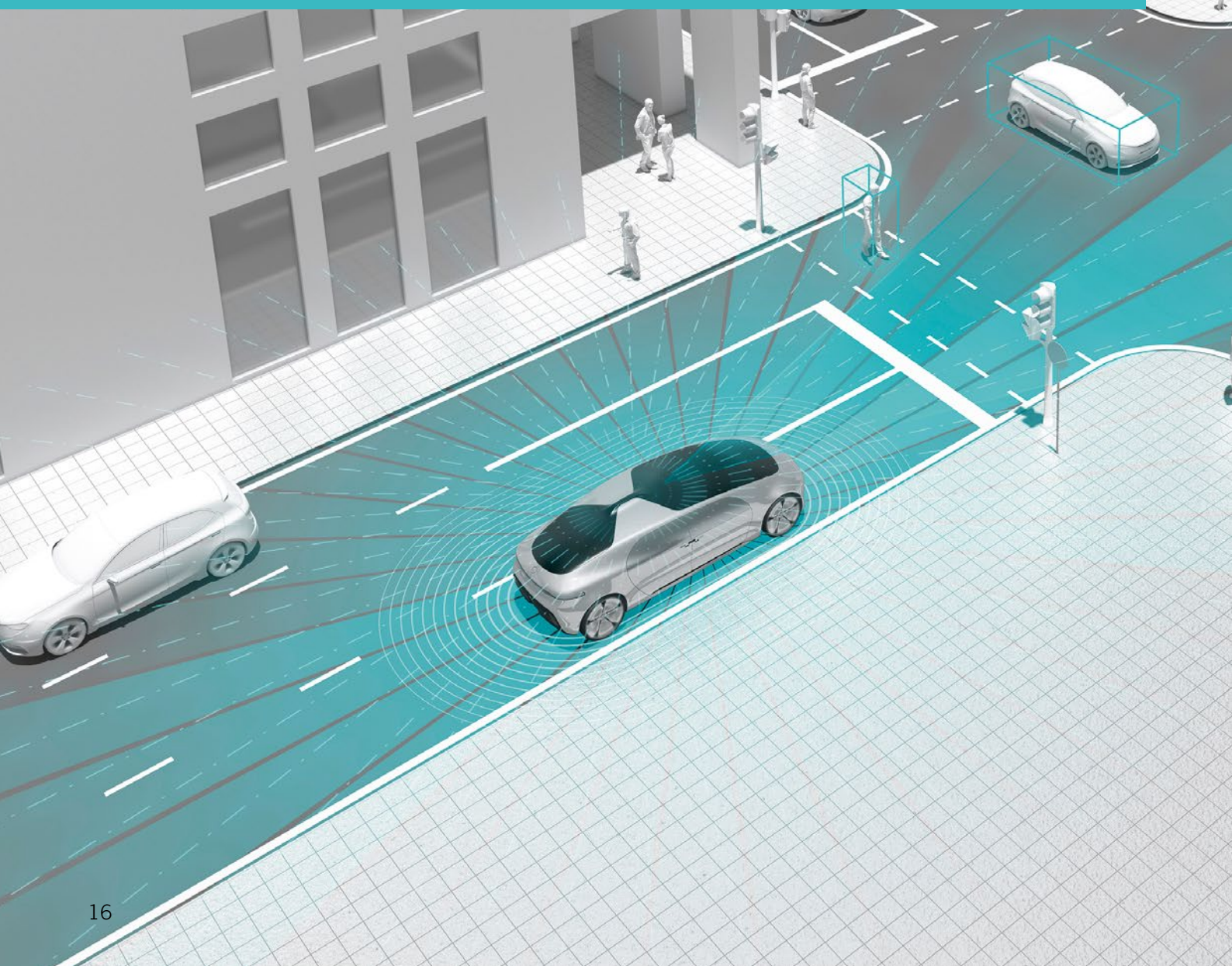
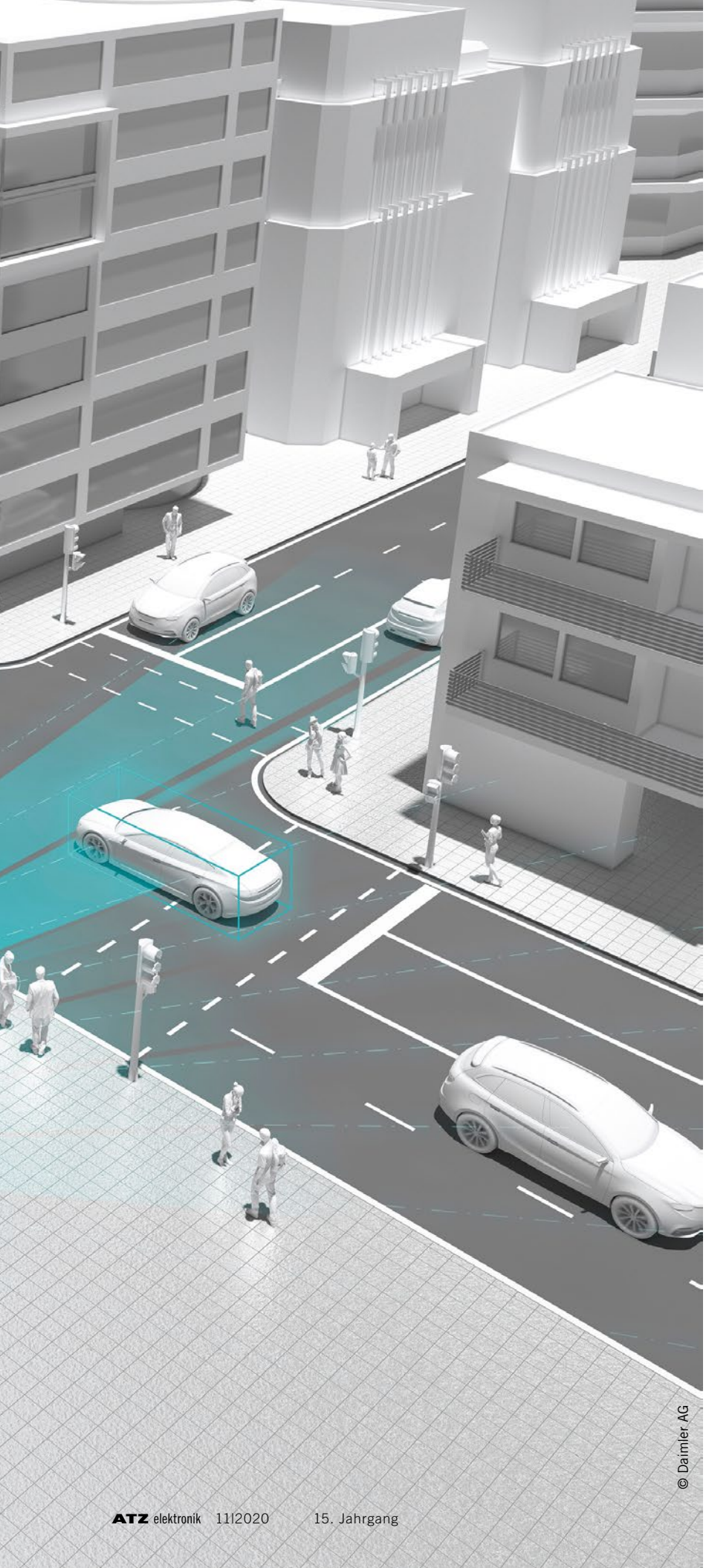
Moderne Speicherarchitekturen für leistungsfähige Infotainmentsysteme und autonomes Fahren

