Sachbericht zum Verwendungsnachweis

Teil I - Kurzbericht

Zuwendungsempfänger: IHP GmbH - Leibniz-Institut für innovative Mikroelektronik	**Förderkennzeichen:** 16ES1002
Vorhabenbezeichnung: Energieeffiziente Datenverarbeitung im autonomen Fahrzeug mittels Mehrprozessorsystem und integrierten KI-Beschleunigern - KI-PRO **Teilvorhaben:** Erforschung von RRAM-basierten KI-Hardwarebeschleunigern für automatisiertes Fahren	
Laufzeit des Vorhabens: 10.2019 – 03.2023	

Ursprüngliche Aufgabenstellung und State of the Art

KI-Lösungen gelten derzeit als vielversprechendster Ansatz zur Beherrschung komplexer Aufgaben wie etwa Bild-, Objekt- und Szenenerkennung oder Regelung dynamischer, nichtlinearer Systeme. Sie sind daher von besonderem Interesse hinsichtlich des Einsatzes in autonomen Fahrzeugen sowohl zur Sensordatenverarbeitung wie auch für die Fahrkontrolle. Für diesen Einsatz sind jedoch zwei Aspekte wesentlich: die Zuverlässigkeit - denn auch in einem potentiellen Fehlerfall muss das Fahrzeug weiterhin funktionssicher sein - und der Energieverbrauch. Dieser Aspekt kommt vor allem bei der E-Mobilität zum Tragen: existierende Lösungen für Echtzeit-KI-Anwendungen haben eine Leistungsaufnahme von 500W und mehr und setzen daher die Reichweite batteriegetriebener Fahrzeuge signifikant herab.

Das IHP adressiert diese Aspekte im KI-PRO Projekt durch die Integration der neuartigen *Resistive Random Access Memory* (RRAM)-Technologie in eine anwendungsspezifische Beschleuniger-Einheit und stellt damit eine leistungsfähige, aber energiesparende Rechenarchitektur für KI-Anwendungen zur Verfügung, die außerdem durch verschiedene Redundanz- und Fehlerkorrekturmechanismen spezielle Robustheit gegen auftretende Fehler zeigt.

Memristoren, von denen der RRAM eine Unterart darstellt, wurden theoretisch bereits in den frühen 1970er Jahren als viertes fundamentales passives Bauelement beschrieben, jedoch erst 2007 erstmalig physikalisch realisiert ([1]). Beim RRAM selbst handelt es sich um eine nichtflüchtige RAM-Speicherzelle mit einem elektrisch verstellbaren Widerstand, der gegenüber der oft verwendeten CMOS-basierten Implementierung nicht nur Größenvorteile zeigt, sondern auch anstatt nur eines digitalen Bits pro Zelle die Speicherung eines analogen und damit mehrstufigen Wertes erlaubt. Angeordnet als sogenannte Schachbrett- oder *Crossbar*-Struktur ermöglichen die RRAM-Zellen die Matrix-Vektor-Multiplikation der in ihnen gesicherten Werte direkt im Speicher und sind damit die ideale Grundlage für Anwendungen im KI-Bereich. Aus diesem Grund bilden memristive Speicherarrays den Gegenstand vieler aktueller Forschungsarbeiten ([2,3,4]). Die hier geplanten Arbeiten sollen darüber hinausgehen und einen RRAM-basierten Crossbar erstmals als in IHP-Technologie gefertigten, digitalen Beschleuniger für die Matrix-Vektor-Multiplikation verfügbar machen.

Ablauf des Vorhabens

Das IHP ist bereits seit über 10 Jahren auf dem Gebiet der RRAM-Technologieentwicklung tätig und hat dem Stand der Technik entsprechende Technologie zur Verfügung gestellt. Das Testvehikel des IHP zu Projektbeginn ist ein integriertes 4-kBit-RRAM-Array, welches in IHPs 250nm-CMOS-Technologie fabriziert wurde. Zur Erreichung der genannten Ziele muss das

[1] Siehe: https://ethw.org/Memristor (zuletzt besucht am 22.08.2023)
[2] 'A Functional Hybrid Memristor Crossbar-Array/CMOS System for Data Storage and Neuromorphic Applications - Nano Letters'. https://pubs.acs.org/doi/abs/10.1021/nl203687n, ACS Publications; accessed: 15-May-2018.
[3] P. Yao et al., 'Face classification using electronic synapses', Nat. Commun., vol. 8, p. 15199, May 2017.
[4] M. Hu et al., 'Memristor-Based Analog Computation and Neural Network Classification with a Dot Product Engine', Adv. Mater., vol. 30, no. 9, p. 1705914, Mar. 2018.

Testvehikel nun entsprechend weiterentwickelt und charakterisiert werden. Dazu zählt die Skalierung des RRAM-Prozesses von 250 auf 130nm, eine entsprechende Erweiterung des IHP PDK, die Fertigung und Charakterisierung eines 130nm Crossbars und die Entwicklung einer digitalen Schnittstelle für den neuromorphen Beschleuniger. Basierend auf der resultierenden Test-Hardware kann anschließend die Charakterisierung der elektrischen Parameter und vor allem der Zuverlässigkeitsmerkmale stattfinden. Die hier gewonnenen Kenntnisse dienen einerseits zur Verbesserung des neuromorphen Beschleunigers. Andererseits werden sie zur Untersuchung der Auswirkung der verschiedenen Variabilitäten und eventueller Fehlereffekte in ein entsprechendes Simulationsmodell integriert. Darauf aufbauend werden entsprechende Methoden der Fehlertoleranz entwickelt und fließen zusätzlich in das Design des Beschleunigers ein. Außerdem finden auf dem Simulationsmodell basierende Untersuchungen zum sogenannten *fault aware training* (FAT) Neuronaler Netze statt. Ziel ist es hier herauszuarbeiten, inwieweit ein Neuronales Netz unter gezieltem Training dazu in der Lage ist, die Variabilitäten und Fehlereffekte der genutzten Hardware zu tolerieren.

