

GEFÖRDERT VOM



Bundesministerium
für Bildung
und Forschung

Teil I

**Rheinisch-Westfälische Technische Hochschule Aachen
und
Forschungszentrum Jülich**

Zukunftscluster NeuroSys

Neuromorphe Hardware für autonome Systeme

Projekt C: Algorithmen-Hardware Co-Design

Prof. Dr.-Ing. Tobias Gemmeke

Förderkennzeichen: 03ZU1106CA/03ZU1106CB

Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

Einleitung

Teilprojekt C zielte auf den Co-Entwurf von Algorithmen und Hardware ab. Dazu wurden in verschiedenen Kooperationen mit anderen Projektteilnehmern die hardwareseitigen Charakteristiken (Projekte A und B) simuliert und auf der Algorithmenseite (Projekt D) angewendet. Durch ausgiebige Explorationen und mehrere Optimierungszyklen konnten hocheffiziente Implementierungen realisiert werden. Im Folgenden findet sich ein Überblick zu genaueren Aktivitäten der einzelnen Projektpartner.

AP C1: Quantitative Entwurfsraumexploration

Prof. Dr. Tobias Gemmeke / RWTH Aachen, IDS

Das Arbeitspaket C1 fokussierte sich auf die Entwicklung und prototypische Erprobung einer Methodik für den Algorithmen-Hardware Co-Entwurf. Ausgangspunkt waren Modellierungsansätze, die physikalische Eigenschaften (Projekte A und B) und algorithmische Anforderungen (Projekt D) verbinden. Dazu wurden zunächst existierende Werkzeuge und Simulatoren identifiziert und analysiert. Anschließend wurde eine Entwurfs- und Explorations-Methodik entwickelt. Diese wurde dann für spezifische neuronale Anwendungsfälle eingesetzt wie beispielsweise automatische Spracherkennung (HLTPR, Projekt D2) und medizinische Überwachung (INDA, Projekt D4). Ergebnis dieser Arbeiten sind verschiedene Hardwarebeschleuniger, die basierend auf einer Exploration algorithmischer sowie hardwarespezifischer Metriken optimiert wurden.

AP C2: Entwurf und Entwurfsmethodik neuromorpher memristiver Schaltkreise

Prof. Dr. Stefan Heinen / RWTH Aachen, IAS

Es wurden Analogschaltungsblöcke zum Betreiben von Memristor-Matrizen entwickelt, die es erlauben, die analogen Eigenschaften dieser Bauelemente mit einer höheren Effizienz als vorherige Implementierung zu nutzen. Zunächst wurde die CMOS-Prozessierung des Partners ELMOS auf die Memristoren vom Partner PGI-10 abgestimmt. Nach Einfügen dieser Bauelemente in der Softwareumgebung wurden Analogblöcke gestützt von Memristor-Modellen der Partner vom IWE-2 entwickelt und ein Demonstrator-Chip gefertigt. Auf diesem können die einzelnen Memristoren integriert geformt, programmiert und gelesen werden. Damit können Matrix-Vektor-Multiplikationen durchgeführt werden. Die einzelnen Blöcke wurden mit Messungen charakterisiert und die Funktionalität des Systems bestätigt.

AP C3: Entwurf und Optimierung der analogen neuromorphen Schaltungen und elektro-optischen Schnittstellen

Prof. Dr. Renato Negra / RWTH Aachen, HFE

In Projekt B wurde die Implementierung eines photonischen neuronalen Netzes angestrebt. Das Teilprojekt C3 zielte darauf ab, in enger Zusammenarbeit mit den Projektpartnern des IPH aus Projekt B eine optoelektronische Schnittstelle für dieses photonische neuronale Netz zu gestalten, welches die neuronale Aktivierungsfunktion implementiert und die photonischen Schaltungen um elektronische Funktionen ergänzt. Dazu wurden zunächst die Anforderungen an die optoelektronische Schnittstelle von photonischer Seite definiert, um im Anschluss einen elektronischen CMOS-Chip auszulegen, welcher diesen Anforderungen gerecht wird. Dieser elektronische Chip ist für ein Flip-Chip Packaging mit dem photonischen Chip des IPH abgestimmt worden. In einer zweiten Iteration wurde dann der Funktionsumfang des elektronischen Chips optimiert und erweitert, basierend auf neu gewonnenen Erfahrungen.