Halbleiterspeicher wie DRAMs oder Flash sind ein oft wenig beachteter Aspekt, der aber in zukünftigen Fahrzeugarchitekturen zum limitierenden Faktor wird, wenn in der Konzeption wichtige Aspekte vernachlässigt werden. Das Fraunhofer IESE, die TU Kaiserslautern und Mercedes-Benz Research & Development zeigen, worauf dabei zu achten ist.





© Daimler AG

AUTOREN

Dr. Matthias Jung
ist Expert Virtual Hardware Engineering beim Fraunhofer-Institut für Experimentelles Software Engineering IESE in Kaiserslautern.

Michael Huonker
ist Engineer Design High-Computing Platforms Architecture bei der Mercedes-Benz AG R&D in Sindelfingen.

Ralf Kalmar
ist Business Area Manager Automotive & Commercial Vehicles beim Fraunhofer-Institut für Experimentelles Software Engineering IESE in Kaiserslautern.

Prof. Dr. Norbert Wehn
ist Professor am Lehrstuhl Entwurf mikroelektronischer Systeme der Technischen Universität Kaiserslautern.

NEUE STRUKTUREN DURCH NEUE KOMPONENTEN

Die Menge an Daten, die in heutigen Fahrzeugen in Echtzeit verarbeitet werden müssen, nimmt stetig zu. Dabei gibt es Anforderungen an einen niedrigen Leistungsverbrauch sowie einen großen Kostendruck, der dazu führt, dass verstärkt Komponenten eingesetzt werden, die ursprünglich für den Consumer-Markt entwickelt wurden. Dies führt zu heterogenen Rechenplattformen, die aus GPUs, dedizierten Beschleunigern, CPUs

sowie insbesondere DRAM und Flash-Speicher bestehen.

Eine große Rolle für zukünftige Automobilanwendungen spielt dabei die Interaktion mit dem Fahrer und die Entlastung des Fahrers bei seinen Fahraufgaben. Somit kann in einem Auto zwischen den Bereichen Fahrerassistenz (ADAS und autonomes Fahren) und Infotainment unterschieden werden. Auf den ersten Blick zeigen sich für beide Bereiche ähnliche Anforderungen, denn sowohl DRAM als auch Flash-Speicher werden in der Speicherhierarchie von beiden Bereichen benötigt. Jedoch ergeben sich beim genaueren Hinsehen deutliche Unterschiede: Im Infotainment kommt wegen der hohen Speicherkapazität verstärkt Flash zum Einsatz, während bei der Fahrerassistenz überwiegend schnelle DRAM-Speicher benötigt werden. Im Folgenden werden diese Unterschiede in Bezug auf die verwendeten Speichertechnologien näher beleuchtet.

LEISTUNGSFÄHIGE SYSTEME NÖTIG

Im Infotainment-Bereich steht die User Experience im Vordergrund, welche sich an bekannten Konzepten von Smartphones, Smart TV und anderen IoT-Geräten orientiert. Das System soll unmittelbar auf Eingabemöglichkeiten wie Touch, Gesten und Sprache reagieren, sowie Anzeigen auf großen und hochauflösenden Bildschirmen ausgeben. Die Software wird dabei mit einem robusten Updatesystem auf dem aktuellen Stand gehalten, und es ist möglich, neue Funktionen oder Applikationen in kurzer Zeit nachzuladen. Ein modernes Fahrzeug ist heute mit der IoT-Welt vernetzt und an ein Rechenzentrum angebunden. Um die User Experience zu gewährleisten, ist ein leistungsfähiges System on Chip (SoC) wie zum Beispiel ein Nvidia Xavier mit genügend Rechen- und Grafikleistung notwendig. Um die vielen parallelen Anwendungen möglichst schnell zu starten und auszuführen, sind große und schnelle Speicher notwendig. Ein High-End-System weist dabei heute schon bis zu 16-GB-LPDDR4x und 320-GB-Flash-Speicher auf.