Wesentliche Ergebnisse / Zusammenarbeit mit anderen Forschungseinrichtungen

Zusammenfassend konnten wir in diesem sehr erfolgreichen Projekt drei wesentliche Ergebnisse erzielen. Als wichtigstes ist dabei der in 130nm IHP-Technologie gefertigte neuromorphe Beschleuniger selbst zu nennen. Programmierbar über ein digitales SPI-Interface, enthält er ein 16*16 Zellen umfassendes RRAM-Crossbar, in dem jede Zelle mit vier verschiedenen Werten programmiert werden kann. Die in jedem Takt durchführbare spaltenweise Multiplikation der Zellen mit einem vierwertigen Eingangssignal und die anschließend zeilenweise, analoge Aufsummierung der Ergebnisse bildet die Matrix-Vektor-Multiplikation ab . Analoge Bauteile, wie AD-, DA-Wandler und Verstärker wurden von der TU München entwickelt und gemeinsam integriert.

Das zweite wichtige Ergebnis sind die gewonnen Erkenntnisse durch die Charakterisierung. Zusammenfassend hat sich das elektrische Verhalten, im speziellen die Programmierbarkeit der einzelnen Zustände. sowie die Zuverlässigkeit der Zellen bei der Umstellung von 250 auf 130nm deutlich verbessert. Außerdem konnten im Test-Array weder *stuck-at-open* noch *stuck-at-short* Effekte gemessen werden und auch die Variabilität zwischen verschiedenen Zellen ist gesunken. Der nun dominierende Fehlereffekt ist das sogenannte *read-disturb*, bei dem sich der Wert in den Zellen durch wiederholte Lesevorgänge verändert. Dies lässt sich jedoch durch regelmäßiges Auffrischen der Werte relativ leicht verhindern.

Das dritte Resultat, welches spezielle Erwähnung finden soll, ist das Simulationsmodell des RRAM Crossbar. Dieses in SystemVerilog und SystemC zur Verfügung stehende Modell beschreibt im Wesentlichen das zeitliche und logische Verhalten der Zellen, sowie der analogen und digitalen Baugruppen. Außerdem werden basierend auf den Messdaten aus der Charakterisierung auch die Fehlereffekte realistisch abgebildet. Basierend auf diesem Modell konnte einerseits die Inferenz mehrerer neuronaler Netze validiert werden, die so auf dem physischen Beschleuniger nicht möglich gewesen wäre. Andererseits erlaubte dieses Modell die Untersuchung des FAT und führte zu interessanten Ergebnissen, was die Robustheit größerer Netze auf imperfekter Hardware angeht. Weiterhin konnten wir in Zusammenarbeit mit der Universität zu Lübeck einen RISC-V basierten Systemdemonstrator erstellen, der die tatsächliche Anwendbarkeit des neuromorphen Beschleunigers in einem realistischen Szenario unter Beweis stellt.

Sachbericht zum Verwendungsnachweis

Teil II – Eingehende Darstellung

Zuwendungsempfänger: IHP GmbH - Leibniz-Institut für innovative Mikroelektronik	**Förderkennzeichen:** 16ES1002
Vorhabenbezeichnung: Energieeffiziente Datenverarbeitung im autonomen Fahrzeug mittels Mehrprozessorsystem und integrierten KI-Beschleunigern - KI-PRO **Teilvorhaben:** Erforschung von RRAM-basierten KI-Hardwarebeschleunigern für automatisiertes Fahren	
Laufzeit des Vorhabens: 10.2019 – 03.2023	

Aufgabenstellung

KI-Lösungen gelten derzeit als vielversprechendster Ansatz zur Beherrschung komplexer Aufgaben wie etwa Bild-, Objekt- und Szenenerkennung oder Regelung dynamischer, nichtlinearer Systeme. Sie sind daher von besonderem Interesse hinsichtlich des Einsatzes in autonomen Fahrzeugen sowohl zur Sensordatenverarbeitung wie auch für die Fahrkontrolle. Für diesen Einsatz sind jedoch zwei Aspekte wesentlich: die Zuverlässigkeit - denn auch in einem potentiellen Fehlerfall muss das Fahrzeug weiterhin funktionssicher sein - und der Energieverbrauch. Dieser Aspekt kommt vor allem bei der E-Mobilität zum Tragen: existierende Lösungen für Echtzeit-KI-Anwendungen haben eine Leistungsaufnahme von 500W und mehr und setzen daher die Reichweite batteriegetriebener Fahrzeuge signifikant herab.

Das IHP adressiert diese Aspekte im KI-PRO Projekt durch die Integration der neuartigen *Resistive Random Access Memory* (RRAM)-Technologie in eine anwendungsspezifische Beschleuniger-Einheit und stellt damit eine leistungsfähige, aber energiesparende Rechenarchitektur für KI-Anwendungen zur Verfügung, die außerdem durch verschiedene Redundanz- und Fehlerkorrekturmechanismen spezielle Robustheit gegen auftretende Fehler zeigt.

Diese Zielstellung lässt sich in vier Teilaufgaben untergliedern:

1. Das Testvehikel des IHP zu Projektbeginn ist ein integriertes 4-kBit-RRAM-Array, welches in IHPs 250nm-CMOS-Technologie fabriziert wurde. Im ersten Schritt muss das Testvehikel entsprechend weiterentwickelt und charakterisiert werden. Dazu zählt die Skalierung des RRAM-Prozesses von 250 auf 130nm, eine entsprechende Erweiterung des IHP PDK, die Fertigung und Charakterisierung eines 130nm Crossbars und die Entwicklung einer digitalen Schnittstelle für den neuromorphen Beschleuniger.
2. Basierend auf der resultierenden Test-Hardware findet anschließend die Charakterisierung der elektrischen Parameter und vor allem der Zuverlässigkeitsmerkmale statt. Die hier gewonnenen Kenntnisse dienen der Verbesserung des neuromorphen Beschleunigers und der Untersuchung der Auswirkung der verschiedenen Variabilitäten und eventueller Fehlereffekte auf das System.
3. Zur Durchführung der Untersuchung werden die bei der Charakterisierung gewonnenen Erkenntnisse und Parameter in ein entsprechendes Simulationsmodell integriert. Darauf aufbauend werden entsprechende Methoden der Fehlertoleranz entwickelt und fließen zusätzlich in das Design des Beschleunigers ein.
4. Als letzte Aufgabe wird einerseits der finale Demonstrator gefertigt, auf ein Testboard gebracht und getestet. Außerdem finden auf dem Simulationsmodell basierende Untersuchungen zum sogenannten fault aware training Neuronaler Netze statt. Ziel ist es hier herauszuarbeiten, inwieweit ein Neuronales Netz unter gezieltem Training dazu in der Lage ist, die Variabilitäten und Fehlereffekte der genutzten Hardware zu tolerieren.

