VDE SPEC

Neuromorphic Computing auf Basis neuartiger Bauelemente – ein Schichtenmodell für die Entwicklung von KI-Hardware

VDE SPEC 90033 V1.0 (de)

VDE

Vorwort

Veröffentlichungsdatum dieser VDE SPEC: 01.11.2024.

Zur vorliegenden VDE SPEC wurde kein Entwurf veröffentlicht.

Diese VDE SPEC wurde nach dem VDE SPEC-Verfahren erarbeitet. Die Erarbeitung von VDE SPEC erfolgt in Projektgruppen und nicht zwingend unter Einbeziehung aller interessierten Kreise.

Diese VDE SPEC ist nicht Bestandteil des VDE-Vorschriftenwerks oder des Deutschen Normenwerks. Diese VDE SPEC ist insbesondere auch keine Technische Regel im Sinne von § 49 EnWG.

Verfasser dieser VDE SPEC sind:

- Beyer, Sven, GlobalFoundries
- Bolzani Pöhls, Leticia, RWTH Aachen
- Dittmann, Regina, FZ Jülich
- Dudek, Damian, VDE ITG
- Gemmeke, Tobias, RWTH Aachen
- Gude, Michael, CologneChip
- Joseph, Jan Moritz, Roofline AI
- Kohlstedt, Hermann, Universität Kiel
- Leupers, Rainer, RWTH Aachen
- Mikolajick, Thomas, TU Dresden
- Nielen, Lutz, aixACCT Systems
- Paintz, Christian, Melexis
- Thiem, Steffen, X-FAB Semiconductor Foundries
- Waser, Rainer, FZ Jülich
- Wehn, Norbert, RPTU Kaiserslautern
- Wenger, Christian / IHP
- Wiefels, Stefan, FZ Jülich
- Wirth, Matthias, VDE WIN
- Ziegler, Martin, Universität Kiel

Zurzeit gibt es in keiner deutschen Norm eine Regelung zu diesem Thema.

Trotz großer Anstrengungen zur Sicherstellung der Korrektheit, Verlässlichkeit und Präzision technischer und nicht-technischer Beschreibungen kann die VDE SPEC-Projektgruppe weder eine explizite noch eine implizite Gewährleistung für die Korrektheit des Dokuments übernehmen. Die Anwendung dieses Dokuments geschieht in dem Bewusstsein, dass die VDE SPEC-Projektgruppe für Schäden oder Verluste jeglicher Art nicht haftbar gemacht werden kann. Die Anwendung der vorliegenden VDE SPEC entbindet den Nutzer nicht von der Verantwortung für eigenes Handeln und geschieht damit auf eigene Gefahr.

Im Zuge der Herstellung und/oder Einführung von Produkten in den Europäischen Binnenmarkt muss der Hersteller eine Risikoanalyse durchführen, um zunächst festzustellen, welche Risiken das Produkt möglicherweise mit sich bringt. Nach Durchführung der Risikoanalyse bewertet er diese Risiken und ergreift gegebenenfalls geeignete Maßnahmen, um die Risiken wirksam zu eliminieren oder zu minimieren (Risikobewertung). Die vorliegenden VDE SPEC entbindet den Nutzer nicht von dieser Verantwortung.

Es wird auf die Möglichkeit hingewiesen, dass einige Elemente dieses Dokuments Patentrechte berühren können. VDE ist nicht dafür verantwortlich, einige oder alle diesbezüglichen Patentrechte zu identifizieren

Executive Summary

Die Entwicklung unserer digitalen Welt geht mit einem immer höheren Rechenleistungs- und Energiebedarf einher – mit entsprechenden Folgen für unser Klima. Es wird prognostiziert, dass bis 2030 rund ein Fünftel der weltweiten elektrischen Energieproduktion für die Informationstechnik (IT) benötigt wird (Jones, 2018). Eines der am schnellsten wachsenden Felder innerhalb der IT ist derzeit die Künstliche Intelligenz (KI), die jedoch bisher eine sehr ineffiziente Energiebilanz aufweist (Dhar, 2020). Grund hierfür ist die konventionelle CMOS-Technologie, die uns als Computing-Hardware zur Verfügung steht. Doch nicht nur die Hardware weist Überarbeitungspotential auf, sondern auch die auf ihr zum Einsatz kommenden Algorithmen, die klassisch auf dem sogenannten „von-Neumann-Konzept“ realisiert sind.

Diese Limitierungen machen die Entwicklung neuer, ressourcenschonender und energieeffizienter Technologien wichtiger, bei denen die Entwicklung neuer Materialien, Technologien und Mikroelektroniken für Informationsverarbeitung, -speicherung und -übertragung neue Konzepte erfordert.

Mit dem Ansatz des Neuromorphic Computing (NMC) orientiert man sich an Aufbau und Wirkungsweise biologischer Nervensysteme und versucht damit neuronale Netze naturgetreu nachzubilden. Biologische Neuronen (Nervenzellen) können Informationen sowohl verarbeiten als auch speichern. Dabei hat ein menschliches Gehirn eine Leistungsaufnahme von etwa 20 Watt, während Serverfarmen mit GPU-basierten Rechenknoten auf mehrere Mega-Watt kommen. Auch wenn der Vergleich stark vom jeweiligen Anwendungsfeld abhängt, kann man ihn für bestimmte Aufgaben heranziehen und damit ist eindeutig, wie eklatant der Unterschied bei der Energieaufnahme zwischen technischen Systemen und der Natur ist.

Die vorliegende VDE SPEC zu NMC-Technologien dient der Festlegung von einheitlichen Begrifflichkeiten in diesem Forschungs- und Entwicklungsbereich sowie der Entwicklung eines abgestimmten NMC-Schichtenmodells (Bild 1) zur Einteilung der unterschiedlichen Technologieebenen von der Materialzusammensetzung bis hin zur Systemebene. Das NMC-Schichtenmodell ist ein wesentlicher Bestandteil dieser VDE SPEC und soll für weitere Entwicklungen als Referenzmodell für den Entwurf von Bauelementen, den Schaltungsentwurf, die Algorithmenverwendung und auch die Systemarchitektur dienen. Mithilfe dieses NMC-Schichtenmodells soll Forschungsaktivitäten der Weg hin zu Entwicklungen und Bedarfen in der Anwendung geebnet werden – vorrangig in der industriellen Umgebung.

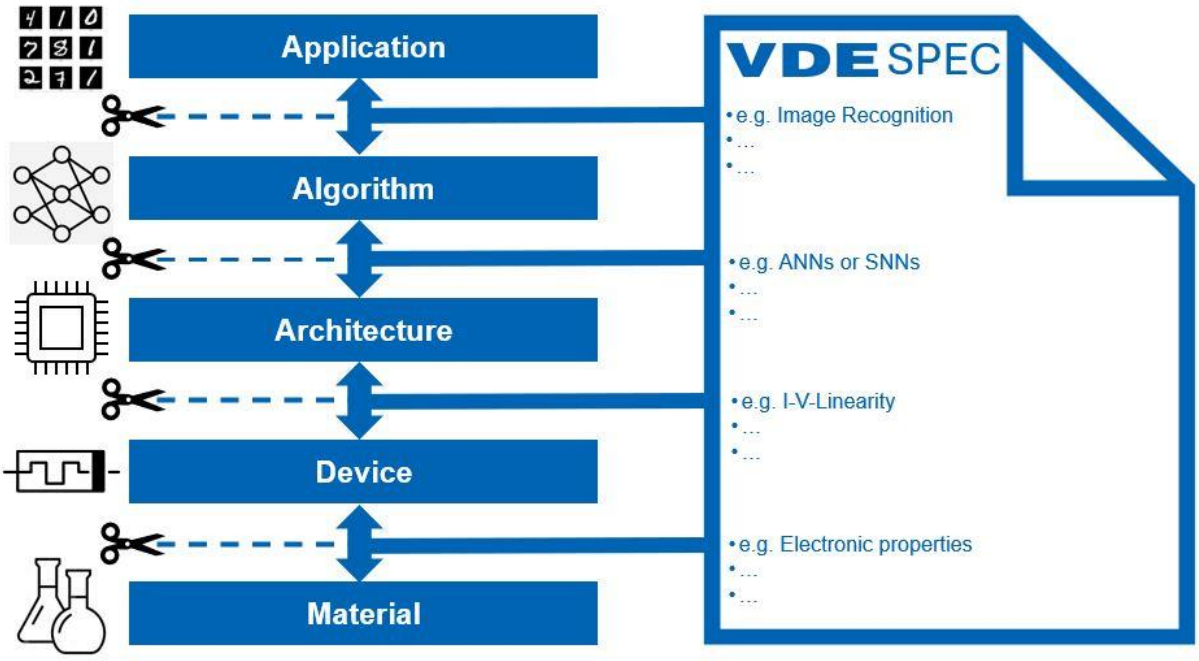

Bild 1 – Schichtenmodell mit den entsprechenden Schnittstellen

Ansatz des NMC-Schichtenmodells ist es, Sensor- und Informationsverarbeitung über unterschiedlichste technische Entwicklungen zu beschreiben und diese in die entsprechenden Schichten einzusortieren. Dabei wird gerade auf die Schnittstellen zwischen den Schichten ein besonderer Schwerpunkt gelegt, da an diesen Stellen die Übergabe selten klar definiert ist oder gar fehlt. Dies dient der Weiterentwicklung der gesamten NMC-Technologien und liefert Ansätze für definierte Anwendungen.

Im Schichtenmodell sind fünf übereinanderliegende Schichten mit ihren jeweiligen Schnittstellen definiert. Je nach Funktion sind in der gleichen Schicht mit klar definierten Schnittstellen die Bauelemente untereinander austauschbar.

Anmerkung: Die in dieser VDE SPEC enthaltenen Tabellen 1–4 sind aufgrund der allgemeinen, international genutzten Fachtermini in englischer Sprache belassen.

Während in der akademischen Entwicklung verschiedene Forschungsrichtungen angegangen werden, können lediglich die technologisch belastbaren und verifizierten Ansätze in Systeme überführt werden, die dann für den weiteren Entwicklungsprozess in Produkten und Dienstleistungen als Mehrwert zur Verfügung stehen. Daher ist diese VDE SPEC zum NMC eine erste Festlegung von Rahmenparametern, um diesen Transfer aus der Forschung für die Forschung und Anwendung bis hin zur Produktentwicklung auf einem strukturierten Weg zu vereinfachen. Der VDE in seiner Rolle als unabhängiger Technologieverband mit dem Schwerpunkt auf Elektrotechnik, Elektronik und Informationstechnik veröffentlicht diese abgestimmte VDE SPEC zu NMC und nimmt auf deren Grundlage den Aufbau einer Prüf- und Validierungsplattform auf. Die Validierungsplattform wird vorläufig auf einer Controller-basierten Platine (Device-under-test, DUT) mit einem Bussystem aufgebaut sein, bei der Adapter für integrierte NMC-Schalteinheiten vorrangig auf memristiven Bauelementen vorhanden sind (Bild 2). Für die Erfassung der Energieeffizienz ist ein Einsatz unter typischen Bedingungen notwendig, d. h. Testpattern werden auf dem Chip generiert und validiert. Zunächst geht man von X-Bar-Arrays aus arbeitet mit unterschiedlichen Pulsformen – nicht nur mit DC-Signalen.