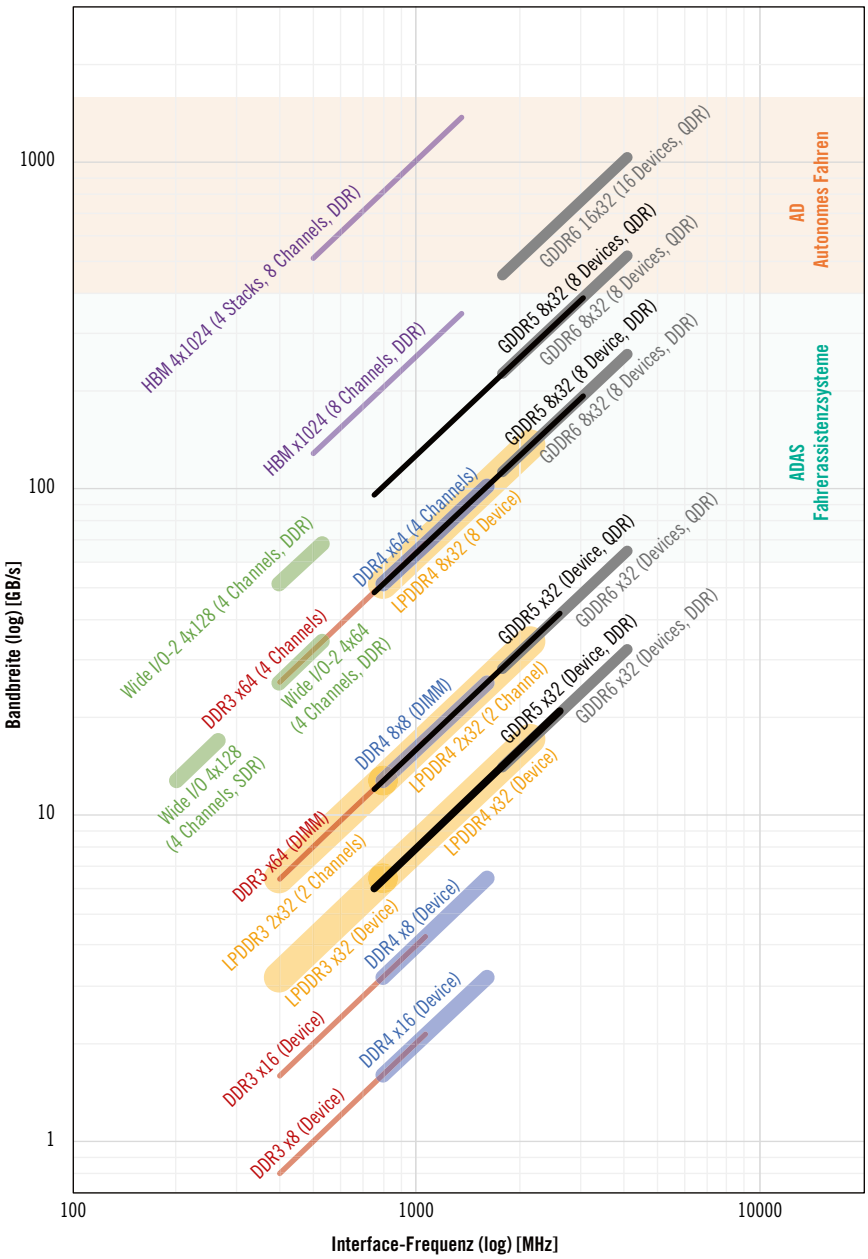


BILD 1 DRAM-Bandbreite in Abhängigkeit von der Interface-Frequenz (© Fraunhofer IESE)

KAPAZITÄTSSTEIGERUNG BEI FLASH DURCH 3-D-ZELLEN

Flash ist ein nichtflüchtiger Speicher, der in den letzten Jahren die mechanischen Festplatten fast vollständig verdrängt hat. Grund dafür ist die kontinuierliche Kapazitätssteigerung von 20 % pro Jahr. Die Verkleinerung der Flashzelle ist bis zu einer Strukturgröße von 15 nm in planarer Technologie fortgeschritten. Ab diesem Wert wird die Zahl der Elektronen zur Informationsspeicherung so klein, dass die Zuverlässigkeit darunter leidet.