Voraussetzungen und Stand der Technik

Die Technologie, die zur Herstellung von memristiven Bauelementen benötigt wird, ist entscheidend für die Realisierung von neuromorphen Schaltungen. Insbesondere wird eine gute Technologie benötigt, die es ermöglicht, eine große Anzahl von memristiven Bauelementen mit einer geringen Streuung der Parameter zu erzeugen, und die es zusätzlich ermöglicht, diese Bauelemente in VLSI-Schaltungen zu integrieren.

Im Jahr 2011 haben Seo et al. unter der Verwendung einer 45-nm-CMOS-Technologie mit SRAM-Crossbar-Arrays, die als binäre synaptische Gewichte dienen, einen digitalen neuromorphen Chip zur Musterklassifizierung entwickelt [1]. Bei der Nutzung dieses nicht auf nicht-flüchtigem Speicher (NVM) basierenden CMOS-Technologie-Ansatzes wird die synaptische Gewichtungs-funktion in den digitalen SRAM-Bauelementen erzeugt. Die essentielle Integration von Neuronen und Synapsen in Netzwerken wurde erstmalig gezeigt. Im Jahr 2012 wurde von Chen et al. eine kosteneffektive 3D-Crossbar-Architektur ohne Selektor-Bauelemente auf der Basis von HfO2-RRAM-Bauelementen demonstriert [2]. Verglichen mit dem NVM-freien CMOS-basierten Implementierungsansatz zeigten RRAM-Bauelemente das Potenzial, die Größe der synaptischen Schaltung zu reduzieren. Zusätzlich können memristive RRAM-Bauelemente möglicherweise mehrere digitale Bits durch Speichern analoger synaptischer Gewichte ersetzen. Das RRAM-Bauelement besteht aus einer einfachen Metall-/Isolator-/Metall-Struktur, die kompakt, CMOS-kompatibel und hoch skalierbar ist. Zusätzlich kann der Energieverbrauch pro synaptischer Operation optimiert werden.

Im Jahr 2013 schlugen Ambrogio et al. ein synaptisches Konzept mit einer HfO2-basierten RRAM-Zelle vor, die mit einem selektierenden NMOS-Transistor gekoppelt ist [3]. Die vollständige Zelle besteht aus einem Transistor und einem Widerstand (1T-1R-Architektur). Der Transistor dient sowohl als Selektor als auch als spannungsabhängige Stromquelle. Die Spike-Timing-abhängige Plastizitätsmodulation des RRAM-Widerstands wurde durch Simulationen unter Verwendung eines kompakten RRAM-Modells demonstriert. Die Simulation basierte allerdings auf der experimentellen Auswertung eines einzelnen nach Transistor ausgewählten HfO2-RRAM-Bauelements.

Kim et al. erzeugten ein hybrides Crossbar-/CMOS-Speichersystem mit Multi-Level Memristive Devices und CMOS-Decodern. Ein 40×40 memristives Crossbar-Array wurde mittels einer Back-End-of-Line (BEOL) -Prozessierung auf einer vorgefertigten CMOS-Schaltungsplattform integriert [4]. Binäre Bitmap-Bilder wurden erfolgreich gespeichert und eine beträchtliche Wiederer-kennungsrate demonstriert. Darüber hinaus wurde ein neues Programmierschema entwickelt, um es dem integrierten Crossbar-Array zu ermöglichen, bis zu 10 verschiedene Ebenen zu speichern.

Inzwischen haben viele Forschungsteams memristive Crossbar-Arrays für neuronale Netzwerkanwendungen erforscht. Die meisten dieser Arbeiten beruhen jedoch auf Simulationen zur Vorhersage der Leistung für Berechnungen innerhalb von Crossbar-Arrays. Im Jahr 2016 wurde von Yao et al. eine optimierte Speicherzellenstruktur implementiert, die mit dem CMOS-Prozess kompatibel ist. [Yao]. Die RRAM-Bauelemente wurden in einem 1024-Zellen-Array integriert, das in

---

[1] J. s Seo et al., 'A 45nm CMOS neuromorphic chip with a scalable architecture for learning in networks of spiking neurons', in 2011 IEEE Custom Integrated Circuits Conference (CICC), 2011, pp. 1–4.

[2] H. Y. Chen, S. Yu, B. Gao, P. Huang, J. Kang, and H. S. P. Wong, 'HfOx based vertical resistive random access memory for cost-effective 3D cross-point architecture without cell selector', in 2012 International Electron Devices Meeting, 2012, p. 20.7.1-20.7.4.

[3] S. Ambrogio, S. Balatti, F. Nardi, S. Facchinetti, and D. Ielmini, 'Spike-timing dependent plasticity in a transistor-selected resistive switching memory', Nanotechnology, vol. 24, no. 38, p. 384012, 2013.

[4] 'A Functional Hybrid Memristor Crossbar-Array/CMOS System for Data Storage and Neuromorphic Applications - Nano Letters'. https://pubs.acs.org/doi/abs/10.1021/nl203687n, ACS Publications; accessed: 15-May-2018.

einem neuromorphen Netzwerk eingesetzt wird. Dieses analoge RRAM-Array demonsrierte eine bemerkenswerte Energieeinsparung.

Im Jahr 2018 wurde die hochpräzise analoge Abstimmung und Steuerung von memristiven Zellen in einem 128×64-Array für die Vektor-/Matrix-Multiplikation (VMM) demonstriert [5]. Die memristive Baulementintegration wurde auch hier durch einen CMOS-kompatiblen BEOL-Prozess durchgeführt. Es wurde gezeigt, dass ein großes memristives Crossbar-Array einstufige VMM- Operationen im Speicher ausführen kannt. Es wurde zudem gezeigt, dass die Rechengenauigkeit für Anwendungen im Bereich Machine Learning akzeptabel ist.