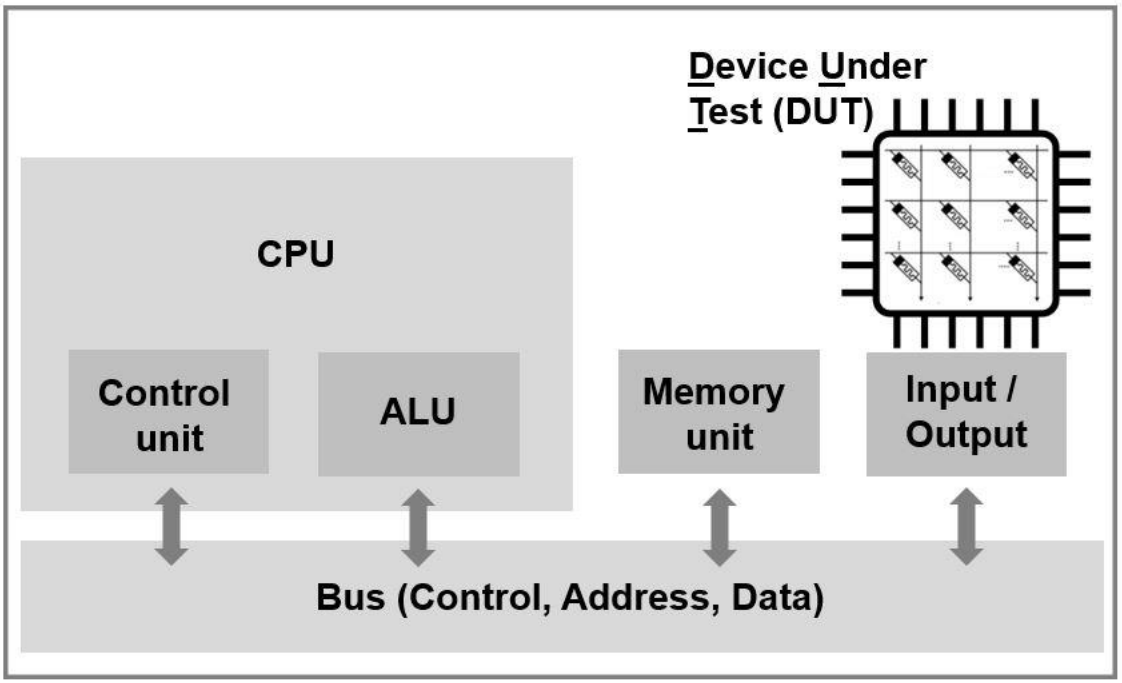

Bild 2 – Validierungsplattform mit Device-under-test (DUT) und Bussystem

Diese Einheit fungiert als austauschbares Modul, das an weitere Peripherie (Bild 3) angeschlossen werden kann, um Signalmessungen und die Validierung von Schaltzuständen durchzuführen, aber auch um Algorithmen und festgelegte Anwendungsbeispiele gegenüber konventionellen Systemen der Informationsverarbeitung auf Energieeffizienz und Benchmarking prüfen zu können.

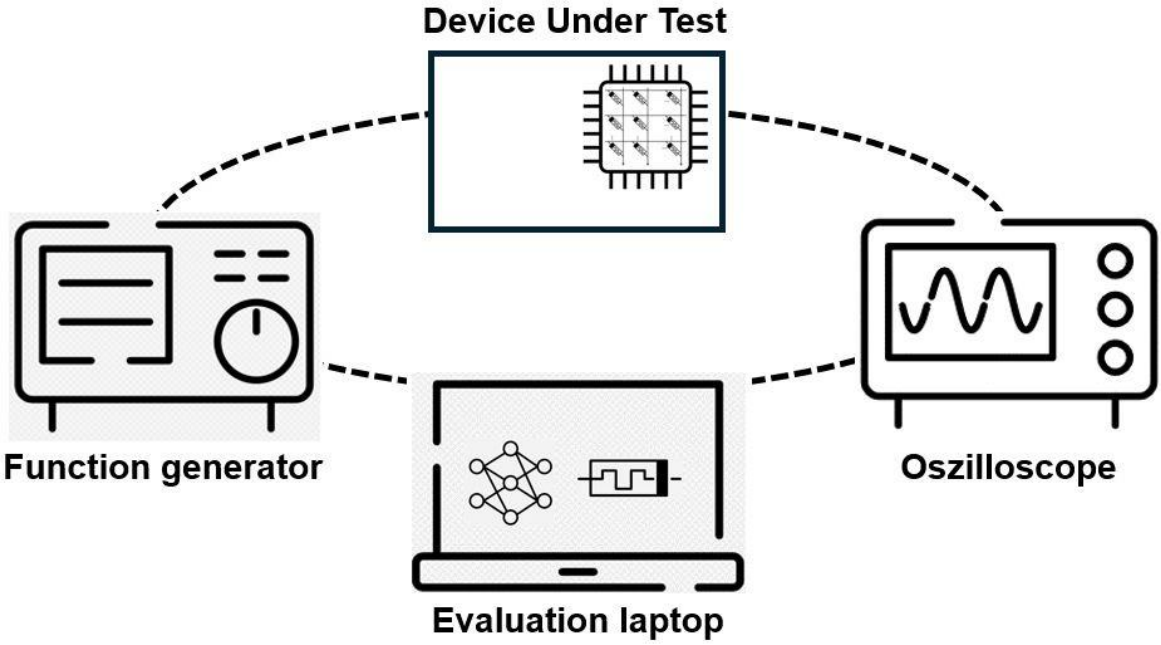

Bild 3 – Device-under-test (DUT) mit Peripheriegeräten

Für die Materialprüfung und Parametrisierung von Bauelementen wird vorweg auf externe Analytik zugegriffen. Obwohl die „on wafer“ und „on chip“ Messungen ein wesentlicher Bestandteil der weiteren Planung sind, werden diese zunächst zurückgestellt, da man sich zunächst auf die technologische Entwicklung mit dem hier beschriebenen Ansatz vorbereitet und sowohl den akademischen Bereich als auch die Industrie beim Erkenntnistransfer unterstützen möchte.

Inhalt

Abbildungsverzeichnis

Tabellenverzeichnis

1 Anwendungsbereich

Potenzielle Anwender neuromorpher Bauelemente (Devices) und Systeme sollten auf neutraler Basis die Vorteile und Leistungsfähigkeit verschiedener Konzepte überprüfen und bewerten können. Dies setzt einen zuvor abgestimmten Satz von Metriken voraus, der hier durch die vom VDE aufgesetzte Expertenkommission in Gemeinschaftsarbeit zugrunde gelegt wurde. Der Inhalt dieses Dokumentes ist ein wesentlicher Bestandteil dieser Vergleichbarkeit, um einen effizienteren Transfer in die Anwendung und Kommerzialisierung dieser Technologie zu erlauben.

Mit dieser VDE SPEC sollen Anforderungen und Kriterien für die Bewertung von Devices und Systemen im Kontext memristiv schaltender Devices und Systeme basierend auf neuromorphen Konzepten definiert werden.

Die Anwender dieser zukunftsweisenden Technologien profitieren in den unterschiedlichen technischen Ebenen, indem ein transparenter Vergleich für Ihre Entwicklungen zugrunde gelegt wird. Hiermit wird das Ziel verfolgt, Innovationszyklen effizienter und transparenter bewerten zu können. Die hier definierten Spezifikationen dienen im nächsten Schritt dem Aufbau eines Prüfzentrums und der Zertifizierung von Baugruppen. Hierdurch wird die Entwicklung innovativer Anwendungen wirkungsvoll unterstützt, indem ein Benchmarking auf Basis neutraler Prüfungen möglich ist.

2 Normative Verweisungen

Es gibt keine normativen Verweisungen in diesem Dokument.

3 Begriffe

Für die Anwendung dieses Dokuments gelten die folgenden Begriffe.

DIN und DKE stellen terminologische Datenbanken für die Verwendung in der Normung unter den folgenden Adressen bereit:

- DIN-TERMinologieportal: verfügbar unter https://www.din.de/go/din-term
- DKE-IEV: verfügbar unter https://www.dke.de/DKE-IEV

3.1
Algorithmus

Im Kontext künstlicher neuronaler Netze (KNN) sind Algorithmen abstrahierte Modelle von verbundenen künstlichen Neuronen, die biologischen lebenden Systemen nachempfunden sind. Diese werden dazu eingesetzt, um praktikable Lösungsansätze für komplexe Aufgaben aus verschiedenen Anwendungs-bereichen zu bieten.

3.2
Funktionseinheit

In der betrachteten Entwurfshierarchie ein monolithischer, physikalischer Block (z. B. ein Crossbar einschließlich Peripherie und Anbindung an ein Bussystem).

3.3
Neuromorph

altgriechisch: νεῦρον – *neuron*, Nerv und μορφή – *morphé*, Gestalt, Form

3.4
Neuromorphes Computing

Das Konzept des Neuromorphic Computing ist eine technische Nachbildung biologischer Systeme zur Informationsverarbeitung. Durch diese Ansätze verspricht man sich eine Steigerung der Rechenleistung und auch eine Steigerung der Energieeffizienz.

3.5
System

Im vorliegenden Kontext einer Entwurfshierarchie zum Hardware-/Software-Co-Design wird unter System das Zusammenspiel zwischen System-on-Chip, Sub-Systemen, Funktionseinheiten und Devices verstanden.

Weitere Definitionen und Abkürzungen finden sich im Anhang der einzelnen Tabellen.

4 Abkürzungen

Abkürzung	Bedeutung
ADC	Analog-to-Digital Converter (de: Analog-Digital-Wandler)
AFM	Atomic Force Microscopy (de: Rasterkraftmikroskopie)
CIM	Computing In-Memory
CMOS	Complementary Metal-Oxide-Semiconductor
DAC	Digital-to-Analog Converter (Digital-Analog-Wandler)
DDR	Double Data Rate
DUT	Device-under-test
HW	Hardware
ITG	Informationstechnische Gesellschaft (im VDE; www.vde.com/de/itg)
KI	Künstliche Intelligenz
KNN	Künstliche Neuronale Netze
MAC	Multiply-Accumulate (Operation)
ML	Maschinelles Lernen
MTJ	Magnetic tunnel junction
MVM	Matrix-Vector-Multiplikation
NMC	Neuromorphic Computing
NoC	Network-on-Chip
REM	Rasterelektronenmikroskopie (en: Scanning Electron Microscopy)
SoC	System-on-Chip
SPEC	Spezifikation (hier: VDE SPEC)
STM	Scanning Tunneling Microscopy (de: Rastertunnelmikroskopie)
SW	Software
VDE	Verband der Elektrotechnik, Elektronik und Informationstechnik e.V. (www.vde.de)
XRD	X-Ray Diffraction (de: Röntgenbeugung/-diffraktion)

5 Inhalt

Der inhaltliche Bereich der hier vorliegenden VDE SPEC gliedert sich, wie folgt.

In **Abschnitt I.I.** wird die Bedeutung des NMC bzw. neuromorpher Konzepte für aktuelle Anwendungen in der Informationstechnologie und hier insbesondere im Bereich der Künstlichen Intelligenz beleuchtet.

Abschnitt I.II. adressiert die Bedeutung eines Hardware-/Software-Co-Designs für die effiziente Entwicklung komplexer Architekturen und Systeme.

In den **Abschnitten 5.1. bis 5.5**. wird im Sinne einer ganzheitlichen Systembetrachtung der Bogen gespannt **von der Anwendung über Algorithmen, Architekturen und Devices bis hin zu den Materialien**. Ein wesentlicher Aspekt liegt dabei auf der Beschreibung der jeweiligen Schnittstellen zwischen zwei Schichten bzw. den Anforderungen an die technischen Baugruppen bzw. Systeme dieser Schichten.

■ I.I. Neuromorphes Computing und KI-Hardware

Mit dem NMC orientiert man sich an der energieeffizienten und leistungsfähigen Informationsverarbeitung in biologischen Systemen, und sieht Potenzial einer Alternative konventioneller CMOS-basierter von-Neumann-Architekturen.