AP C4: Software-Mapping und Neuromorphic Security

Prof. Dr. Rainer Leupers / RWTH Aachen, SSS

Im Rahmen des Arbeitspakets C4.1 wurden Software Development Kit (SDK) und die Hardwaresicherheit neuromorpher Systeme untersucht. Dazu wurde ein SDK für Systems-on-Chip (SoCs) mit neuromorphen Hardware-Beschleunigern entworfen und implementiert. Die Arbeiten umfassten zunächst die Analyse bestehender Machine-Learning-Frameworks sowie die Definition einer Software-Zwischenrepräsentation, die als Schnittstelle zur Hardware dient. Darauf aufbauend wurden schrittweise drei unterschiedliche Compilerprototypen entwickelt. Abschließend wurde das SDK durch Benchmarks verifiziert und hinsichtlich seiner Leistungsfähigkeit evaluiert.

Für Arbeitspaket C4.2 wurde die Sicherheit neuromorpher Systeme insbesondere im Hinblick auf neuartige Angriffsvektoren für Hardware-Trojaner untersucht. Nach Analyse der typischen Hardware-Architekturen wurden mögliche Payloads für Trojaner identifiziert und die Vielversprechendsten detailliert analysiert. Dabei konnte gezeigt werden, dass ein Trojaner Leistungsseitenkanalattacken ermöglichen und erleichtern kann, womit sensible Zwischenergebnisse aus zukünftigen Chips extrahiert werden können. Es wurde ein Algorithmus entworfen und evaluiert, der diese Zwischenergebnisse zusammensetzen kann. So könnte ein Angreifer sensibles geistiges Eigentum stehlen.

AP C5: Anforderungen an Optimierungsmöglichkeiten durch Lokalität und Spärlichkeit

Prof. Dr. Markus Diesmann / FZJ, INM-6

Prof. Dr. John Paul Strachan / FZJ, PGI-14

Prof. Dr. Emre Neftci / FZJ, PGI-15

In diesem Teilprojekt ging es darum, die Anforderungen an einen neuromorphen Rechner herauszuarbeiten, die es ermöglichen, die Lokalität und Spärlichkeit der Berechnungen in natürlichen neuronalen Netzwerken zu nutzen. Es sind diese beiden Eigenschaften, die erheblich zu der Energieeffizienz von Gehirnen beitragen. Lokalität bedeutet, dass für das Lernen des Systems an jeder Synapse nur Information der vorgeschalteten und der nachgeschalteten Nervenzelle sowie unspezifische Fehlersignale notwendig sind. Spärlichkeit bedeutet, dass ein Neuron selbst mit den Neuronen in der unmittelbaren Umgebung nur mit einer Wahrscheinlichkeit von zehn Prozent eine Verbindung hat. Besteht eine Verbindung wird über diese nur einmal pro Sekunde ein kurzes Signal ausgetauscht. Es handelt sich also um eine Spärlichkeit in Raum und Zeit. Das Teilprojekt hat erforscht, wie die natürliche Spärlichkeit in einem neuromorphen Computer genutzt werden kann, um die verschiedenen an der Rechnung beteiligten Komponenten voneinander zu entkoppeln und eine gute Lastverteilung zu erreichen. Weiterhin wurde erforscht, wie mit den spärlichen und punktaktigen Signalen im Gehirn ein effizienter Lernalgorithmus implementiert werden kann, der tatsächlich nur lokale Information verwendet. Weiterhin zeigt das Teilprojekt, wie solche sogenannte Gradienten-basierte Algorithmen mit Memristoren, neuartigen energiesparenden Bauelementen, implementiert werden können. Sowohl die natürlichen als auch die neuromorphen Bauelemente weisen ein stochastisches Verhalten auf. Das Teilprojekt zeigt, wie dieses mathematisch erfasst und kontrolliert werden kann. An einem Beispiel wird gezeigt, wie die geschaffenen Grundlagen dazu genutzt werden können, neuromorphe Systeme zu konstruieren, die lernen, konkrete Anwendungsprobleme zu lösen.