Das IHP arbeitet bereits seit mehr als einem Jahrzehnt aktiv an der CMOS-Integration der RRAM-Technologie und gehört in diesem Bereich zu den führenden Forschungseinrichtungen in Europa. Infolgedessen wurden große Erfahrungen in Bezug auf die Implementierung und Charakterisierung von 1T1R-RRAM-Zellen, die Entwicklung der Treiber- und Ausleselogik sowie die Bewertung und Verbesserung der Zuverlässigkeit gesammelt. In den letzten Jahren wurde ein integriertes 4-kbit-RRAM-Array auf Basis von HfO_2 entwickelt und charakterisiert (vgl. Abb. 2). Darüber hinaus wurde sogar das komplexe Mbit-Array für Raumfahrtanwendungen getestet. In den letzten Jahren lag der Fokus der Untersuchung auf neuromorphem Computing und der Verwendung der Plastizität der Speicherzellen für analoge Speicherfunktionen und in der In-Memory-Array-Verarbeitung. In diesem Zusammenhang werden bereits signifikante Ergebnisse erzielt, wenn das Array erfolgreich zur Mustererkennung in einem maschinellen Lernszenario verwendet wird.

Planung und Ablauf

Das Projekt war ursprünglich für eine Laufzeit von drei Jahren geplant und war in insgesamt acht Arbeitspakete unterteilt. Der Balkenplan für die ersten beiden Projektjahre kann Abbildung 1 und für das dritte Jahr Abbildung 2 entnommen werden.

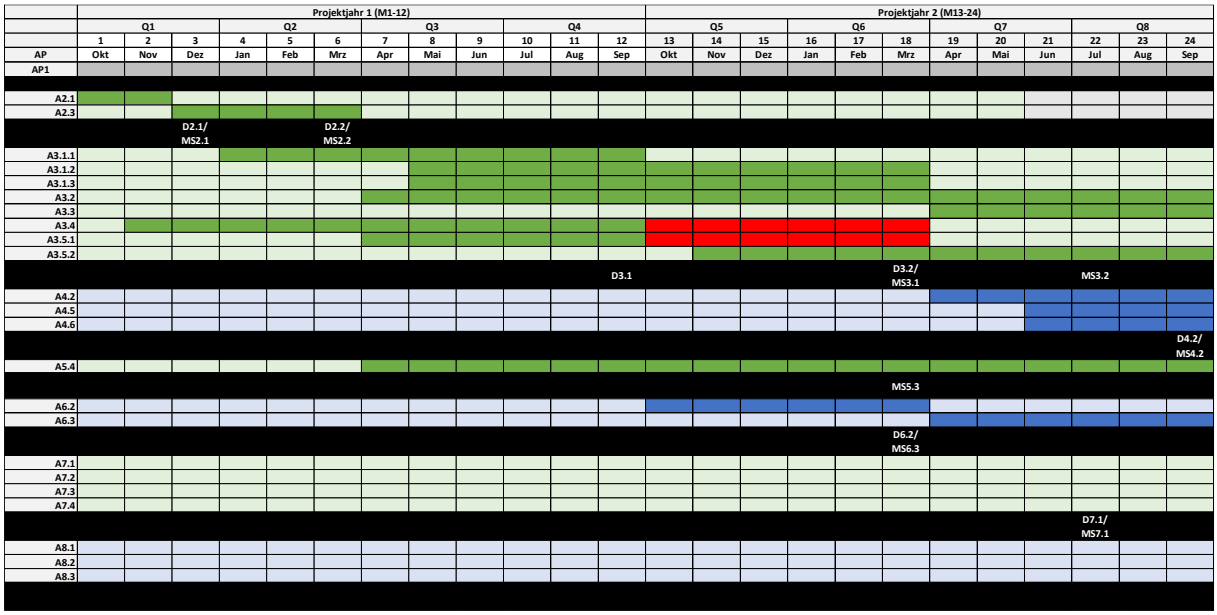

Abbildung 1: Projektplan Jahr 1 + 2

[5] M. Hu et al., 'Memristor-Based Analog Computation and Neural Network Classification with a Dot Product Engine', Adv. Mater., vol. 30, no. 9, p. 1705914, Mar. 2018.

Wie sich aus den Abbildungen erkennen lässt, konnte der ursprüngliche Zeitplan jedoch nicht eingehalten werden. Aufgrund Pandemie-bedingter Einschränkungen in den Lieferketten sowie unseren Laboren, kam es vor allem in AP3.4 und AP3.5 (detailliertere Erläuterung folgt) zu stärkeren Verzögerungen. Aus diesem Grund wurde das Projekt in Absprache mit dem Projektträger um 6 Monate verlängert (siehe Abbildung 2, rechts) und das zusätzliche Deliverable D-EX eingefügt.