Einen praktikablen Ansatz, um die Ziele des NMC zu erreichen, bieten sogenannte memristive Devices. Bei diesen handelt es sich vorrangig um elektronische Bauelemente mit variabler, widerstandsbasierter Speicherfunktion (engl.: memory, Speicher und resistor, Widerstand) und sie implizieren durch diese Wirkung die Speicherung einer Information aber auch die Verarbeitung dieser Information durch die flexible Anpassung weiterer Signaleingänge. Damit ist man näher an der Informationsverarbeitung in biologischen Systemen, beispielsweise am Neuron (Ziegler, 2020).

Hardware für das NMC soll in erster Linie nicht die konventionelle CMOS-Technologie für Digitalschaltungen ersetzen, sondern soll diese in den entsprechenden Anwendungsfeldern ergänzen. Damit sind Materialklassen für die Entwicklung von memristiven Bauelementen von besonderem Interesse, die sich in die existierende CMOS-Technologie einbinden lassen. In diesem Kontext sind Übergangsmetall-Oxide und – in jüngster Zeit – vermehrt Übergangsmetall-Dichalcogenide potentielle Materialien, aber auch Memristoren auf Basis von Phasenwechselmaterialien und ferroelektrische Materialien kommen in Betracht (Waser, 2019).

■ I.II. Hardware-Software-Co-Design

Die technische Prüfung von Systemen für das NMC benötigt adäquate Methoden und Messgeräte sowie die eindeutige Beschreibung einer Prüfprozedur (Leupers, 2024)**.** Ein Brückenschlag zwischen den Anforderungen in den Systemanwendungen, sowie der Schaltungsarchitektur und elektronischen Bauelementen mit den dazu zur Verfügung stehenden Materialien ist notwendig und dieser setzt eine Festlegung von Systemschichten mit deren Schnittstellen voraus. Als praktisches Utensil dient eine Validierungsplattform mit dem Device-under-test, bei der der integrierte Schaltkreis, die sogenannte „neuromorphe Hardware“, mit Ein- und Ausgängen verbunden wird. Durch diese flexible Prüfumgebung wird ein Software-Development ermöglicht, um Anwendungen der Künstlichen Intelligenz auf die neuartigen memristiven bauelementbasierten Schaltungen zu migrieren. Dies ist eine Variante der Evaluation und Validierung neuromorpher Systemarchitekturen sowie der dazugehörigen Software.

Aufgrund ihrer kompakten Bauweise lässt sich diese Mess- und Charakterisierungsplattform ohne aufwendige Aufbauten und spezielle Anforderungen nutzen. Besonders hervorzuheben ist die Flexibilität der Plattform: Sie bietet Ein- und Ausgänge, die dank einer eigens entwickelten Verbindungsmatrix variabel zugewiesen werden können. Dadurch ist die Plattform unabhängig von spezifischen Chip-Packages und Pin-Belegungen. Mittels dieser Platine ist man in der Lage, vollautomatisch alle notwendigen Eingangspulse zu generieren. Die eigens implementierten Transimpedanzverstärker ermöglichen es, Ströme mit hoher Genauigkeit zu messen. Für die korrekte Bewertung der Leistungsfähigkeit und Energieeffizienz der neuromorphen Schaltung müssen periphere Elemente, wie ADCs oder Transimpedanzverstärker separat vermessen und berücksichtigt werden.

Gesteuert wird die Plattform von einem Mikrocontroller, der eine Python-Schnittstelle beinhaltet. Diese ermöglicht eine einfache Implementierung der erforderlichen Messroutinen. Im Fokus steht dabei die Entwicklung entsprechender Software für das NMC, die durch die eigens implementierten Computing-in-Memory-Schnittstellen erleichtert wird. Dies ermöglicht es, entwickelte Mapping- und Scheduling-Algorithmen frühzeitig auf skalierbarer Halbleitertechnologie zu testen und zu verifizieren.

Für die Anwendung von Modellen und Methoden der KI auf Systemen für das NMC ist ein mehrstufiges Verfahren nötig.

- **In Stufe 1** wird ein KI-Produkt als Software geladen. Das bedeutet konkret, dass das Tensorflow- oder PyTorch-Modell eingelesen wird. Dann wird das Modell mit drei Funktionen angepasst: Erstens wird das Modell analysiert und z. B. durch geeignete Quantisierung an die Eigenschaften der integrierten Schaltung angepasst. Zweitens wird eine erste Schätzung der Leistungsfähigkeit der Anwendung (z. B. die zu erwartende Latenz) für die spätere Validierung angegeben. Drittens liegt das Modell in einem formalen Datenformat vor, sodass es vom Compiler in den nächsten Schritten weiterverarbeitet werden kann.

- **In Stufe 2** wird die eingelesene und ggf. quantisierte KI-Software (typischerweise das Modell eines konkreten neuronalen Netzes) vom Compiler verarbeitet. Hierbei wird das Modell durch unterschiedliche mathematische Abstraktionen betrachtet und für die Charakteristika des NMC optimiert. Ein konkretes Beispiel hierfür ist, dass das Modell als Netz von Neuronen betrachtet wird, die in Memristoren gespeichert sind; der Ort der Speicherung bestimmt, wie die Daten zwischen den Neuronen durch den integrierten Schaltkreis bewegt werden müssen.

- **In Stufe 3** erstellt der Compiler eine allgemeine Abstraktion des KI-Modells für das Computing-in-Memory-Konzept der integrierten Schaltkreise des NMC. Die Abstraktion wird dabei so gewählt, dass alle integrierten Schaltkreise adressiert werden können. Hierfür sind Speicheroperationen (LOAD/STORE) und Matrix-Vektor-Multiplikationen (MVM) exemplarisch. Auch ist die Synchronisation zwischen mehreren Rechenkernen nötig.

- **In Stufe 4** wird die allgemeine Computing-in-Memory-Abstraktion für einen konkreten integrierten Schaltkreis spezifiziert. Dazu werden die abstrakten Befehle (z. B. LOAD) in konkrete Befehle übersetzt (z. B. einen Bit-Code). Die Trennung von Stufe 3 und 4 ermöglicht es, einen „retargetable“ Compiler für verschiedene integrierten Schaltkreise zu erstellen.

- **In Stufe 5** wird der generierte Code für ein KI-Modell ausgeführt. Dies ist entweder auf dem integrierten Schaltkreis des Anwenders möglich, um dessen KI-Modell zu verifizieren, oder (früher im Entwurfszyklus) in einer hybriden Simulationsplattform inklusive der Testplattform. Letzteres ermöglicht es, Systeme zu verifizieren, bevor die skalierbare Technologieimplementierung erfolgt, oder verschiedene Architekturen von integrierten Schaltkreisen für eine serienreife Entwicklung des NMC präzise zu evaluieren.

5.1 Anwendungs-Schicht

Die Anwendungsräume für das NMC sind vielfältig und man sieht in dieser Technologie ein großes Potenzial, um die heutigen, konventionell zur Verfügung stehenden Technologien für Modelle der Künstlichen Intelligenz (KI) und des Maschinellen Lernens (ML) hinsichtlich Leistungsfähigkeit, Komplexität, aber auch einer geringeren Energieaufnahme bei größerem Mehrwert in der Informations-verarbeitung zu verbessern.

Ob sich diese Anforderungen umsetzen lassen, ist vor allem von einem transparenten Vergleich von unterschiedlichen technologischen Ansätzen abhängig, so auch von der Metrik und einer strukturierten Einteilung des technologischen Bedarfs in einzelne Schichten und Gruppen. Die Anwendungsgebiete spielen hier eine wesentliche Rolle in der Festlegung dieser Metriken, denn an ihnen richtet sich der Bedarf an Speicher, Logik und Datenverarbeitungskapazität aus. Dafür stehen heute unterschiedliche Ansätze aus Forschung und Entwicklung zur Verfügung, die jedoch nur schwer zusammengeführt werden können. Mithilfe dieser VDE SPEC möchten wir den Weg hierfür ebnen und geben exemplarisch einige wichtige Beispiele für Anwendungsfelder.

- **Edge Computing und Internet-of-Things (IoT)**

NMC eignet sich insbesondere für Anwendungen im Edge-Computing und auf IoT-Geräten, bei denen eine geringe Leistungsaufnahme durch den dezentralen Betrieb und die Energieversorgung gegeben ist. Dabei ist die Echtzeitverarbeitung von großer Bedeutung, damit die Informationsverarbeitung direkt auf dem SoC vorgenommen wird. Mit integrierten Schaltungen, basierend auf dem NMC, wäre eine Sensordatenverarbeitung, Mustererkennung und eine effiziente Entscheidungsfindung in einem definierten Kontext direkt „on the edge“ durchführbar.

■ Sensorik

Mithilfe von NMC können Sensordaten und sensorischen Verarbeitungssysteme effizienter aufgebaut werden, indem lernende Algorithmen mit der Sensordatenerfassung direkt in derselben Schaltung integriert werden. Dies erübrigt den Einsatz eines zusätzlichen Mikrocontrollers oder einer CPU. Zudem sind die Algorithmen an biologischen Mechanismen zur Wahrnehmung der sensorischen Daten ausgelegt, die eine Echtzeitdatenverarbeitung unter einer effizienten Ausnutzung der Primärenergie erlauben. Zudem ist der technologische Aufwand geringer, da auf eine komplexe von-Neumann Rechnerarchitektur verzichtet werden kann. Dies bietet die Möglichkeit der Skalierung zur multimodalen Datenerfassung über mehrere Kanäle und einer effizienten Verarbeitung, um Muster erkennen und relevante Informationen extrahieren zu können.

■ NMC-Plattformen

Mit zunehmender Reife der NMC-Technologie ist zu erwarten, dass spezialisierte Plattformen entstehen, die Hardware, Software und Entwicklungswerkzeuge für verschiedenste Einsatzbereiche in der Sensordatenverarbeitung und Prozessierung von großen Datenmengen – auch in Echtzeit – zur Verfügung stellen werden. Mit dem NMC werden von-Neumann Architekturen ergänzt, erweitert oder gar ersetzt. Dies führt zu mehr Flexibilität im Entwurf von informationsverarbeitenden Systemen, wie beispielsweise im Entwurf von integrierten Schaltungen für das „edge computing“. Diese NMC-Plattformen bieten die Möglichkeit, die vorhandenen Technologien zu skalieren und zu verifizieren. Zudem eröffnen Sie einen Experimentierraum in der akademischen und industriellen Forschung.

■ Verarbeitung von audiovisuellen Daten

Eine der vorrangigsten Anwendungen für das NMC ist Zeichen- und Bilderkennung mit den bisherigen Komplexitätsstufen der Schaltungstechnik. Es ist jedoch möglich, auch andere Anwendungsfelder einzubeziehen, die beispielsweise in der Medizintechnik liegen. Informationsverarbeitung von Audiosignalen für Hörhilfen und Hörimplantate wie das Cochlea-Implantat bieten einen guten Anwendungsbereich, da hier klare Leistungsparameter vorliegen, die Informationsverarbeitung klar definiert ist und die Energieaufnahme dezentral erfolgt. Obwohl sich andere Anwendungsgebiete auftun, wie die Sensordatenverarbeitung für das „autonome Fahren“, so sind die Entwicklungshorizonte hier noch zu weit, um sie in dieser VDE SPEC zum NMC einzubeziehen. Je nach Reifegrad und Entwicklung werden diese Anwendungsfelder sicherlich eine wichtige Rolle bei dem Einzug des NMC einnehmen, doch in dieser VDE SPEC widmen wir uns klar abschätzbaren Parameterräumen, mit einem überschaubaren Komplexitätsgrad.