Eine Lösung wurde durch die 3-D-Zelltechnologie gefunden, bei der vertikal aufgebrachte Ebenen (Layer) zur Strukturierung benutzt werden. Die eigentlichen Transistoren zur Informationsspeicherung sind dabei vertikal angeordnet. Im Moment sind Bausteine mit 64 bis 128 Lagen erhältlich. Durch diese 3-D-Anordnung konnte die Zahl der Elektronen in der Zelle erhöht werden. Eine Steigerung der Kapazität konnte durch die Nutzung der Multi-Level-Cell-Technik eingeführt werden. Dabei wird die Information mit analogen Spannungspegeln in der Zelle abgespeichert. Momentan sind im Automobil Bausteine mit 2 Bit (MLC) und 3 Bit (TLC) pro Zelle im Einsatz (4 oder 8 Zustände). In der 3-D-TLC-Technologie stehen für jedes Bit drei Mal so viele

Elektronen zur Verfügung wie in einer vergleichbaren 2-D-planaren Zelle.

Jeder Programmiervorgang beschädigt die Flashzelle geringfügig. Der Begriff der Endurance gibt an, wie oft die Zelle beschrieben werden kann. Typischerweise liegt die Zahl der Zyklen für Multi-Bit-Zellen (MLC/TLC) im Bereich von 3000 und für Einzel-Bit-Zellen (SLC) bei über 30.000. Die Sperrschicht des FET in der Zelle altert mit jedem Beschreiben und beeinträchtigt dabei die Fähigkeit der Zelle, die Ladung aufzunehmen und zu halten. Die Data Retention sagt dabei aus, wie lange die Zelle in der Lage ist, die Information zu speichern. Für eine Anwendung im Automobil muss diese dabei deutlich länger als ein Jahr gespeichert bleiben, während im Serverbereich eine SSD über Wochen selten stromlos ist und die Haltedauer der Information oft im Bereich von wenigen Wochen liegt. Ein weiterer Einflussfaktor auf die Qualität der geschriebenen Information ist die Temperatur.

Der Flashspeicher muss in einem Temperaturbereich zwischen -40 und +105 °C über 15 Jahre zuverlässig arbeiten. Die Lebensdauer wird dabei nicht nur durch die verbauten Speicherchips und deren Technologie bestimmt. Ein eingebauter Controller übernimmt wesentliche Teile der Steuerung und der Kaschierung von Fehlern. Er fügt in den Datenstrom eine ECC-Information ein, mit der falsche Informationen korrigiert werden können. Für 2-D-NAND-Bausteine sind herkömmliche BCH-Codes und bei 3-D-TLC-Bausteinen LDPC-Codes im Einsatz. Weitere Aufgaben des Controllers sind das Wear Leveling, das eine gleichmäßige Abnutzung der Zellen sicherstellt, und die Garbage Collection, mit der gelöschte Speicherbereiche wieder nutzbar werden.