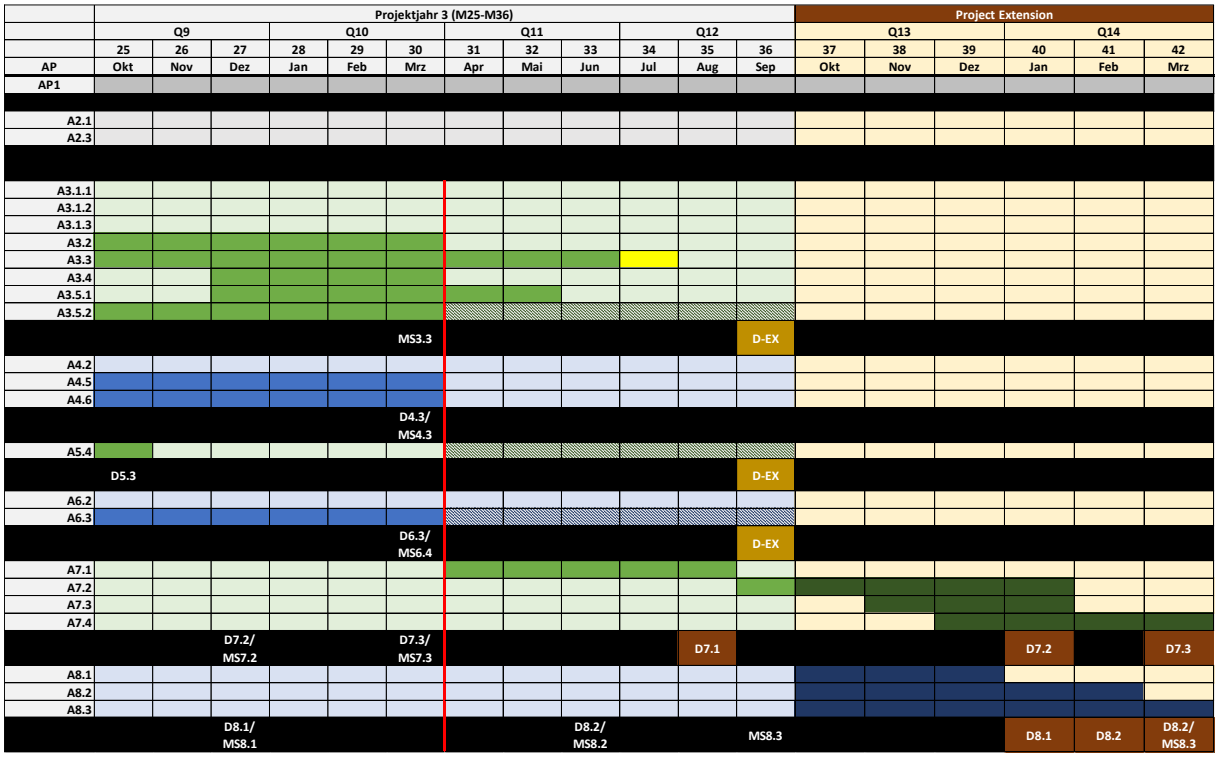

Abbildung 2: Projektplan Jahr 3 + Verlängerung

	Delay in Workpackages
	Originally planned finalization
	Planned extension
	Planned Tape-In
	Finalized, but possible refinement after characterization
	Finalized, but possible refinement after characterization
D / M	Deliverable / Milestone on time
D / M	Deliverable / Milestone delayed

Abbildung 3: Legende zum Projektplan

AP1: Verbundprojektleitung

Das Projektmanagement übernahm der Koordinator Universität zu Lübeck und wurde dabei vom IHP bei allen Aufgaben unterstützt, die in IHP-Verantwortung lagen.

AP2: Systemanalyse und Entwurfsspezifikation

Basierend auf den branchenspezifischen Anforderungen und der Beschreibung der Testfälle, wurde die Architektur der RRAM-Beschleunigerschaltung definiert. Diese umfasst die Partitionierung der in ASIC und FPGA implementierten Module, sowie die Schnittstellendefinition. Das IHP erarbeitete hier die Basisspezifikation der RRAM-Funktion und spezifizierte die initiale Architektur des Hardwarebeschleunigers. Die Ergebnisse dieses AP waren Basis für alle anderen technischen APs. Die hier definierten Spezifikationen inklusive Verifikationsmethoden, wurden zur Validierung des KI-PRO-Konzepts im finalen AP 8 verwendet.

AP3: Entwurf und Fertigung des RRAM-basierten Speicherarrays

Dieses Arbeitspaket widmete sich der Implementierung des neuromorphen Beschleunigers. Aus technologischer Sicht war die erste Aufgabe die Skalierung der RRAM-Zelle von einem 250 nm- auf einen 130 nm-Prozess. Basierend auf dieser technologischen Entwicklung wurde das neuartige RRAM-Array entwickelt. Der Schwerpunkt lag dabei auf neuen Lese-/Schreibschnittstellen, die die Kommunikation mit dem in einem FPGA integrierten RISC-V-Prozessor ermöglichen. Das Array wurde als Crossbar organisiert, um eine Matrixmultiplikationsoperation zu ermöglichen. Im Zuge dessen mussten zuverlässige und energieeffiziente Schreib- und Lesealgorithmen entwickelt werden. Dies beinhaltete auch den Entwurf von analogen Schaltungsteilen zur peripheren Beschaltung des Arrays, der von dem Projektpartner TU München durchgeführt wurde. Im Zuge dessen wurde ein Array mit Multi-Bit-Fähigkeit der einzelnen Elemente, sowie im späteren Verlauf des Arbeitspaketes ein auf Matrixmultiplikation optimiertes Array entwickelt. Die entworfene Schaltung wurde dem regulären 130 nm-Tape-out vorgelegt und hergestellt. Nach der Produktion wurde der Chip mit spezifischen Testgeräten getestet und charakterisiert. Parallel zu diesen Entwurfsaktivitäten wurde basierend auf den Zelleigenschaften ein RRAM-Beschleunigermodell entwickelt, welches beim Systementwurf der komplexen neuromorphen KI-PRO-Lösung Anwendung fand. Auf Grundlage des Funktions- und Fehlermodells der RRAM-Bauelemente, wurden schließlich fehlertolerante Mechanismen untersucht. Ziel ist, die technologische Unvollkommenheit zu überwinden und den Einsatz solcher Elemente für äußerst kritische Anwendungen, beispielsweise für automatisiertes Fahren, zu ermöglichen.

Wie bereits angedeutet, kam es Aufgrund der Corona-Pandemie zu unvorhergesehenen Verzögerungen in den Lieferketten und in unseren Laboren. Aus diesem Grund konnten die Chips nicht rechtzeitig erstellt werden, was direkte Auswirkungen auf die Arbeitspakete AP3.4 „RRAM-Modellierung“ und AP3.5.1 „Analyse der Zuverlässigkeitsmerkmale von RRAM-Zellen und –Arrays“ hatte. Erst ca. ein Jahr später konnten diese Arbeiten erfolgreich abgeschlossen und mit den zusätzlich geplanten Arbeiten für D-EX sogar deutlich erweitert werden.

AP4: Verlässliche Hochleistungs-Rechenplattform mit HW-Beschleunigern für KI-gestützte Sensordatenverarbeitung

Die Hardwareplattform besteht aus einer RISC-V-basierten Hostumgebung, an welche sowohl klassische digitale KI-Beschleuniger, als auch die im Rahmen des Projekts entwickelten neuartigen RRAM-basierten Beschleuniger angeschlossen wurden. Dies bedurfte der Entwicklung einer geeigneten SPI-Hardwareschnittstelle zur Steuerung des Beschleunigers, sowie Arbeiten hinsichtlich spezieller Trainingssoft- und –hardware.

Auf AP4 hatten die Verzögerungen im Ablauf keine Auswirkungen und alle Aufgaben konnten wie geplant abgeschlossen werden.

AP5: Betriebssicherer Einsatz von KI-Algorithmen

In diesem Arbeitspaket wurde die Fehlertoleranz von ADAS-relevanten KI-Algorithmen untersucht.