Insgesamt hat NMC das Potenzial einer Technologie, die Innovation und Fortschritt in einem breiten Spektrum von Bereichen vorantreibt und neue Möglichkeiten zur Verbesserung der Effizienz, Autonomie und Intelligenz von Systemen und Geräten bietet. Da die F&E-Bemühungen weitergehen, ist davon auszugehen, dass NMC zunehmend in großvolumigen Anwendungen eingesetzt wird, die von seinen einzigartigen Fähigkeiten profitieren.

Schnittstellen im Schichtenmodell

Im Folgenden werden die Schichten mit den dazugehörigen Schnittstellen des NMC-Schichtenmodells erläutert.

In diesem Schichtenmodell werden die Funktionseinheiten definiert und kategorisiert sowie die Begrifflichkeiten festgelegt. Dies dient der strukturierten Übersicht und des Austausches zwischen den einzelnen Fachgebieten bzw. Disziplinen, die in den jeweiligen Schichten ihre Forschungs- und Entwicklungsarbeit vorantreiben. Diese Festlegung erlaubt eine reibungsfreie Kommunikation und ermöglicht die Weitergabe von Parametern, die angefordert werden für die Optimierung oder die Bereitstellung sowie Implementierung neuer Parameter in das Gesamtsystem.

Viele dieser Parameter, die ausgetauscht werden, werden in ein komplexes Gesamtsystem integriert, das nur auf den ersten Blick einfach erscheint. Jedoch wird eine Vielzahl von Details hinsichtlich Funktionalität, Zuverlässigkeit, Sicherheit und Effizienz erfüllt sein müssen. Die zu lösenden Aufgabenstellungen in den jeweiligen Schichten reichen vom elektronischen Verhalten der Materialien zur Signalführung in den Bauelementen und Schaltungen bis hin zur Informationsverarbeitung in Hard- und Software auf den höheren Abstraktionsebenen. Dazu ist eine geregelte Reihenfolge im Schichtenmodell des NMC vorgesehen, damit die ausgewählten Anwendungsfelder bearbeitet werden können. Im Nachgang werden hieraus Prüfungs- und Testprozedere im VDE abgeleitet.

Schnittstelle 1 <> 2: Anwendungs-Schicht <> Algorithmik-Schicht

Aus der vorangegangenen Diskussion ist klar abzuleiten, dass die Anwendungen ein wesentlicher Schwerpunkt für den Aufbau des Schichtenmodells sind. Mit ausgewählten und festgelegten Anwendungen lassen sich die Anforderungen einer höheren Abstraktionsebene auf die darunterliegenden Ebenen einer geringeren Abstraktion ableiten. Zweck dieser Strukturierung ist es, Ablauf und Übergabestellen zu definieren, die für alle Beteiligten, welche dieses System aufbauen klar sind und ein gemeinsames Verständnis festzulegen, bei dem die Parameter in den jeweiligen Schichten und in den Schnittstellen bekannt sind. Dies wird im Folgenden als Spezifikationsrahmen bezeichnet.

Spezifikationsrahmen
Spezifikation der auszuführenden Aufgabenstellung als Ableitung für den dann auszuführenden Algorithmus Zur Bewertung der Messergebnisse, werden ausgewählte Algorithmen verwendet, die auf der jeweiligen NMC-Hardware implementiert sind. Dazu wird ein geeigneter Datensatz für die Bewertung der Energieeffizienz und des Durchsatzes (MNIST, ggf. CIFAR-10) festgelegt. Der Fokus liegt hier auf der quantitativen Bewertung der neuartigen Systemkomponenten. Die Komplexität der Problemstellung ist zweitrangig, insofern diese keinen Einfluss auf die funktionalen Eigenschaften der Systemkomponente an sich und im Zusammenspiel mit anderen Systemkomponenten hat. Die Festlegung des Nutzungsprofils sollte ebenfalls berücksichtigt werden, um einen einheitlichen Ablauf der Algorithmen zu den Anwendungen sicherzustellen. Hiermit ist zum Beispiel eine Parameteraktualisierung 1x täglich über einen festgelegten Zeitraum gemeint. Es sind jedoch auch andere Szenarien denkbar, je nach Anforderungsprofil einer Anwendung. Die zum Einsatz kommende NMC-Hardware und das Product Validation and Testing (PVT) müssen Ausnahmefälle, sogenannte „worst-case corners“ entsprechend dem Anforderungsprofil umfassen.
Bewertung auf Systemebene Die quantitative Erfassung der Messdaten ist wesentlich, um die Ergebnisse durch feste Kenngrößen validieren zu können. Die jeweiligen Messdaten werden anhand des hier vorgestellten Schichtenmodells sowie der Schnittstellen eigeordnet und klassifiziert. Die Genauigkeit des Algorithmus ist bei Ausführung auf der jeweiligen NMC-Hardware festzulegen. Falls nicht deterministisch, so sollten Einheiten wie für N Durchläufe mit individuellen Mittelwerten versehen werden. Die Latenz hat eine signifikante Bedeutung hinsichtlich der Effizienzbewertung von NMC-Systemen. Somit ist die Messung von Zeitintervallen für den Prozess des ersten Datums an Input/Output (IO) von Hardware zu Ergebnis an IO wichtig und muss als Parameter erfasst werden. Der Energiebedarf ist einer der zentralen Optimierungsparameter. Daher ist die Festlegung der Messung dieser Größe von großer Bedeutung für die Effizienzmessung des individuellen Systems, aber auch im Vergleich mit anderen Ansätzen. Hierzu wird das Referenzsystem eines konventionellen von-Neumann-Rechners aufgesetzt und die Anwendung für den Durchlauf des Algorithmus festgelegt. Die Erfassung der Energieaufnahme könnte in Form eines Quotienten Durchlauf eines ersten Datums an IO mit festgelegter Annahme der energetischen Kosten für die IO-Kommunikation, wie beispielsweise Energie / IO-bit gesetzt werden. Der Flächenbedarf auf dem Chip (on-chip) wird als Vergleichsparameter hinzugezogen. Einheiten, wie: on-chip Speicher, Steuereinheit, Recheneinheiten (ALU), IO-Busverstärker oder eben auch integrierte Einheiten wie das Memristive-Crossbar-Array werden berücksichtigt. Der Vorteil der Integration wird ebenfalls während des Prüfbetriebes validiert, sodass sich Rückschlüsse auf die Bauform und den Technologieprozess bis hin zur Material-Schicht ziehen lassen. Die Mehrfachnutzung bei den Operationsschritten im Algorithmus ist ebenfalls möglich und wird beim Benchmarking berücksichtigt. Eine strukturierte Dokumentation der Ergebnisse ist Grundlage der Validierung und Prüfung der NMC-Systeme, wobei nach Modellen, Schaltungssimulationen, Datenbasis für ein virtuellen Tape-Out, Messung physikalischer Parameter der NMC-Hardware auch nach Aufteilung auf Komponenten gemessen wird. Toleranzräume sind ebenfalls fester Bestandteil der Festlegung der Prüfparameter. Zu definieren sind hier unter anderem: berücksichtigte Einbußen in der Ausbeute der Chipfläche auch als Yield benannt, tolerierbare „worst-case“-Ergebnisse unter Einbeziehung der Prozessfähigkeitsindizes, wie beispielsweise der 3-sigma als Standardabweichung zum Mittelwert bei nicht-deterministischer Ausführung.

Insbesondere für memristive und weitere neuartige Bauelemente sind pro-aktiv zusätzliche Maßnahmen zur

- Initialisierung des Algorithmenablaufes und Energieaufnahme für die Nutzung der NMC-Hardware
- Laden der Parameter, Trainingsläufe für die Crossbar-Arrays sowie andere Alternativen gemessen anhand von Latenz, Durchlaufeffizienz und Energiebedarf
- „Data Retention" Verhalten und Auffrischen von Daten in den Speicherzellen und Einheiten anhand der Prüfparameter vielfaches Schreiben, Auslesen, „read disturb", „soft errors" Fehlerkorrekturmethoden
- Belastungsprüfungen zur Fehlertoleranz und Alterung unter mehrfachem Ansprechen der Einheiten u.a. bei höheren Spannungen der Signalpulse
- Festlegung der erforderlichen Testmethodiken Flankensteilheit, Zeitintervalle wie „time-0", „in-the-field" abhängig von der zur prüfenden NMC-Hardware

Zum Einsatz kommender Algorithmus für die Validierung von NMC-Systemen

Grundsätzlich ist der Algorithmus zur Problemlösung frei wählbar, solange dieser die spezifizierten Mindestanforderungen der Anwendung erfüllt und damit eine strukturierte Aufgabenlösung möglich ist. Zur Nachvollziehbarkeit muss der Algorithmus einer Reproduktion zugänglich gemacht werden. Dies geschieht wie folgt:

- verwendeter Algorithmus (z. B. Netzwerkmodell einschließlich der Parameter) zur Lösung der Aufgabenstellung formuliert als Pseudocode
- insofern der Algorithmus spezifisch zur NMC-Hardware passt, müssen auch die Trainingsdaten im entsprechenden Format vorliegen
- Dokumentation des Trainingsalgorithmus basierend auf Trainings-, Validierungs-, Testdatensätzen sowie Hyperparametern
- Daten zur festgelegten Genauigkeit sollten mindestens n>10 unabhängige (emulierte) Eingaben und je n>10 Trainings umfassen

Ergänzung: Sicherheitsaspekte und offene Stellen im Algorithmus

Eine Bewertung von sicherheitsrelevanten Aspekten benötigt zunächst die Definition der betrachteten Angriffsvektoren. Darauf aufbauend müssen Algorithmus, Systemarchitektur und architekturelle Komponenten bzgl. ihrer Schwachstellen bewertet werden. Dies ist nicht Bestandteil dieser VDE SPEC, wird jedoch in einer Ergänzung berücksichtigt, in welcher der Schwerpunkt auf Sicherheitsaspekte gelegt wird.

Allgemeine, international genutzte Begrifflichkeiten in Bezug auf NMC-Systeme und Maschinelles Lernen bezogen auf Mustererkennung:

Tabelle 1 – Characteristics and Characteristic Expressions at the interface between Applications and Algorithms

Characteristic	Characteristic Expressions					
Applications	Classification	Prediction	Clustering	Association	…	
Neural Networks	ANN (SLP, MLP)	BNN	CNN	RNN (LSTM, …)	SNN	…
Datasets	MNIST	CIFAR-10	ImageNet	…		
Inference / Training	Inference	Training	Inference + Training	…		

Definitions and Abbreviations (alphabetical):

- *CIFAR-10* – the CIFAR-10 dataset (Canadian Institute For Advanced Research) is a collection of images that are commonly used to train machine learning and computer vision algorithms
- *Datasets* – integral part in the field of machine learning, consisting of texts, numerical data, audio, images, videos etc. for solving and analysing AI challenges
- *Inference* – applying the formerly trained capability to new data
- *MNIST* – the MNIST dataset (Modified National Institute of Standards and Technology database) is a large dataset of handwritten digits that is commonly used for training various image processing systems
- *Neural Networks (NN)* – different architectures of Artificial-NN (ANN) such as Single-Layer-Perceptron (SLP), Multi-Layer-Perceptron (MLP), Binarized-NN (BNN), Convolutional-NN (CNN), Recurrent-NN (RNN), Long-Short-Term-Memory-NN (LSTM), Spiking-NN (SNN), and others
- *Training* – learning a new capability from existing data

5.2 Algorithmik-Schicht

NMC erfordert die umfassende Berücksichtigung eines großen Entwurfsraumes. Hier werden häufig etablierte Datensätze verwendet, um quantitative Vergleiche bezüglich Genauigkeit bzw. Zielvorgaben zu ermöglichen. (Gemmeke, 2024)

Um die Forschung im Bereich der Entwicklung von neuartigen Bauelementen und Funktionseinheiten durch eine bessere, quantitative Vergleichbarkeit zu fördern, stehen Beispiele für trainierte neuronale Netze öffentlich zur Verfügung, wie unter https://lnkd.in/e-F9wce6 oder https://lnkd.in/ec-9Cicu, die mit binären Werten (+1 und -1) arbeiten und dabei Ergebnisse liefern, die denen von Fließkommaberechnungen für einfache Datensätze wie MNIST oder CIFAR-10 nahekommen. Das oben genannte Repository enthält Parameter, Python-Code sowie Details zum (Re-)Training dieser Netze.