Höhere Anforderungen an Flashspeicher werden in Zukunft vor allem in Bezug auf die Schreib-/Lesegeschwindigkeit und die Zahl der Schreibzyklen gestellt. Die hohe Schreib- und Lesegeschwindigkeit kommt vor allem durch die Konsolidierung von Steuergeräten zustande. Neue Anwendungen wie das Aufzeichnen von Daten für das Trainieren von KI-Netzwerken oder das Aufzeichnen von Kamerabildern (DashCam) machen eine hohe Zahl von Schreibzyklen notwendig. Um die Latenz und die Sicherheitsanforderungen im Hypervisor beziehungsweise Betriebssystem zu verbessern, kommen Server-Technologien zum Einsatz. Dies ist exemplarisch die IO-Virtualisierung im Speicher (SR-IOV). Eine schnelle Datenübertragung vom Flashbaustein zum SoC kann nur über schnellere Schnittstellen erreicht werden. Neue Technologien wie UFS 3.1 oder PCIe Gen4 bieten dabei Datenraten von 2,9 GB/s (UFS3.1 2-Lanes) oder 3,94 GB/s (PCIe Gen4 2-Lane). Nur mit den neuesten Technologien können in den Bereichen geforderte Speicherkapazität, Geschwindigkeit und Funktion die technologischen Anforderungen erfüllt werden.

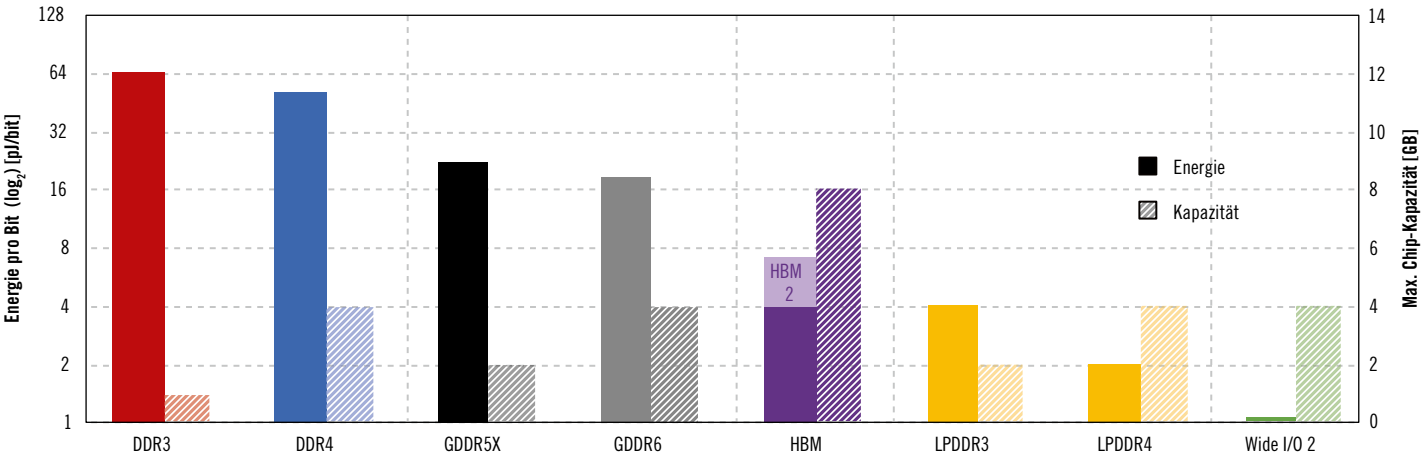


BILD 2 DRAM-Speicherkapazität und Leistungsverbrauch (© Fraunhofer IESE)

DRAM ALS LEISTUNGSFÄHIGE SPEICHER

Das autonome Fahren soll den Benutzer bei seiner Fahraufgabe entlasten. In den höchsten Ausbaustufen Level 4 und 5 fährt dabei das Fahrzeug eigenständig, wobei hier die Sicherheit an höchster Stelle steht. Für die notwendigen Sicherheitsanforderungen werden viele Komponenten redundant eingesetzt und das gesamte System wird sensorzentrisch aufgebaut. Dadurch erhöht sich die Menge an Daten, die in Echtzeit verarbeitet werden müssen. Mit neuronalen Netzen werden die Daten ausgewertet und interpretiert. Große neuronale Netze wie das VGG-16 benötigen pro Durchlauf über 15,5 Milliarden MAC-Operationen (Multiply-Accumulate) und haben 138 Millionen Gewichte. Dies erfordert spezialisierte Recheneinheiten (KI-Beschleuniger) und eine breitbandige Speicheranbindung mit niedriger Latenz, um die Echtzeitanforderungen und damit die Sicherheit gewährleisten zu können. Im Gegensatz zum Infotainment liegt hier der Schwerpunkt nicht auf der Speicherkapazität, sondern auf Latenz und Durchsatz. Diese Anforderungen können nach heutigem Stand der Technik nur DRAM-Speicher erfüllen.