Auch auf AP5 hatten die Verzögerungen im Ablauf keine Auswirkungen, jedoch konnten wir mit den für D-EX geplanten Arbeiten zum fault aware training die ursprünglich geplante Zielstellung deutlich übertreffen.

AP6: Validierungsplattform

Für die Validierungsplattform wurde vom IHP ein spezielles Simulationsmodell erstellt, dass das physische Verhalten der RRAM-Zellen nicht nur in Hinsicht auf Strom- Spannungs- und Verzögerungswerte realistisch abbildet, sondern auch die verschiedenen Fehlereffekte modelliert.

Für D-Ex wurde das Modell über entsprechende Schnittstellen mit dem RISC-V-basierten, virtuellen Prototypen der Universität Lübeck verbunden und hat somit einen Gesamtsystemdemonstrator emuliert, auf dem größere neuronale Netze abgebildet werden konnten.

AP7: Integration der KI-Plattform und Implementierung von Demonstratoren

Für die Integration der einzelnen Komponenten (inklusiv RRAM-Beschleuniger) wurde eine Basisarchitektur für die Demonstratorplattform entwickelt. Dafür wurde für das in AP3 entwickelte RRAM-Array eine separate Platine entwickelt, welche über eine dedizierte Schnittstelle mit dem Basisboard gekoppelt wurde.

In diesem AP, sowie in AP8 traten die größten Verzögerungen auf, da direkt Abhängigkeiten zu AP3 bestanden. Die Entwicklung der Platine wurde erst begonen, als der in AP3 entwickelte Beschleuniger im Wesentlichen fertig war.

AP8: Validierung

In diesem AP erfolgte die Validierung der Projektergebnisse in Hinblick auf die vereinbarten Anwendungsszenarien. In diesem Zusammenhang hat das IHP kleinere neuronale Netze auf dem RRAM-Beschleuniger ausgeführt. Physisch konnte auch dies leider erst nach Projektende geschehen, da der finale Chip nicht rechtzeitig aus der Fertigung kam. Die aus der Simulation gewonnen Ergebnisse waren jedoch vielversprechend und können aufgrund ihrer Genauigkeit als repräsentativ angenommen werden.

Zusammenarbeit mit anderen Stellen

Zunächst ist an dieser Stelle natürlich die Interaktion mit dem Konsortium als Ganzes zu nennen. Innerhalb des Projektes gab es einen regen Informations-, Wissens- und Technologieaustausch mittels abgehaltener Telefonkonferenzen und Emails.

Speziell soll aber auf die Zusammenarbeit mit zwei Projektpartnern hingewiesen werden. Eines der bedeutendsten Ziele des Projekts war die Entwicklung eines RRAM-basierten Beschleunigers, der sich in eine digitale Prozessorstruktur einfügt. Die Matrix-Vektor-Multiplikation in den Zellen findet jedoch im analogen Bereich statt. Die Brücke zwischen diesen Welten wurde von der TU München geschlagen. Sie haben für uns sämtliche analogen Bauteile, sowie die A/D- und D/A-Wandler entworfen, die die Entwicklung des Systems überhaupt erst ermöglichten. Für den virtuellen Prototypen hat uns die Zusammenarbeit mit der Universität zu Lübeck sehr bereichert. Ihr SystemC/TML-basiertes Modell eines RISC-V Prozessors erlaubte in Zusammenarbeit mit unserem SystemC Modul des RRAM-Beschleunigers die Abbildung neuronaler Netze in realistischem Umfang.

Verwendung der Zuwendung

Die Zuwendung wurde im Wesentlichen für die Bezahlung der technischen und administrativen Mitarbeiter, für die Fertigung der verschiedenen Testchips für den RRAM-Crossbar und den RRAM-basierten Beschleuniger sowie zugehörige Evaluationsboards verwendet.

Kosten für Gehälter

Da die wesentliche Aufgabe des IHP im KI-PRO Projekt in der Entwicklung von Methoden und Entwurf der Charakterisierung des RRAM-basierten Beschleunigers lag, ist der größte Anteil der Zuwendung für die Gehälter verwendet worden. Dabei wurden die projektkritischen Aufgaben von erfahrenen Mitarbeitern ausgeführt, deren Gehälter in die Gruppen E12-E14 fallen. Hierfür fielen Kosten in Höhe von 540.568,78€ an, die unter Position 0812 verbucht wurden.

Kosten für Verbrauchsmaterialien

Für die Fertigung der verschiedenen ICs mit RRAM-Komponenten mussten zunächst die Chips selbst im Haus gefertigt werden. Dafür fielen Kosten für Material (Masken, Wafer, Chemikalien), sowie für die Fertigung an sich an. Auch Kosten für Veröffentlichungen und die Anschaffung projektrelevanter Literatur zählen zu den insgesamt 151.468,06€, die unter Pos. 0843 verbucht wurden.

Wichtigste Positionen des zahlenmäßigen Nachweises

Im Folgenden sind die wichtigsten Positionen der Mittelverwendung dargelegt. Insgesamt hatte das Projekt ein Budget von 727.308,59€.

- Für Gehalt und Beschäftigungsentgelte wurden insgesamt 564.048,31€ verbucht, wobei 540.568,78€ auf erfahrene Beschäftigte E12-E14 (Pos. 0812) entfielen, die für projektkritische Aufgaben eingesetzt wurden. Für weniger kritischen Aufgaben wurden zusätzlich 23.479,53€ (Pos. 0822) für studentische Hilfskräfte ausgegeben.
- Für allgemeine Verwendungszwecke, wie Verbrauchsmaterialien, open access Publikationen und Literaturbeschaffung (Pos. 0843) wurden 151.468,06 € ausgegeben.
- Unter Vergabe von Aufträgen (Pos. 0835) wurden 2.013,29€ für PTSL und Multi Circuit Boards, sowie Grindingkosten abgerechnet.
- Schließlich wurden 9.779,19 € für Dienstreisen (Pos 0846) benötigt.
    - Auslandsreisen: Genf SERESSA 2022; Granada Development of Compact Models for Memristors 2020; Montecatini CIMTEC 2020; Granade ELICSIR Training School 2022
    - Inlandreisen: Aachen Summer School 2022; MEMRIS Tec Workshop Dresden 2022; VDE ITG MN Fachtagung Dresden 2022; 10. MEMRISTOR Coloqium Bamberg 2023 und Projektmeetings