In einem binären neuronalen Netz werden üblicherweise die verwendeten Fließkomma-Gewichte durch binäre Gewichte ersetzt. Dies optimiert Speicherplatz sowie Rechenaufwand zugleich und ist daher insbesondere für Geräte mit lokal begrenzten Ressourcen, sogenannte Systems-on-Chip (SoC), vorteilhaft. Die Verwendung binärer Gewichte bewirkt eine Beschleunigung, jedoch erreichen binäre neuronale Netze noch nicht die gleiche Genauigkeit wie ihre Pendants mit Fließkommagewichten in 32-bit Netzwerken.

Grundsätzlich ist und bleibt die Optimierung der verwendeten Zahldarstellung ein hochaktuelles Forschungsfeld – auch im Bereich komplexer Informationsdatennetze. Dieser Bereich ist von großer Relevanz für gesellschaftliche, wirtschaftliche sowie industrielle Anwendungen.

Schnittstelle 2 <> 3: Algorithmik-Schicht <> Architektur-Schicht

Spezifikationsrahmen
Der Zusammenhang zwischen Algorithmus und NMC-Hardware (HW) basiert auf einer Definition der HW-Architektur. Diese HW-Architektur besteht aus verschiedenen Komponenten (Abschnitt 5.3), auf die Instruktionen der Beschreibung des Algorithmus in Pseudo-Code angewendet werden. Neben Funktionseinheiten zur Datentransformation auf die entsprechende Recheneinheiten umfasst die Architektur Komponenten zur Datenspeicherung und für den Datenaustausch.

Allgemeine, international genutzte Begrifflichkeiten in Bezug auf NMC-Systeme und Maschinelles Lernen bezogen auf von-Neumann-Rechnereinheiten oder Funktionsblöcke eines Mikroprozessors.

Tabelle 2 – Characteristics and Characteristic Expressions at the interface between Algorithms and Architectures

Characteristic	**Characteristic Expressions**			
ADCs	on-Chip	off-Chip	…	
DACs	on-Chip	off-Chip	not implemented	…
Sensing Electronics	on-Chip	off-Chip	charge integration	…
Activation Functions	sigmoid	hyperbolic tangent	rectified linear unit	…

Definitions and Abbreviations (alphabetical):

- *Activation Functions* – so-called neuron activation functions – most common are: Sigmoid, Hyperbolic Tangent (tanh) and Rectified Linear Unit (ReLU). Activation functions can be realized either in SW or in HW
- *ADC* – system that converts an analog signal into a digital signal, e.g. converting an analog input voltage or current to a digital number representing the magnitude of the voltage or current
- *DAC* – system that converts a digital signal into an analog signal – performing the reverse function of an ADC
- *Sensing Electronics* – e.g. a transimpedance amplifier (TIA) for converting current to voltage or a charge-based accumulation circuit

5.3 Architektur-Schicht

Die HW-Architektur als Ganzes ist als Auswahl von Teilsystemen (System-on-Chip, Subsysteme, Funktionseinheiten) und deren Konnektivitäten zu sehen, die eine bestimmte Funktionalität bieten muss.

Sinnvoll ist eine Unterscheidung zwischen Funktionseinheiten und Subsystemen, wobei eine Funktionseinheit z. B. ein monolithischer, physischer Block, wie ein RAM-, Cache- oder ROM-Speicher sowie auch ein Crossbar-Array sein kann, der zusätzliche Subsysteme benötigt, um in einem SoC Verwendung zu finden. Solche zusätzlichen Subsysteme können beispielsweise die Steuereinheit (hier: Controller für die Programmierung des Crossbar-Arrays) oder die Schnittstellenlogik zum Netzwerk (hier: Network-on-Chip, NoC) sein.

Die HW-Architektur ist ein wesentliches Element im Chipentwurf und ihre Optimierung ist entscheidend für die effiziente Ausführung eines Algorithmus. Gleichzeitig ist durch Variation und Auslegung der Funktions-einheiten, wie beispielsweise Speichern, Prozessoren und NoCs ein Optimierungsraum gegeben, der anhand der o.g. Parameter (s. Schnittstelle 1<>2:) gemessen werden kann.

Ein gängiger Ansatz im Bereich künstlicher neuronaler Netze ist die Auslagerung von Daten, wie Gewichten, Eingangsdaten und Zwischenergebnissen, in einen Bereich außerhalb der betrachteten Architektur, also einen externen, peripheren Speicher. Um andererseits den relevanten Freiraum in der Optimierung der Chiparchitektur nicht unnötig durch die Festlegung von Konstanten und Parametern einzuschränken, ist eine Bewertung bzw. messtechnische Erfassung der Effizienz anhand der gelisteten Parameter auf Systemebene sinnvoll. Hier ist sichergestellt, dass der Datenaustausch von Eingang bis Ausgang nach Erfüllung der Aufgabenstellung berücksichtigt ist.

Die messtechnische Bewertung der Effizienz des Algorithmus allein wird als nicht zielführend gesehen. Grund dafür ist, wie bereits unter Abschnitt 5.1 ausgeführt, die Betrachtung des Gesamtsystems und der Aufgabenlösung, um dann die Qualität des erzielten Ergebnisses und die damit verbundene Energieaufnahme quantitativ zu prüfen. Die alleinige Prüfung auf Ebene einer einzelnen Rechenoperation auf dem System (hier z. B.: MAC-Operation), ist dabei nicht zielführend. Einerseits hat man in diversen Arbeiten gezeigt, dass für verschiedene Anwendungsfälle und Architekturen der Energiebedarf der einzelnen Rechenoperation vernachlässigt werden kann. Andererseits gibt es zahlreiche NMC-Ansätze, bei denen die MAC-Operation effizienter ausgeführt werden kann. Somit lassen sich Ergebnisse einzelner Rechenoperationen in der Schicht des Algorithmus nicht direkt miteinander vergleichen.

Grundlage für die weitere Entwicklung und Prüfung bietet die Festlegung einer Referenzimplementierung unter Vorgabe der Charakterisierung in einer konventionellen Logikfamilie der CMOS-Technologie. Dieses NMC-System ist die Ausgangsbasis als Referenzsystem, das Optimierungen und proprietäre Lösungen in allen Schichten des Modells zulässt. Eine Prüfung und Bewertung erfolgen dann automatisiert, mit gleichen Annahmen für alle Lösungsvorschläge, in der Schicht. Dieses NMC-HW-System ist das Referenzsystem, um ein einfaches SoC mit Prozessor, Speicher und Beschleuniger einer Prüfung und dem Benchmarking gegenüber konventionellen Systemen zu unterziehen. Wie in Abschnitt 5.2 bereits erwähnt, kommen nach Auswahl der festgelegten Anwendungsroutine ggf. noch periphere Funktionsblöcke, wie DDR, ADC etc. zum Gesamtsystem hinzu.

Um den Eigenschaften, d.h. den Parameterräumen der NMC-Funktionsblöcke Rechnung zu tragen, müssen die Bewertungsmethoden auf diese angepasst und parametrisiert werden. Beispiele sind Schaltfrequenz, Endurance und Stabilität der eingeschriebenen Werte unter Last, subsumiert unter „Zuverlässigkeit“ der verwendeten memristiven oder anderen Bauelemente. Hierbei ist zu unterscheiden, ob es sich um eine deterministische oder eine nicht-deterministischen Ausführung handelt.

Wie bereits aus diesem Kontext entnommen werden kann, ist die Optimierung eines einzelnen memristiven (oder anderen) Bauelementes nicht für die Messung, Prüfung und das Benchmarking ausschlaggebend. Damit würde ggf. lediglich die Charakterisierung und Modellierung eines Crossbar-Arrays als Funktionseinheit durchgeführt werden können. Die Bewertung auf Systemebene dagegen berücksichtigt die Modellierung in ihren festgelegten Parametern, einschließlich der Messung der erzielten Lösung mit dem entsprechenden Vergleich zum Referenzsystem. Die Anpassung und Modifikation der HW-Architektur erfordern damit einen größeren Aufwand, da hierzu ein Algorithmus auf diese neue Architektur angepasst werden muss.

Schnittstelle 3 <> 4: Architektur-Schicht <> Bauelement-Schicht

Spezifikationsrahmen

Die Definition dieser Schnittstelle zwischen den Schichten „Architektur“ und „Bauelemente“ ist in dem Modell wesentlich, da im Bereich der Bauelemente eine Vielzahl von Ansätzen in der Grundlagen-forschung vorliegen, die jeweils unterschiedliche technische Reifegrade erfüllen und daher nur selten in Schaltungen und danach folgend in HW-Architekturen implementiert werden.

Zudem ist für den Entwurf von integrierten Schaltungen der Parameterraum nicht fest definiert, sodass eine Implementierung beliebiger Bauelemente in beispielsweise memristiver Technologie nur selten praktisch umsetzbar ist. Daher ist die Festlegung definierter Parameter an dieser Schnittstelle maßgebend für den Entwurf von Schaltungen sowie die Integration der Peripherie (SoC, Subsysteme, Funktionseinheiten). Durch diese Festlegung wird der Transfer aus der Forschung in den Systementwurf ermöglicht und man stellt dadurch eine inhaltliche Kommunikation zwischen zwei Disziplinfeldern her.

Messgegenstand für die erfolgreiche Integration an dieser Schnittstelle sind je Funktionseinheit die Latenz für die Rechenoperation im dokumentierten HW-System, die Energieaufnahme und die Flächen-belegung der Schaltung auf dem Chip oder im Package.

Das Referenzsystem ist die konventionelle von-Neumann Rechnerarchitektur in CMOS-Technologie. Zugrunde gelegt werden Signalverläufe der binären booleschen Algebra. Sollte die Codierung in den zu prüfenden Architekturen abweichen, erfordert dies einen Abgleich und eine Erweiterung der hier vorliegenden Festlegung.

Da die Anforderungen an die Funktionseigenschaften der Bauelemente durch Parameter beschrieben werden und von der Charakteristik einer HW-Funktionseinheit abhängen, ist eine vollständige Auflistung relevanter Parameter nicht hinreichend erfüllbar. Eine Prüfung und Anpassung auf Systemschicht ist hier maßgebend und praktisch handhabbar. Methoden zur Fehlerkorrektur, Redundanz oder Kalibrierung kommen in der Systemschicht zum Einsatz und werden dort validiert und gemessen.