DRAMs sind flüchtige Speicher, die auf minimale Kosten pro Bit optimiert sind. Daher muss vor allem das DRAM-Package kostengünstig sein, was die verfügbaren Pins und somit die Bandbreite begrenzt. Aus diesem Grund verfügen DRAMs über eine komplexe interne Architektur mit internem Prefetching, um die Lücke zwischen der extern verfügbaren Speicherbandbreite und der internen Latenzzeit zu schließen. In den letzten Jahren wurden viele neue DRAM-Architekturen vorgestellt (zum Beispiel DDR4, LPDDR4, GDDR6, Wide I/O, HMB2), **BILD 1**, welche sich in Bandbreite, Latenz, Kapazität und Leistungsverbrauch unterscheiden, **BILD 2**.

Aktuelle ADAS-Anwendungen erfordern eine DRAM-Bandbreite von circa 100 GB/s. Diese Anforderung ist zurzeit noch mit LPDDR-Lösungen, die ursprünglich für Smartphones und Tablets entwickelt wurden, problemlos zu erfüllen. In den nächsten Jahren werden sich durch den Übergang auf die Autonomielevel 4 und Level 5 die Anforderungen auf bis zu 400 bis 1000 GB/s erhöhen, was langfristig nur mit Grafikspeichern (GDDR) oder HBM2 DRAM umsetzbar ist. Allerdings hängt bei DRAMs die Bandbreite stark von den laufenden Anwendungen ab. Daher wird es eine große Herausforderung für die Entwickler werden, die maximal verfügbare Bandbreite bestmöglich auszunutzen. Darüber hinaus weisen DRAMs aufgrund der komplexen Protokoll- und Laufzeitoptimierungen des Speichercontrollers ein komplexes Latenzverhalten auf. Dies führt zu einer weiteren Herausforderung für die Entwickler, da es schwierig ist, das Zeitverhalten für sicherheitskritische Echtzeitanwendungen vorherzusagen.

Aktuelle autonome Fahrzeugprototypen verbrauchen bis zu 2,5 kW für die Datenverarbeitung. Angestrebt werden integrierte Lösungen, die etwa 75 bis 500 W verbrauchen werden. Unter der Annahme, dass 1000 GB/s Bandbreite benötigt werden, liegt die Leistungsaufnahme von 16-GDDR6-Chips bei 150 W und bei vier HBM2-Stacks bei etwa 60 W [1]. Gerade unter dem Gesichtspunkt der Elektrifizierung der Fahrzeuge ist dies ein wichtiger Parameter. Um niedrige Startzeiten zu garantieren, wird das DRAM nach dem Parken des Fahrzeuges in einen Energiesparmodus, den sogenannten Self-Refresh, versetzt. So kann nach erneutem Starten des Fahrzeugs die Rechenplattform unmittelbar weiterarbeiten und muss nicht komplett hochgefahren werden. Auch hier existiert eine Leistungsaufnahme, die nicht vernachlässigbar ist.