GEFÖRDERT VOM



Bundesministerium
für Bildung
und Forschung

Teil II

**Rheinisch-Westfälische Technische Hochschule Aachen
und
Forschungszentrum Jülich**

Zukunftscluster NeuroSys

Neuromorphe Hardware für autonome Systeme

Projekt C: Algorithmen-Hardware Co-Design

Prof. Dr.-Ing. Tobias Gemmeke

Förderkennzeichen: 03ZU1106CA/03ZU1106CB

Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

Inhalt

Einleitung.....	4
AP C1: Quantitative Entwurfsraumexploration	5
Wissenschaftlich-technische Ergebnisse.....	5
Die wichtigsten Positionen des zahlenmäßigen Nachweises	6
Notwendigkeit und Angemessenheit der geleisteten Projektarbeiten.....	6
Voraussichtlicher Nutzen, fortgeschriebener Verwertungsplan	6
Während der Durchführung des Vorhabens bekannt gewordener Fortschritt auf dem Gebiet des Vorhabens bei anderen Stellen.....	7
Publikationen	7
AP C2: Entwurf und Entwurfsmethodik neuromorpher memristiver Schaltkreise.....	9
Wissenschaftlich-technische Ergebnisse.....	9
Die wichtigsten Positionen des zahlenmäßigen Nachweises	10
Notwendigkeit und Angemessenheit der geleisteten Projektarbeiten.....	10
Voraussichtlicher Nutzen, fortgeschriebener Verwertungsplan	10
Während der Durchführung des Vorhabens bekannt gewordener Fortschritt auf dem Gebiet des Vorhabens bei anderen Stellen.....	10
Publikationen	10
AP C3: Entwurf und Optimierung der analogen neuromorphen Schaltungen und elektro- optischen Schnittstellen.....	11
Wissenschaftlich-technische Ergebnisse.....	11
Publikationen	13
AP C4: Software-Mapping und Neuromorphic Security	14
Wissenschaftlich-technische Ergebnisse.....	14
Die wichtigsten Positionen des zahlenmäßigen Nachweises	16
Notwendigkeit und Angemessenheit der geleisteten Projektarbeiten.....	16
Publikationen	16
AP C5: Anforderungen an Optimierungsmöglichkeiten durch Lokalität und Spärlichkeit	17
Wissenschaftlich-technische Ergebnisse.....	17
C5.1: Lernalgorithmen, die Spärlichkeit in Raum und Zeit ausnutzen.....	17
C5.2: Lokales Lernen durch Gradient-basierte Methoden.....	17
C5.3: Memristor-basierte Modelle des gradientenbasierten Lernens	19
C5.4: Lokale Lernalgorithmen für adaptive Kantenüberwachung und Steuerungsprobleme mit niedriger Leistungsaufnahme	19
C5.5: Nutzung von stochastischem Verhalten für industriell relevante probabilistische Inferenzprobleme	20
Die wichtigsten Positionen des zahlenmäßigen Nachweises	21

Notwendigkeit und Angemessenheit der geleisteten Projektarbeiten.....	21
Voraussichtlicher Nutzen, fortgeschriebener Verwertungsplan	22
Während der Durchführung des Vorhabens bekannt gewordener Fortschritt auf dem Gebiet des Vorhabens bei anderen Stellen.....	23
Publikationen	24
Zusammenfassung.....	26

Einleitung

Das Projekt NeuroSys hat sich der Herausforderung gestellt, innovative Technologien zu entwickeln und zu implementieren, die sowohl wissenschaftliche als auch praktische Lösungen für gesellschaftliche Herausforderungen bieten. In einer Zeit des rasanten Fortschritts im Bereich der künstlichen Intelligenz und neuromorpher Systeme zielt das Projekt darauf ab, die technologische Unabhängigkeit Europas zu stärken und gleichzeitig energieeffiziente Lösungen zu erforschen. Die Arbeit erstreckt sich über mehrere Teilprojekte, die von der Entwicklung effizienter Co-Entwurfsmethoden bis hin zur Optimierung neuromorpher Schaltkreise reichen. Durch enge Zusammenarbeit mit Industriepartnern und dem Einsatz modernster Technologien wird eine Brücke zwischen Forschung und praktischer Anwendung geschlagen. Im Folgenden findet sich ein detaillierter Überblick zu wissenschaftlich-technischen Arbeiten und Ergebnissen der einzelnen Unterprojekte von Projekt C. Sofern relevant werden auch weitere Aspekte erläutert in Bezug auf wichtige zahlenmäßige Positionen, Notwendigkeit und Angemessenheit der Arbeiten, den weiteren Verwertungsplan sowie relevante Veröffentlichungen von anderen Stellen. Am Ende jedes Teilprojektes befindet sich eine Übersicht über alle Veröffentlichungen der jeweiligen Forschungsgruppe.