Durch den gemeinsamen Entwurf (HW/SW-Co-Design für das NMC) ergeben sich insbesondere in der Algorithmen-Schicht zusätzliche Freiheitsgrade, die durch neuartige Funktionalitäten memristiver Bauelemente bzw. weiterer möglicher Technologien, den Parameterraum bereichern. Um den möglichen Gewinn in der Energieeffizienz, Fehlerkorrektur, Latenz und Flächenbelegung durch diese neuen Bauelementetechnologien bewerten zu können, sind Festlegungen von Prüfparametern zwingend und daher Bestandteil dieser VDE SPEC. Ob es sich dann um lokale Optimierungen in der jeweiligen Schicht handelt oder gar schichtübergreifende, „globale“ Optimierungen des Gesamtsystems, ist abhängig vom Ansatz und dem Bezug zum Schichtenmodell. Für die „globale“ Optimierung und das Benchmarking im hier vorgestellten Schichtenmodell ist eine Modellierung über die Schichten *Anwendung* <> *Algorithmik* <> *Architektur* <> *Bauelement* <> *Material* erforderlich. Leitlinien für diese Modellierung in den Schichten Architektur und Bauelemente sind Parameter, die in die jeweilige Schicht bzw. die Schnittstelle eingeordnet werden und der übergeordneten Lösung der Aufgabenstellung (deterministisch und geeigneter Zahlendarstellung) aus der festgelegten Anwendung heraus genügen.

Zur Prüfung der NMC-HW-Architektur sind Modelle unter Berücksichtigung der Parameter der Bauelemente wichtig, die neben ihren I/U-Nichtlinearitäten sowie dem zeitlichen dynamischen Verhalten der Bauelemente auch deren Variabilität und Latenz abbilden. Entsprechende Modelle werden durch Ermittlung der Bauelementparameter festgelegt und erweitert. Diese Parameter sind in Tab. 3 zusammengefasst.

Die Messungen an den jeweiligen Bauelementen werden durch Spannungspulse mit vorgegebener Signalfolge durchgeführt, um so die Parameter zu ermitteln. Für diese Pulsmessungen wird ein einheitliches Messprotokoll definiert, damit Vergleichbarkeit zwischen verschiedenen Bauelement-gruppen gegeben ist. Diese, durch Messungen ermittelten, physikalischen Parameter sind Grundlage für die beschriebenen Modelle, die zum Entwurf der NMC-HW-Architektur verwendet werden. Toleranzräume sind dort genauso berücksichtigt, wie Unsicherheiten und das dynamische Verhalten durch transiente Analyse. Diese Modelle enthalten physikalische, aber auch abstrakte Bauelemente und Schaltungseigenschaften für die dann in der Architektur-Schicht (Abschnitt 5.3) Schaltungssimulationen zur Auslegung des Systems durchgeführt werden. In der Schicht der Bauelemente und der hier beschriebenen Schnittstelle ist das Abstraktionsniveau von den durch Messung ermittelten Parametern gering.

Messtechnische Ansätze zur Charakterisierung resistiver Speicherelemente (Nielen, 2023)

Die messtechnische Charakterisierung resistiver Speicherelemente, hier auch memristive Bauelemente, erfordert eine spezielle Analytik. Es wird zwischen der Charakterisierung von Einzelzellen und der von Speicherarrays/-matrizen unterschieden.

Hierbei bedürfen Einzelzellen, je nach Konfiguration, entweder zwei (hier ein resistiver Widerstand, 1R) oder bis zu vier (hier eine Kombination aus einem Transistor und einem resistiven Widerstand, 1T1R) Kontaktmöglichkeiten.

Die Charakterisierung von Zusammensetzungen dieser Einzelzellen (Crossbar-Arrays), wie sie für NMC- Anwendungen relevant sind, erfordert in der Schicht der Bauelemente entsprechend mehr Kontakte bzw. Anschlussmöglichkeiten (bisherige Obergrenze liegt bei 64 – 128 – State of the Art). Sowohl Einzelzellen als auch Speichermatrizen müssen – idealerweise automatisiert – kontaktiert werden. Die Messung erfolgt durch Positionierung von Kontaktnadeln mit Hilfe eines Mikroskops.

Zur Messung temperatur-abhängiger Charakteristiken wird ein ThermoChuck eingesetzt. Die Automatisierung über sog. Waferprober („on-wafer-Messung“) gewährleistet einen sicheren und zuverlässigen Messprozess.

Der hohe Grad an Flexibilität und Konfigurationsmöglichkeiten gewährleistet dem Prüfenden einen hohen Grad an Flexibilität, so dass beliebig viele Testszenarien für die Charakterisierung memristiver Bauelemente in hoher Auflösung vorgenommen werden können. Dazu dienen speziell dafür ausgelegte Matrix-Testsysteme, welche die Erfassung der genannten physikalischen Parameter erlauben. Von essentieller Bedeutung ist die Adaption der Verstärkerschaltungen an die Erfordernisse einer schnellen Strombegrenzung sowie einer breitbandigen Abdeckung von Spannungsbereichen bis zu Pegeln von hier 10 V, wie in den nachfolgend dargestellten Spezifikationen erläutert.

Charakterisierung von Einzelzellen

Für die Charakterisierung resistiver Einzelzellen sind hohe Bandbreiten, insbesondere für ultraschnelle Pulsmessungen hier bis 250 MHz Bandbreite mit schneller Strombegrenzung hier bis zu 30 ns Reaktionszeit, sowie Strommessungen hier Bandbreite bis zu 100 MHz, vorzusehen.

Die Messroutinen umfassen hierbei die Aspekte:

- Elektroforming, abgeleitet aus der Galvanik zur Beschreibung der Diffusionsprozesse im Festkörper
- Unipolares und bipolares Schalten (I/U-Kennlinien und Pulsformen)
- Retention, hier Parameter zur Langzeitstabilität eines eingespeicherten Datums
- Endurance, hier Parameter zur Beschreibung der Anzahl an Lösch-/Programmierzyklen, ohne Degradation des Datums
- Life-Time Acc Testing (in Kombination mit opt. ThermoChuck), Degradationstest unter Belastung und Stresstests
- Schaltkinetik, hier auch Bauelementdynamik und transiente Analyse
- Quantisierte Leitfähigkeiten (Histogramm), hier Parameter der Leitfähigkeit und Widerstandsverhalten

Charakterisierung von Speichermatrizen

Die Charakterisierung resistiver Speichermatrizen hier bis zu 32 x 32 Kanäle, für digitale, analoge und NMC-Anwendungen erfordert Bandbreiten bis zu 100 MS/s inkl. schneller Strombegrenzungen mit einer Reaktionszeit bis zu <100 ns und schneller Strommessung bis zu 100 MHz.

Die Messroutinen umfassen hierbei die Aspekte:

- Elektroforming
- Unipolares und bipolares Schalten (I/U-Kennlinien und Pulsformen)
- Schreib- und Lese-Methoden für analoge Spannungs-Gewichtung, Programmierung des Crossbar-Arrays
- zeitlich aufgelöste Signalfolgen wie hier „Spike-Current Integration“ für Crossbar-Arrays
- Computing-In-Memory (ACN, seq. Logik)
- Spike Timing Dependent Plasticity (STDP, vollständig variable PRE- und POST-synaptische Signalerzeugung)
- Short und Long Time Plasticity

Tabelle 3 – Characteristics and Characteristic Expressions at the interface between Architectures and Devices

Characteristic	Characteristic Expressions (Beyer, 2024)						
Crossbar Dimensions*	8x8	32x32	64x64	128x128	256x256	512x512	…
Cell Elements	1R (passive CB)	1D1R (active CB)	1T1R (active CB)	…			
CMOS Process Nodes	14 nm	55 nm	65 nm	90 nm	130 nm	150 nm	180 nm	…
Feature Size	0,5 µm x 0,5µm	1 µm x 1µm	…				
Forming Voltage	suitable for the CMOS node, e.g. 130nm/VForm <1.5 V						
Write Voltage	<1 V	<3 V	<5 V	<10 V	>10 V	…	
Write Time - Switching Time	<10 ns	<100 ns	>100 ns	…			
Read Time	<10 ns	>10ns	…				
Write Energy - VoltagexTime	<1 fJ/bit	<10 fJ/bit	<100 fJ/bit	<1 pJ/bit	<10 pJ/bit	>10 pJ/bit	…
Retention Time	<10 years (at 125°C)	>10 years (at 125°C)	…				
Number of States	1-10	10-100	100-1000	…			
I-V-Linearity	none	low	medium	high	…		
Endurance Cycles	10^4	…	10^{15}	…			
Variability	Device-to-Device (D2D) Variability	Cycle-to-Cycle (C2C) Variability	…				
on/off-Ratio	1	10	100	1000	10^4	…	

Definitions and Abbreviations (alphabetical):

- *1R* – single resristive switch in passive crossbar arrays, feature size $4F^2$, easy to fabricate, excellently applicable for 3D integration
- *1T1R* – combination of a transistor (1T) and a resistive switch (1R) in resistive random access memory (ReRAM), feature size 6 − $8F^2$, 3D integration difficult
- *CMOS Process Node* – refers to a specific semiconductor manufacturing process and its design rules
- **Crossbar Dimensions* – crossbar array consisting of n word lines (WL), m bit lines (BL) and n x m memory cells
- *Device Variability* – deviation of the main chosen parameters and figures of merit of the device(s)
- *Endurance Cycles* – ability of a memristive device to sustain a certain number of operational cycles before its memristive states become unstable and difficult to maintain
- *Feature Size* – is determined by the width of the smallest lines that can be patterned in a semiconductor fabrication proces
- *I-V-Linearity* – relationship between the current through an electronic device and the voltage across its terminals (⇒ I-V-characteristic of the device)
- *Number of States* – distinguishable conductance levels on each memristive device
- *on/off-Ratio* – ratio of the on-state and off-state current without any applied gate voltage. High on/off ratio means a low leakage current (= improved device performance)
- *Retention Time* – measures the duration for which memristive states can persist without significant degradation or relaxation
- *Write Time / Switching Time* – time taken by a switch to go from an ON state to an OFF state or vice versa

5.4 Bauelement-Schicht

Zur Nachbildung neurobiologischer Mechanismen der Informationsverarbeitung und -speicherung inner-halb memristiver Bauelemente ist das Verständnis quantenmechanischer und physikalischer Materialeigenschaften vonnöten. (Ziegler, 2024) Durch Skalierung einiger dieser Materialien ist es möglich, geometrische Dimensionen mitunter in der Größenordnung der Materiewellenlänge der Elektronen in den memristiven Bauelementen zu erreichen. Dies bedarf aktueller technologischer Herstellungsverfahren, Grenzflächen-untersuchungen bis in den atomaren Bereich hinein sowie einer angemessenen Messtechnik. In diesem Zusammenhang gliedern sich die Aktivitäten in Forschung, Entwicklung und Anwendung in die zwei folgenden, miteinander verbundenen, Sektionen.

Sektion 1 – Parametrisierung memristiver Bauelemente

Durch etablierte Verfahren sowie neu entwickelte Methoden der Materialanalytik und der Materialcharakterisierung wird ein grundlegendes Verständnis der chemischen und physikalischen Mechanismen memristiver Bauelemente erhalten. Hierbei ist anzumerken, dass bei den meisten memristiven Bauelementen die Funktionsweise auf makroskopischer Skala modellierbar ist, ein fundamentales Verständnis auf atomarer Skala jedoch nur bedingt verfügbar ist. Dies beruht insbesondere auf der Tatsache, dass es sich um komplexe, gekoppelte chemische und physikalische Prozesse handelt.

Elektronische Messungen am Einzelbauelement charakterisiert durch U/I- und C/V-Messungen sowie automatisierte Messungen auf „on wafer“-Ebene, temperaturabhängige und zeitabhängige Messungen, als auch Analytikverfahren zur Charakterisierung und strukturellen Untersuchung aktiver Schichten mit Hilfe üblicher materialwissenschaftlicher Verfahren wie AFM, STM, REM, XRD (Abk. siehe Abschnitt 4) sind eng miteinander zu verzahnen. Dies ist notwendig, um die Materialeigenschaften nachzuvollziehen und diese dann in die entsprechenden Technologieprozesse einfließen zu lassen, sodass eine Optimierung der memristiven Bauelemente für die hier beschriebenen nachfolgenden Schichten des NMC-Schichtenmodells erfolgen kann. Diese Parametermessungen werden kategorisiert und dienen der Erweiterung einer fundierten Statistik für die erwähnte Optimierung der Bauelemente. Dieses Vorgehen schafft die Grundlage für eine weitergehende Forschung über die anderen Schichten des NMC-Modells hinweg und bietet die Möglichkeit eines Transfers in die technologisch industrielle Umgebung.