GRENZEN DER DRAM-TECHNOLOGIE

DRAM-Technologien weisen eine große Parametervariation auf und die Speicherzellen müssen wegen eines Leckstroms regelmäßig aufgefrischt werden (Refresh). Die Speicherzellen sind außerdem sehr empfindlich gegenüber hohen Temperaturen, da dieser Leckstrom mit steigender Temperatur exponentiell zunimmt. Um die Zuverlässigkeit aufrechtzuerhalten, muss die Refresh-Frequenz bei hohen Temperaturen somit erhöht werden. Vor allem im Automobilbereich, wo, wie bereits erwähnt, Temperaturen bis zu +105 °C möglich sind, hat dies wiederum eine große Auswirkung auf die Bandbreite und Latenz, da während eines Refreshvorgangs der Zugriff auf die Daten nur eingeschränkt möglich ist.

Um den Sicherheitsanforderungen der ISO 26262 gerecht zu werden, müssen Diagnosemechanismen wie zum Beispiel Speichertests oder Fehlerkorrekturmechanismen (ECC) eingesetzt werden.

Da die Technologieskalierung weit vorangeschritten ist, implementieren die Speicherhersteller bereits im DRAM selbst ECC-Mechanismen, um ihre Ausbeute zu erhöhen und die korrekte Funktionalität des Bausteins zu gewährleisten. Auch Gegenmaßnahmen zur Verhinderung von Row-Hammer-Angriffen müssen betrachtet werden [2]. Somit ergeben sich beim Einsatz von DRAMs für das autonome Fahren auch hier große Herausforderungen in Bezug auf Safety und Security.

AUSBLICK UND ALTERNATIVEN

Daimler hat bereits frühzeitig diese neuen Herausforderungen erkannt und untersucht deshalb systematisch im Entwicklungsprozess, ob die Leistungsfähigkeit der Speicher den hohen Anforderungen der Systeme einer neuen Baureihe genügt. In Zusammenarbeit mit dem Fraunhofer IESE und der TU Kaiserslautern werden Lösungen vermessen und geeignete Speichertechnologien ausgewählt. Die Optimierungsmöglichkeiten reichen bis zur Entwicklung anwendungsspezifischer DRAM-Controller. Dazu werden hardwarenahe Simulationen genutzt, die mit dem Open-Source-Tool DRAMSys [3] umgesetzt werden.

Die zunehmende Menge an Daten in Fahrzeugen führt dazu, dass auch Elektronikkomponenten verwendet werden, die ursprünglich für den Consumer-Markt entwickelt wurden. Dazu gehören Flash- und DRAM-Speicher, die somit verstärkt in künftigen E/E-Architekturen vorkommen werden. Sowohl DRAM- als auch Flash-Speicher sind mit Einschränkungen behaftete Technologien, die hauptsächlich für den Consumer-Bereich entwickelt wurden. Für den Einsatz in sicherheitskritischen Systemen für das autonome Fahren sind geschickte Mechanismen für das Management der Komponenten zwingend erforderlich, um die Anforderungen an Sicherheit, Performanz und Leistungsverbrauch zu gewährleisten. Durch die Einführung von Storage Class Memories (SCM) wie zum Beispiel STT-MRAM oder RRAMs stehen langfristig Alternativen für DRAM und Flash zur Verfügung, welche sowohl nichtflüchtig sind, als auch eine hohe Performanz bieten. Bis dahin muss sich die Automobilindustrie allerdings weiterhin auf die aktuellen Speichertechnologien verlassen.

LITERATURHINWEISE

[1] Jung, M.; McKee, S. A.; Sudarshan, C.; Dropmann, C.; Weis, C.; Wehn, N.: Driving Into the Memory Wall: The Role of Memory for Advanced Driver Assistance Systems and Autonomous Driving. Tagung ACM International Symposium on Memory Systems (MEMSYS 2018), Washington, DC, USA, Oktober 2018

[2] Kim, Y. et al.: Flipping bits in memory without accessing them: An experimental study of DRAM disturbance errors. Tagung ACM/IEEE 41st International Symposium on Computer Architecture (ISCA), Minneapolis, MN, USA, 2014

[3] DRAMSys 4.0. Online: https://github.com/tukl-msd/DRAMSys, aufgerufen: 17.08.2020