Notwendigkeit und Angemessenheit der geleisteten Projektarbeiten

Das Vorhaben war geplant als vorwettbewerbliche industrielle Forschung zur Definition und Umsetzung der Architektur eines Komponentensystems für die KI-basierte Sensordatenverarbeitung für autonome Kraftfahrzeuge. Die Entwicklungsrisiken begründen sich in der Entwurfsphase des Projektvorhabens auf die flächen- und leistungsspezifische Umsetzung der redundanten Verarbeitungsarchitektur, die optimale Auslegung und Anbindung von Beschleunigern für KI-Algorithmen, sowie die erreichbare Fehlertoleranz. Dies soll durch detaillierte Analyse der Entwurfsanforderungen und Planung miteinbezogen und somit verringert werden. Die erreichbare Rechenleistung und Fehlertoleranz wird durch Validierung und Tests an Demonstratoren verifiziert.

Das Vorhaben mit eindeutigem industriellen Forschungscharakter, – auch unter Einbeziehung der notwendigen Forschungskompetenz von TUBS auf dem Gebiet der Prozessorarchitektur, DNDEs Expertise im Bereich betriebssicherer Hardwareplattformen und IHP auf dem Gebiet der Speicherarchitekturen, – wird benötigt als Vorlauf zu einer mögliche späteren ASIC-Umsetzung. Die ASIC-Umsetzung ist die eigentliche Zielsetzung eines Gesamtvorhabens, die aber den Vorlauf durch KI-PRO in einer Phase 1 zur Reduzierung der technischen Risiken benötigt.

Nutzen und Verwertbarkeit des Ergebnisses inkl. konkrete Planungen

Im Projektantrag, wurden mögliche verwertbare Ergebnisse definiert, die in den Zwischennachweisen wie folgt konkretisiert und fortgeführt wurden:

Ergebnis	Verwertung	Zeithorizont

RRAM-Prozess	Es wird erwartet, dass nach Abschluss des Projekts einige Jahre ausreichen würden, um den RRAM-Prozess für die IHP-Foundry kommerziell zu qualifizieren. Das IHP verfügt über beträchtliche Erfahrung in der kommerziellen Nutzung der Technologie und in diesem Prozess spielt das Tochterunternehmen IHP-Solutions eine wichtige Rolle. Es wird erwartet, dass dieser Technologieprozess für die Kunden aus Industrie und Wissenschaft für MPW- und Engineering-läufe angeboten werden kann. Dieser RRAM-Prozess könnte für neuromorphe, aber auch für klassische NVM-Anwendungen verwendet werden.	mittelfristig
RRAM-Crossbar-Array	Aufgrund der erfolgreichen Projektergebnisse könnte das implementierte RRAM-Array kommerziell genutzt werden. Voraussetzung ist in diesem Fall eine erfolgreiche Chip-Qualifizierung und Bestätigung der Zuverlässigkeitsaspekte. Zu den Zielanwendungen zählen automatisiertes Fahren, aber auch Industrie 4.0, Medizin, Big Data usw.	langfristig
Neue Methoden für fehlertolerantes Computing mit RRAM	Es wird erwartet, dass die neuen Verfahren zur Erhöhung der Zuverlässigkeit eines RRAM-Arrays ebenfalls einen kommerziellen Wert haben. Alle relevanten Erfindungen werden geschützt und die Verwertung kann in Form von Lizenzen oder Industrieaufträgen erfolgen.	mittel-/langfristig

Der erwartete Markt und die Chancen für eine Marktdurchdringung haben sich gegenüber dem Berichtszeitraum nicht geändert. Die Hauptaspekte des Verwertungsplans sind Weiterhin gültig.

Weiterführung des Stands der Technik anderer Stelle

Memristoren sind aufgrund ihrer Vielseitigkeit natürlich Gegenstand aktueller Forschung in vielen Einrichtungen. Jedoch sind die wenigsten davon in der Lage, diese selbst zu fertigen und beschränken sich deshalb auf rein theoretische Betrachtungen. Dies verleiht den Forschungen im Projekt große Relevanz.

Die Autoren von [6] hingegen sind durchaus in der Lage, die Memristoren zu fertigen, nutzen aber eine auf Wolfram-Oxid (WO_x) basierende Technologie. Abgesehen von den technologischen Unterschieden setzen sie auch auf eine sehr starre Crossbar-Architektur ohne Tiles, die wenig Skalierbarkeit verspricht. Außerdem widmen sie sich nicht dem Thema Fehlertoleranz, welches im KI-PRO Projekt zentral ist.

Zusätzlich dazu sind keine Ergebnisse Dritter bekannt geworden, die für die Durchführung und das Erreichen der Ziele des Vorhabens relevant waren.

Veröffentlichungen

[1] S. Pechmann; T. Mai, M. Völkel, M.K. Mahadevaiah, E. Perez, E. Perez-Bosch Quesada, M. Reichenbach, Ch. Wenger, A. Hagelauer; "A Versatile, Voltage-Pulse Based Read and Programming Circuit for Multi-Level RRAM Cells"; Electronics (MDPI)

[2] M. Ulbricht; J. Wen, A. Veronesi; "An Accelerator for Neuromorphic Computing Based on Memristive Crossbar Arrays in IHP Technology"; ITI Oberseminar (2022), Lübeck, January 17, 2022, Germany

[3] R. Romero-Zaliz; A. Cantudo, E. Perez, F. Jimenez-Molinos, Ch. Wenger, J.B. Roldan; "An Analysis on the Architecture and the Size of Quantized Hardware Neural Networks based on Memristors"; Electronics (MDPI)

[4] J. Wen; M. Ulbricht, E. Perez, X. Fan, M. Krstic; "Behavioral Model of Dot-Product Engine Implemented with 1T1R Memristor Crossbar Including Assessment"; Proc. 24th International Symposium on Design and Diagnostics of Electronic Circuits and Systems (DDECS 2021), 29 (2021)

[5] J. Wen; M. Ulbricht, X. Fan, M. Krstic; "Behavioral Simulation of Dot-Product Engine Implemented with 1T1R Memristor Crossbar Including Assessment"; Proc. 33. GI/GMM/ITG-Workshop Testmethoden und Zuverlässigkeit von Schaltungen und Systemen (TuZ 2021), 59 (2021)