Sektion 2 – Modellierung und Simulation memristiver Bauelemente

Für den zielgerichteten, von der Anwendungsschicht abhängigen, Entwurf der Bauelementmaterialien ist eine physikalische Beschreibung gemäß Sektion 1 (s.o.) notwendig.

Modelle reichen dann von äquivalenten Ersatzschaltkreisen für die entsprechenden Bauelemente über datenbasierte Modelle, bis hin zu detaillierten analytisch physikalischen Beschreibungen basierend auf Drift-Diffusionsgleichungen der Materialübergänge in den jeweiligen memristiven Bauelementen.

Es werden an dieser Stelle Einzelzellen (1R aber auch 1T1R) berücksichtigt, sodass deren Parameter an höhere Schichten übergeben werden können. Die hier festgelegten Begrifflichkeiten sollen zusammen mit diesen Parametern einen hinreichenden Austausch mit den nachfolgenden Schichten sicherstellen und damit eine Optimierung hinsichtlich der Anforderungen aus der Anwendungsschicht erreichen.

Schnittstelle 4 <> 5: Bauelement-Schicht <> Material-Schicht

Spezifikationsrahmen
Anhand dieser Spezifikationen werden einheitliche Parameter zum Austausch zwischen den Schichten Bauelemente und Materialien festgelegt. Diese dienen dem einheitlichen Austausch innerhalb des gesamten Schichtenmodells zur Optimierung, aber auch der Vergleichbarkeit von Forschungs-ergebnissen und der darauffolgenden Forschung. Die Anforderungen an die Materialien und Bauelemente resultieren aus den darüberliegenden Schichten unter der Maßgabe, dass die anvisierte Anwendung erfüllt wird. In der hier beschriebenen Schnittstelle zwischen Materialsystemen und NMC-Bauelementen ist der technische Reifegrad ausschlaggebend für die weitere Integration und Verwendung von Materialien, die in dem CMOS-Fabrikationsprozess zur Verfügung stehen. Daher sollten Forschung und Entwicklung von Bauelementen in der industriellen Umgebung folgenden Parametern im Rahmen des Schichtenmodells dieser VDE SPEC genügen: ■ Forming Voltage – Spannungspuls zur Ausbildung des Speicherzustandes ■ Write Voltage – Spannungspuls zum Schreiben des Speicherzustandes ■ Write Time – Switching Time – Zeitintervall für den Schreibzyklus und allgemein für das Schalten der memristiven Einzelzelle ■ Read Time – Zeitintervall für den Lesezyklus und das Auslesen der memristiven Einzelzelle ■ Write Energy – Voltage x Time – Energieeintrag zum Schreiben der memristiven Einzelzelle ■ Retention Time – Parameter zur Langzeitstabilität eines eingespeicherten Datums ■ Number of States – Festlegung der Speicherzustände der memristiven Einzelzelle ■ Current Limitation Value – Festlegung der Strombegrenzung der memristiven Einzelzelle Dazu müssen Materialparameter gemäß u. g. Tabelle kategorisiert werden. Eine Validierung der Materialauswahl nach Reifegrad, Reproduzierbarkeit und Skalierung ist entscheidend für die Auslegung und Integration der memristiven Bauelemente.

Die Anpassung der hier vorliegenden Parameter obliegt dem Stand der Forschung, denn neuartige Bauelementekonzepte mit den dazugehörigen Materialien können u. U. nicht durch die Definition dieser VDE SPEC abgedeckt werden und müssen in einer aktualisierten Fassung angepasst werden.

Tabelle 4 – Characteristics and Characteristic Expressions at the interface between Devices and Materials/Mechanisms

Characteristic	**Characteristic Expressions**							
Crystallinity	crystalline	poly-crystalline	amorph	…				
Morphology	Shape	Size	Structure	…				
Electronic Properties	Resistivity (10^{-6}-10^{-7} Ωm)	Band Gap (1–3 eV)	Resistance Change (R_{on}/R_{off} >10)					
	Mobilities Electrons (700-8500 cm²/Vs) and Ions	…						
Mechanical Properties	Strength	Stiffness	Elasticity	Plasticity	Ductility	Toughness		
	Brittleness / Malleability	Resilience	Hardness	…				
Photonic Properties	Reflection	Adsorption	Transmission	Refraction	…			
Magnetic Properties	Type of magnetic Material	Saturation Magnetization	Magnetic Anisotropy	…				
Thermal Properties	Thermal Conductivity (0.5–5 W/Kcm)	…						
Technolog. Properties	Process Temperature	CMOS Compatibility	…					
Switching Mechanism/ Devices	ECM	PCM	VCM	ReRAM	FeRAM	FeFET	STT-MRAM	…

Definitions and Abbreviations (alphabetical):

- *ECM* – Electrochemical Metallization; *PCM* – Phase Change Memory; *VCM* – Valence Change Memory; *ReRAM* - Resistive Random-Access Memory; *FeRAM* – Ferroelectric Random-Access Memory; *FeFET* – Ferroelectric Field-Effect Transistor; *STT-MRAM* – Spin-Transfer Torque Magnetic Random-Access Memory

5.5 Material-Schicht

Ausgewählte Materialverbünde von Ferroelektrika, Metalloxiden, Chalkogeniden, 2D van-der-Waals-Materialien oder organischen Materialien für memristive Bauelemente, die ihre Leitfähigkeit in Abhängigkeit der elektrischen Vorspannung ändern, sind vielversprechende Kandidaten für eine energieeffiziente Informationsverarbeitung. Sie werden in der Forschung als künstliche Synapsen und Neuronen in neuromorphen Schaltungen verwendet. Dies ist ein neuer Ansatz im Vergleich zu konventionellen, binären, nichtflüchtigen Speicherzellen, in denen memristive Bauelemente bereits Verwendung finden. (Ziegler, 2024) (Wiefels, 2024)

In der weiteren Entwicklung von Speicherbauelementen sind zwei technische Faktoren ausschlaggebend:

- die **Skalierung** der Materialien von der sub-Nanometerskala hin zu makroskopischen Größen eines 300 mm Wafers und
- die **Kompatibilität** in bestehenden kontaminationsfreien CMOS-Technologien.

Die Materialien für memristive Bauelementen müssen in einer definierten Schichtdicke und Rauigkeit abstimm- und reproduzierbar sein und den Anforderungen an die Schaltdynamik – hier der transienten Analyse – genügen, sodass auf allen Abstraktionsebenen ein physikalisches Schalt- und Auslesever-halten gesichert ist. Die Vielzahl an unterschiedlichen physikalischen Phänomenen der memristiven Schaltdynamik werden in elektronische Effekte, ionische Effekte sowie strukturelle oder magnetische bzw. ferroelektrische Effekte eingruppiert.

Einige der Materialien für memristive Bauelemente haben mehrere vorteilhafte Eigenschaften, wie beispielsweise eine schnelle Zugriffszeit in der Größenordnung von einigen zehn Picosekunden. Diese

sind um Größenordnungen höher als bei konventionellen nichtflüchtigen Speicherzellen, wie Flash-Speicher. Auch die Retention kann durch die Wahl geeigneter Materialien modifiziert werden.

Neben den vorgestellten Vorteilen stehen jedoch noch einige Herausforderungen an. Eine davon ist die breite Verteilung von Speicherzuständen innerhalb der Einzelzelle. Diese Verteilungen sind von Nachteil bei der Auswertung logischer Operationen. Die zentrale Herausforderung im Zusammenhang mit speicherinternen Logikoperationen ist die begrenzte Präzision, die durch Signalrauschen und Leitwertdrift entsteht. Auch temperaturbedingte Leitwertschwankungen können ein Problem darstellen. Eine weitere Herausforderung ist die stöchiometrische Stabilität während des Schreibimpulses, bei dem Ionenmigrationseffekte auftreten können.

Die Integration memristiver Bauelementen in etablierte CMOS-Techologieprozesse ist ein wesentlicher Schwerpunkt dieses Abschnittes, wobei die Materialauswahl nicht nur darauf beschränkt ist. Abhängig davon, ob das Abschnitt 5.3 beschriebene Crossbar-Array auf dem Chip, auf dem IC, im Package (Aufbau und Verbindungstechnik) oder als periphere Funktionseinheit an den IO-Bus angeschlossen wird, ergeben sich unterschiedliche Anforderungen an die hier festgelegten Parameter.

Hochintegration

Wie bereits in dem Abschnitt zur Schicht „Bauelemente“ (Abschnitt 5.4) erwähnt, sind zwei wesentliche Ziele für die Optimierung von adäquaten Materialien für den Einsatz in elektronischen Bauelementen und Funktionseinheiten in x-bar-arrays für das NMC die **Hochintegration** und die **Steigerung der Energieeffizienz**. Für STT-MRAM wurde die Skalierung auf 11 nm Zellen sowie die Realisierung von 2 Mbit eingebettetem MRAM in 14 nm FinFET CMOS demonstriert. Aufgrund der geringeren Widerstandsauflösung ist das Auslesen von magnetischen Tunnelübergängen (MTJs) jedoch technisch schwieriger zu kontrollieren. Dennoch wurde kürzlich ein 64 x 64 MTJ-Array in 28nm CMOS-Technologie hergestellt. Um das Widerstandsverhältnis von MTJs in Zukunft zu erhöhen, sind Fortschritte auf der Materialseite erforderlich.

Eine Herausforderung ist es ferroelektrische Bauelemente auf Basis von HfO_2 zu skalieren, da die Dicke des Materials, um 3D Kapazitäten mit einem 10 nm Knoten zu ermöglichen und eine einheitliche Polarisierung auf der Nanoskala eines Materials zu erreichen, derzeit noch nicht reproduzierbar gelingt. Grund dafür ist, dass sich verschiedenen Phasen in dem Material einstellen. Daher geht der derzeitige Weg in Richtung ultradünner Schichten mit der reinen ferroelektrischen orthorhombischen Phase und ohne tote Schichten an den Grenzflächen, um sich dem Bereich unter 20 nm für ferroelektrische Bauelemente auf HfO_2 zu nähern.

Phasenwechsel-Bauelemente (PCM Devices) können im Bereich unter 10 nm hergestellt werden. Der begrenzende Faktor für CMOS-integrierte PCM Devices ist der hohe RESET-Strom, der für die Implementierung größerer Zugangstransistoren erforderlich ist. Kommerziell erhältliche ReRAM-Zellen mit konventionellen Geometrien wurden in 28 nm CMOS-Technologie kointegriert. Durch den Einsatz einer Seitenwandtechnik und nanodünner Pt-Elektroden wurden kleine Arrays mit 1 nm × 3 nm großen HfO_2-Zellen und 3 × 3 Arrays aus Pt/HfO_2 /TiO_x /Pt-Zellen mit einer Strukturgröße von 2 nm bzw. einem halben Pitch von 6 nm hergestellt.

Im Hinblick auf die ultimative Skalierung könnte der Verlust von Sauerstoff an die Umgebung die Retentionszeiten für ReRAM Bauelemente mit einer Skalierung von unter 10 nm einschränken. Filamente in einer Größe von 1 – 2 nm können jedoch stabil sein, wenn sie durch strukturelle Defekte wie Korngrenzen oder Versetzungen stabilisiert werden. Daher könnte die Suche nach einer Materiallösung für die Begrenzung von Sauerstoffleerstellen auf der Nanoskala die erforderliche Retention für Bauelemente im Größenbereich von wenigen nm gewährleisten.

