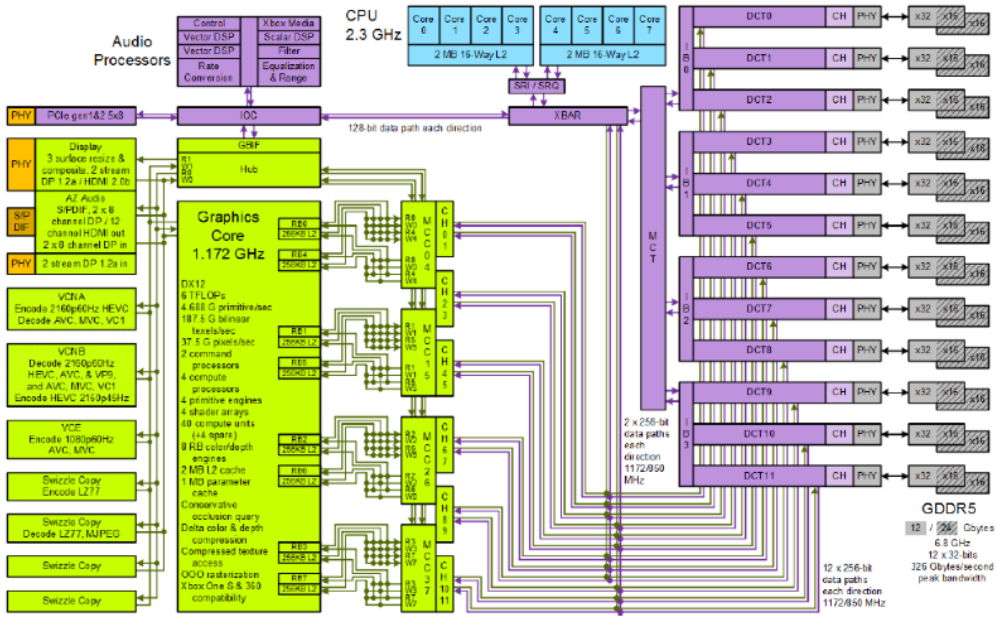

Abbildung 1.7: Architektur des SoC Scorpio [Microsoft17]

Ein weiterer Trend ist die Nutzung der neuartigen nichtflüchtigen Speicher (emerging NVRAM) als Ersatz des Flash-Speichers für Solid-State-Drives (SSD). Dies erklärt sich aus den wesentlich höheren Schreib- sowie Lesegeschwindigkeiten dieser NVRAMs. Abbildung 1.8 zeigt, dass diese Speicher eine bis zu 50-fach höhere Lesegeschwindigkeit, sowie eine bis zu 100-fach höhere Schreibgeschwindigkeit als NAND-Flash (NAND-MLC) erzielen können. So stellten Intel und Micron Technology im Jahre 2015 als Flash-Alternative für SSDs einen Speicher unter dem Eigennamen 3D XPoint (gesprochen „Crosspoint“) vor. Die Bezeichnung 3D weist auf die räumliche Gitterstruktur hin, an deren Kreuzungspunkten (Xpoint) die eigentliche Information sitzt.

Abbildung 1.8 zeigt auch, dass an neuartigen Flash-Speicherzellen gearbeitet wird: So hat die Firma Renesas Electronics im Jahr 2016 eine Split-Gate-MONOS-Flash-Speicherzelle (SG-MONOS) vorgestellt, die eine extrem hohe Lesegeschwindigkeit, aber sehr geringe Schreibgeschwindigkeit zeigt. Werden diese Speicherzellen zusammen mit einem Mikrocontroller auf einem Chip integriert, sogenanntes Embedded-Flash, so eignet sich dieser Chip insbesondere für komplexe Motorsteuerungsaufgaben in der Automobilindustrie. Der Code für die Steuerung ist dann im Embedded-Flash gespeichert und kann sehr schnell ausgelesen werden.

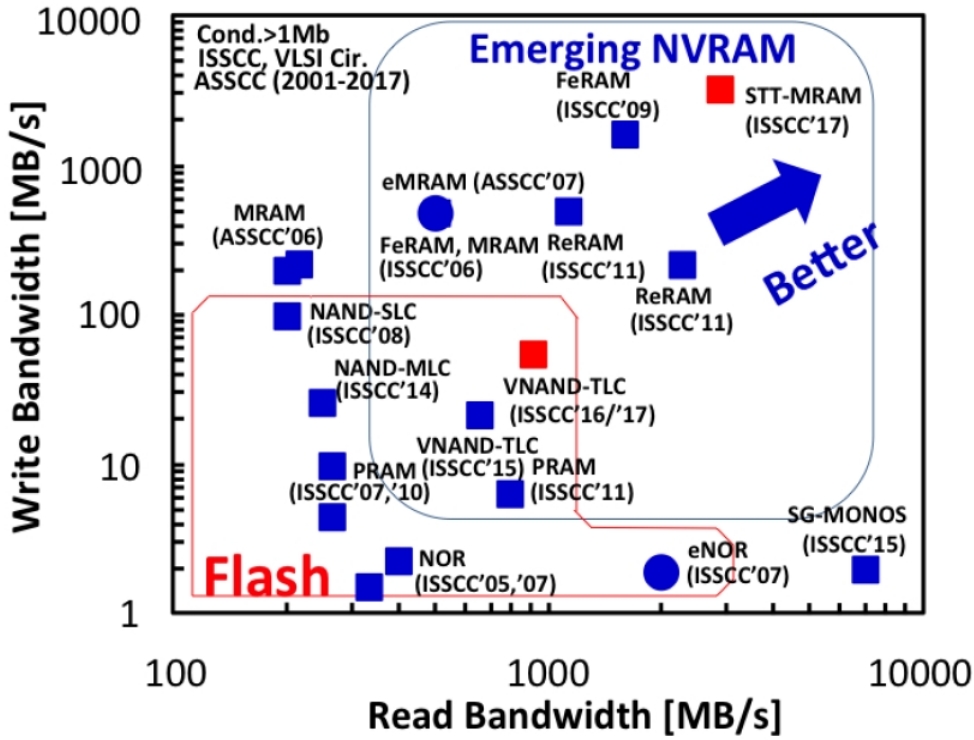

Abbildung 1.8: Schreib-/Lesegeschwindigkeit im Vergleich [ISSCC17]

Die Hürden für den Einzug neuer Halbleiterspeicher in den Massenmarkt liegen hoch. Neben den technischen Eigenschaften wie

- Lese-/Schreibgeschwindigkeit,
- Zuverlässigkeit,
- Geringer Leistungsbedarf,

sind es insbesondere die

- Kosten pro Speicherbit,

die für den Erfolg auf dem Massenmarkt entscheidend sind. Diese Kosten ergeben sich insbesondere durch die

- Speicherzellengröße

sowie die

- Skalierfähigkeit der Speicherzelle.

Übungsaufgaben

1.1. Das Zellenfeld eines 4Gb-Speicherchips soll aus gleicher Anzahl von Zeilen- und Spaltenadressen aufgebaut werden. Es soll eine hierarchische Architektur mit 8 Blöcken zum Einsatz kommen. Wie viele Adressbits sind für den Zeilen- bzw. Spaltendekoder erforderlich?

1.2. Warum und für welchen Zweck wird eDRAM eingesetzt? Warum eignet sich eDRAM nicht generell für alle Cache-Speicher?