[6] A. Veronesi; "Characterizing the Hardware/Software Stack of the NVIDIA Deep Learning Accelerator for FPGA and ASIC Technologies"; Masters Thesis; University of Ferrara, Ferrara, Italy (2020)

[7] Veronesi; M. Krstic, D. Bertozzi; "Cross-Layer Hardware/Software Assessment of the Open-Source NVDLA Configurable Deep Learning Accelerator"; Proc. 28th IFIP/IEEE International Conference on Very Large Scale Integration (VLSI-SoC 2020), (2020)

[8] J. Wen; A. Baroni, E. Perez, M. Ulbricht, Ch. Wenger, M. Krstic; "Evaluating Read Disturb Effect on RRAM based AI Accelerator with Multilevel States and Input Voltages"; Proc. 35th IEEE International Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems (DFTS 2022), (2022)

[9] E. Perez-Bosch Quesada; M.K. Mahadevaiah, T. Rizzi, J. Wen, A. Baroni, M. Ulbricht, M. Krstic, Ch. Wenger, E. Perez; "Experimental Assessment of Multilevel RRAM-Based Vector-Matrix Multiplication Operations"; Memristor-Symposium 2023, Bamberg, February 27 - 28, 2023, Germany

[10] E. Perez-Bosch Quesada; M.K Mahadevaiah, T. Rizzi, J. Wen, M. Ulbricht, M. Krstic, Ch. Wenger, E. Perez; "Experimental Assessment of Multilevel RRAM-based Vector-Matrix Multiplication Operations for In-Memory Computing"; IEEE Transactions on Electron Devices

[6] Cai, F., Correll, J.M., Lee, S.H. *et al.* A fully integrated reprogrammable memristor–CMOS system for efficient multiply–accumulate operations. *Nat Electron* **2,** 290–299 (2019). https://doi.org/10.1038/s41928-019-0270-x

[11]E. Perez-Bosch Quesada; E. Perez, Ch. Wenger; "From Telecommunication to Memristors and all the Steps in Between"; Invited presentation at University of Jaen, Jaen, October 28, 2022, Spain
[12]R. Romero-Zaliz; E. Perez, F. Jimenez-Molinos, Ch. Wenger, J.B. Roldan; "Influence of Variability on the Performance of HfO2 Memristor-based Convolutional Neural Networks"; Solid State Electronics
[13]M. Fritscher; J. Knödtel, M. Mallah, S. Pechmann, E. Perez-Bosch Quesada, T. Rizzi, Ch. Wenger, M. Reichenbach; "Mitigating the Effects of RRAM Process Variation on the Accuracy of Artifical Neural Networks"; Proc. 21st International Conference on Embedded Computer Systems: Architectures, Modeling and Simulation (SAMOS 2021), in: Lecture Notes in Computer Science, Springer, LNCS 13227, 401 (2022)
[14]A.J. Perez-Avila; E. Perez, J.B. Roldan, Ch. Wenger, F. Jimenez-Molinos; "Multilevel Memristor based Matrix-Vector Multiplication: Influence of the Discretization Method"; Proc. 13th Spanish Conference on Electron Devices (CDE 2021), 66 (2021)
[15]M. Krstic; "Neue Technologien für energieeffiziente und zuverlässige hardwarebasierte KI"; 2. KI Tag Brandenburg, Luckenwalde, June 19, 2023, Germany
[16]E. Perez; A.J. Perez-Avila, R. Romero-Zaliz, M.K. Mahadevaiah, E. Perez-Bosch Quesada, J.B. Roldan, F. Jimenez-Molinos, Ch. Wenger; "Optimization of Multi-Level Operation in RRAM Arrays for In-Memory Computing"; Electronics (MDPI)
[17]V. Milo; F. Anzalone, C. Zambelli, E. Perez, M.K. Mahadevaiah, O.G. Ossorio, P. Olivo, Ch. Wenger, D. Ielmini; "Optimized Programming Algorithms for Multilevel RRAM in Hardware Neural Networks"; Proc. International Reliability Physics Symposium (IRPS 2021), (2021)
[18]O.G. Ossorio; G. Vinuesa, H. Garcia, B. Sahelices, S. Duenas, H. Castan, E. Perez, M.K. Mahadevaiah, Ch. Wenger; "Performance Assessment of Amorphous HfO2-based RRAM Devices for Neuromorphic Applications"; ECS Transactions
[19]M. Krstic; "Reliability Evaluation of General Purpose and AI Processing Architectures"; 10th Prague Embedded Systems Workshop (PESW 2022), Horomerice, June 30 - July 02, 2022, Czech Republic
[20]M. Krstic; "Reliability in AI Processing"; Workshop "Com-In-AI", Faculty of Electronic Engineering, Nis, May 31, 2022, Serbia
[21]T. Rizzi; "Scalable Simulation Methodologies and Tools for the Design and Characterization of Hardware Accelerators based on Resistive RAMs"
[22]Glukhov; V. Milo, A. Baroni, N. Lepri, C. Zambelli, P. Olivo, E. Perez, Ch. Wenger, D. Ielmini; "Statistical Model of Program/Verify Algorithms in Resistive Switching Memories for In-Memory Neural Network Accelerators"; Proc. International Reliability Physics Symposium (IRPS 2022), 3C.3-1 (2022)
[23]R. Romero-Zaliz; E. Perez, F. Jimenez-Molinos, Ch. Wenger, J.B. Roldan; "Study of Quantized Hardware Deep Neural Networks Based on Resistive Switching Devices, Conventional versus Convolutional Approaches"; Electronics (MDPI)
[24]E. Perez-Bosch Quesada; R. Romero-Zaliz, E. Perez, M.K. Mahadevaiah, J. Reuben, M.A. Schubert, F. Jimenez-Molinos, J.B. Roldan, Ch. Wenger; "Toward Reliable Compact Modeling of Multilevel 1T-1R RRAM Devices for Neuromorphic Systems"; Electronics (MDPI)
[25]Al Beattie; E. Perez-Bosch Quesada, M. Uhlman, E. Perez, G. Kahmen, E. Solan, K. Ochs; "Wave Digital Emulation of an Enhanced Compact Model for RRAM Devices with Multilevel Capability"; IEEE Transactions on Nanotechnology