Zugriffszeiten

Neben der Steigerung der Zugriffszeit ist die zunehmende Parallelisierung ein wichtiger Optimierungsfaktor, dabei geht es nicht vorranging um immer höhere Taktfrequenzen. Dennoch ist es sinnvoll, die ultimativen Geschwindigkeitsgrenzen von NVM-Konzepten festzuhalten, um die maximalen Lernraten abzuschätzen und die Auswirkungen kurzer Spiking-Stimulationen zu untersuchen. Darüber hinaus könnten neuartige Rechnerkonzepte, wie in der Schicht *Algorithmen* (Abschnitt 5.2) erörtert, von höheren Zugriffszeiten profitieren. Für MRAM wurde ein zuverlässiges Schalten mit 250 ps durch den Einsatz von MTJ mit doppeltem Spin-Torque nachgewiesen, die aus zwei Referenzschichten, einer Tunnelbarriere und einem nichtmagnetischen Abstandshalter bestehen.

Speicher, die auf FeRAM basieren, können erfolgreich mit 14 ns bei 2,5 V schalten. Ferroelektrische Feldeffekttransistoren (FeFETs) schalten nachweislich mit Pulsen < 50 ns in 1 Mbit Speicher-Arrays. PCM-Devices können mit Impulsen < 10 ns geschaltet werden. Im Allgemeinen wird ihre

Geschwindigkeit durch die Kristallisationszeit des Materials begrenzt. An Ge_x-Sn_y-Te-Proben wurde beispielhaft gezeigt, dass diese Zeit durch Anpassung der Materialzusammensetzung in einem weiten Bereich von 25 ns bis zu 10 ms eingestellt werden kann. Damit besteht ein hohes Potenzial, die Betriebszeit eines NC-Systems an die jeweilige Anwendung anzupassen. Für VCM ReRAM wurden SET- und RESET-Schaltungen mit 50 ps und 400 ps demonstriert. Beide sind bisher eher durch extrinsische Effekte und Bauelementfehler als durch intrinsische, physikalische geschwindigkeitsbegrenzende Effekte begrenzt.

Beständigkeit der Zustände (Endurance)

Alle memristive Speicherbauelemente haben eine begrenzte Lebensdauer, da die Speicherung auf der Bewegung oder Verschiebung von Atomen beruht, eben wie ReRAM, PCM und ferroelektrische Effekte in diesen Bauelementen. Bei FeFETs auf Siliziumbasis liegt die Lebensdauer in der Regel in der Größenordnung von 10^5, was hauptsächlich durch einen dielektrischen Durchbruch im SiO_2 an der Si-HfO_2-Grenzfläche bedingt ist. Was die Lebensdauer von VCM ReRAM betrifft, so wurde mit überzeugenden Statistiken nachgewiesen, dass $>10^6$ Zyklen realistisch sind.

In einigen Berichten wird von maximalen Zyklenzahlen von mehr als 10^{10} Zyklen ausgegangen. Je nach Materialsystem werden verschiedene Ausfallmechanismen für die Dauerhaftigkeit diskutiert. Die Mikrostruktur des Schaltmaterials könnte sich verschlechtern oder irreversibel von Metallatomen durchdrungen werden. In VCM ReRAM wurde eine übermäßige Erzeugung von Sauerstoffleerstellen als Faktor diskutiert, der die endurance limitiert. Neuartige Materiallösungen, welche die Ionen auf den vorgesehenen Aktionsradius beschränken, könnten ein Weg sein, um die Lebensdauer von ReRAM-Bauteilen zu erhöhen. Für PCM wurde vorgeschlagen, Multi-PCM-Synapsen zu implementieren. Die Verteilung über mehrere Speicherelemente könnte sowohl die Endurance- als auch die Variabilitätsprobleme umgehen.

Typische Einschränkungen im Hinblick auf einen zuverlässigen Betrieb ferroelektrischer Speicher sind der sog. „Wake-up-Effekt“, der nach einigen Zyklen eine zunehmende Polarisation bewirkt, und die Ermüdung, die bei hohen Zyklenzahlen zu einer Abnahme der Polarisation führt. Beides wird durch die Bewegung von Defekten wie Sauerstofflücken hervorgerufen und muss in Zukunft durch intensive Materialforschung auf diesem Gebiet angegangen werden.

Dauerspeicherung (Retention)

Nach dem Training muss der Zustand der nichtflüchtigen Speichersynapse bei einer Betriebstemperatur von 85°C für 10 Jahre stabil sein. Diese Anforderungen sind jedoch stark von der Anwendungsumgebung des NMC-Systems abhängig. Aus thermodynamischer Sicht könnten die Zustände in ferroelektrischen oder ferromagnetischen Speichern beide stabil sein. Im Gegensatz dazu speichern ReRAM- und PCM-Bauelemente Informationen in Form von Atom-Konfigurationen, wobei sowohl LRS als auch HRS metastabile Zustände sind und die Speicherung durch Materialparameter wie den Diffusionskoeffizienten der jeweiligen Spezies bestimmt wird. Hier ist die Degradation kein digitales Hin-und-her-Flippen von Zuständen, sondern ein allmählicher Prozess. Bei PCM wird die Drift des Widerstandszustands durch die strukturelle Entspannung der schmelzgehärteten amorphen Phase verursacht.

Abgesehen von einem Verwischen oder einer Drift des Zustandes wird bei ReRAM typischerweise eine Verbreiterung der Verteilung des programmierten Zustands (z. B. des Widerstands) beobachtet. Da die analoge oder mehrstufige Programmierung für NC von großer Bedeutung ist, sollte außerdem berücksichtigt werden, dass Zwischenwiderstandszustände im Vergleich zu den Grenzfällen von hoch- und niederohmigen Zuständen, wie sie für PCM-Bauelemente nachgewiesen wurden, eine geringere Retention aufweisen könnten.

Auslesestörungen

Während der Inferenz ist ein häufiges Lesen der Speicherelemente erforderlich, das den gelernten Zustand nicht verändern sollte. Bei einem bipolaren ReRAM-Speicher tritt eine Lesestörung im HRS/LRS hauptsächlich beim Lesen mit SET/RESET-Polarität auf, da die Lesestörung als Extrapolation der SET/RESET-Kinetik auf niedrigere Spannungen angesehen werden kann. Dennoch hat sich der HRS-Zustand in bipolaren filamentären VCM durch Extrapolation über Jahre hinweg bei Lesespannungen bis zu 350 mV als stabil erwiesen.

Variabilität / Anpassungen der Zustände

Die Variabilität ist bei Systemen, die auf der stochastischen Bewegung und Umverteilung von Atomen beruhen, wie ReRAM und PCM, besonders ausgeprägt. Hier ist die Variabilität von Bauelement zu Bauelement (D2D), von Zyklus zu Zyklus (C2C) und sogar von einem Lesevorgang zum nächsten (R2R) zu unterscheiden. Durch Optimierung der Fertigungsprozesse kann die D2D-Variabilität vergleichsweise geringgehalten werden. Im Gegensatz dazu kann die C2C-Variabilität bei filamentären resistiven ReAM und PCM aufgrund der Zufälligkeit des Filament- bzw. Kristallwachstums erheblich sein. Mit intelligenten Programmieralgorithmen kann die C2C-Variabilität jedoch sehr gut auf ein Minimum reduziert werden.

R2R-Schwankungen bleiben jedoch in Form von Ausleserauschen in filamentären VCM bestehen. Sie werden in der Regel auf die Aktivierung und Deaktivierung von Fallen oder die zufällige Umverteilung von Defekten zurückgeführt und hängen stark vom Material ab. Bei PCM werden die R2R-Schwankungen durch 1/f-Rauschen und temperaturbedingte Widerstandsschwankungen verursacht. Ein Ansatz zur Lösung dieser Probleme sowie der Drift ist die Verwendung des so genannten projizierten Phasenwechsel-Speichers mit einem nicht isolierenden Projektionssegment parallel zum PCM-Segment.

Obwohl die Variabilität eine Herausforderung für Speicheranwendungen darstellt, ist es möglich, NMC-Systeme so zu gestalten, dass sie diese ausnutzen.

Für die meisten in der Schicht *Algorithmen* (Abschnitt 5.2) beschriebenen Computerkonzepte ist der Betrieb mit binären Speichergeräten stark eingeschränkt, und die Möglichkeit, mehrere Zustände einzustellen, eröffnet einen neuen Komplexitätsraum. Bei Bauelementen mit thermodynamisch stabilen Zuständen wie ferroelektrischen oder magnetischen Speichern hängen die Zwischenzustände von der Anwesenheit von Domänen ab. Infolgedessen hängt die Leistung stark von der spezifischen Domänenstruktur ab, und die Skalierung kann durch die Größe der Domänen begrenzt sein. Nichtsdestotrotz wurde sowohl für FTJ als auch für FeRAM das Schalten auf mehreren Ebenen nachgewiesen.

Obwohl es keine direkte Verknüpfung zwischen der Anwendungsschicht (Abschnitt 5.1) im vorgestellten Modell gibt, sind die Parameter in den Schnittstellen dennoch ausschlaggebend für die Funktionsweise und Erfüllung der Vorgaben durch die spezifizierte und festgelegte Anwendung. Somit hat die Auswahl und Verwendung eine implizite Wirkung auf die Auslegung des NMC-Gesamtsystems. Dies ist das übergeordnete Ziel dieser VDE SPEC.

6 Literatur- und Quellenhinweise

Beyer, S. (2024), Wenger, C., Ziegler, M. *Informationen von GlobalFoundries, IHP (Frankfurt/O.) und Universität Kiel.*

Dhar, P. (2020). *The carbon impact of artificial intelligence. Nat. Mach. Intell., 2(8), 423-425.*

Gemmeke, T. (2024). *Information des Lehrstuhls für Integrierte digitale Systeme und Schaltungsentwurf (IDS) der RWTH Aachen.*

Jones, N. (2018). *How to stop data centres from gobbling up the world's electricity. Nature, 561(7722), 163-167.*

Leupers, R. (2024). *Information des Instituts für Kommunikationstechnologien und eingebettete Systeme (ICE) der RWTH Aachen.*

Nielen, L. (2023). *aixACCT Systems, Datenblatt RS/FE-Memory Analyzer - Vollintegriertes, modulares und halbautomatisches Charakterisierungssystem für resistive und ferroelektrische Speicher.*

Waser, R. (2019), Dittmann, R., Menzel, S., & Noll, T. *Introduction to new memory paradigms: memristive phenomena and neuromorphic applications. Faraday discussions, 213, 11-27.*

Wiefels, S. D. (2024), Dittmann, R.. *Materials challenges and perspectives, in: Roadmap to Neuromorphic Computing with Emerging Technologies, Page 40, arXiv:2407.02353.*

Ziegler, M. (2020). *Novel hardware and concepts for unconventional computing. Scientific reports, 10(1), 1-3.*

Ziegler, M. (2024). *Information des Lehrstuhls Energy Materials and Devices Department of Materials Science, Kiel University.*

7 Gremien

DKE/K 631: Halbleiterbauelemente

DKE/K 682: Aufbau- und Verbindungstechnik für elektronische Baugruppen

GMM Fachgruppe 1.1.4: Testequipment und -verfahren

VDE Verband der Elektrotechnik
Elektronik Informationstechnik e.V.

Merianstraße 28
63069 Offenbach am Main
Tel. +49 69 6308-0
service@vde.com
www.vde.com

VDE