AP C1: Quantitative Entwurfsraumexploration

Prof. Dr. Tobias Gemmeke / RWTH Aachen, IDS

Wissenschaftlich-technische Ergebnisse

Das Teilprojekt C1 setzt sich aus drei Arbeitspaketen zusammen, die eng miteinander verknüpft sind. Der erste Schritt hat sich mit der *Entwicklung einer Methodik für den effizienten Co-Entwurf* beschäftigt. Dazu wurden zunächst umfassend bestehende Simulationswerkzeuge und die von ihnen modellierten physikalischen Eigenschaften identifiziert und untersucht. Betrachtet wurden dabei die Modellierung verschiedener Nicht-Idealitäten, die Simulation der verschiedenen Hardware Metriken, die Unterstützung für Inferenz bzw. Training neuronaler Netze und die Anwendbarkeit auf bestehende Netzwerkimplementierungen. Mit dem Ziel, diese in eine gesamtheitliche Methodik mit Berücksichtigung algorithmischer und auch hardwareseitiger Effekte zu integrieren, wurden diese Werkzeuge daraufhin entsprechend ihrer Abstraktionsebenen einer von drei Kategorien zugeordnet (System, Crossbar und Device Level). Darauf basierend wurde eine mehrstufige Entwurfsmethodik entwickelt, die mit marktüblichen Werkzeugen sowohl auf Algorithmen- (PyTorch, Tensorflow) als auch der Hardware-Seite (Verilog, VHDL mit entsprechenden Simulatoren) erweitert wurde. Damit knüpft diese Methodik allgemeingültig an bekannte Techniken auf beiden Seiten an und stellt somit einen geeigneten Ansatz zur Co-Exploration dar, um effiziente Hardwarebeschleuniger zu entwickeln [1]. Parallel dazu begannen bereits die ersten Arbeiten zur Verwendung dieser Methodik in Hinblick auf die Anwendungen aus Projekt D.

Das zweite Arbeitspaket widmete sich dann vollständig der *Anwendung der Methodik auf Vehikel aus Projekt D* und wurde auf konzeptionell ähnliche Weise für verschiedene End-Anwendungen genutzt jedoch mit jeweils eigenen Optimierungen und Charakteristiken, die sich aus sehr unterschiedlichen Hardware- und Performanzanforderungen ergaben. Dieser Schritt beinhaltete zunächst die Ausführung der Netze auf der IT-Infrastruktur des IDS gefolgt von einer ausführlichen Entwurfsraumexploration, die hardware-definierte Optimierungen berücksichtigten. Spezifische Fallbeispiele werden in den nächsten Absätzen präsentiert. Diesem Schritt folgte der dritte Schritt, die *Konzeption Hardwarearchitektur mit Validierung und Charakterisierung*. In dieser wurden hochoptimierte Hardwarebeschleuniger für die optimierten Netze aus der vorherigen Exploration entworfen, validiert und charakterisiert. Der Methodik aus dem ersten Schritt folgend, wurden mehrere Optimierungszyklen durchgeführt, um einzelne Komponenten, die als kostenintensive Hardwareoperationen identifiziert wurden, weiter zu verbessern.

Als ein Anwendungsfall wurde die automatische Spracherkennung aus Projekt D2 betrachtet. Diese verwendet eine Conformer-Architektur, wobei es sich um eine Erweiterung der klassischen Transformer Architektur inklusive ihres Aufmerksamkeitsmechanismus mit Faltungsschichten handelt. Derzeitige Implementierungen für Grafikkarten sind kostenintensiv bezogen auf Chipfläche und Energiebedarf. Andererseits haben die einzelnen Komponenten solcher Conformer-Netze sehr unterschiedliche Anforderungen bezüglich Datenflusses und Dimensionsgrößen, welche durch die Flexibilität von Grafikkarten abgedeckt werden kann. Unsere Arbeiten betrachteten daher eine hardware-bewusste algorithmische Exploration. Dazu wurde eine Quantisierung der Gewichte und Aktivierungen auf 8 bit Ganzzahlwerte betrachtet und eine Sensitivitätsanalyse der Dimensionsgrößen für die Zwischenrepräsentationen exploriert. Anschließend wurden weitere Details für eine effiziente Implementierung von nicht-linearen Sub-Modulen untersucht. Dabei wurden, unter anderem, geteilte Skalierungsfaktoren für nicht-lineare

Funktionen und eine vereinfachte Normalisierungsfunktion analysiert, die auf die Berechnung der Wurzelfunktion verzichtet. Dadurch wird eine effiziente Hardwareimplementierung ermöglicht.

Als weiterer Ansatz wurde eine ebenfalls hochoptimierte Implementierung für tiefenweise separierbare Faltungsschichten betrachtet, welche eine Operation in Conformer-Netzen darstellt. Auch hier wurden entsprechende algorithmische Explorations durchgeführt und zwei verschiedene Architekturen implementiert, eine mit zwei getrennten Recheneinheiten für tiefenweise und punktweise Faltung und eine mit einer kombinierten Einheit [2, 4]. Zudem wurde eine weitere hardwarebasierte Optimierung auf der Algorithmenseite entwickelt, welche ein gruppenweise gleichmäßiges Pruning anwendet, um eine erhöhte Energieeffizienz durch hohe Hardwareausnutzung bei reduzierter Operationsanzahl zu erreichen [12].

In Kooperation mit Projekt D4 wurde zudem ein Anwendungsfall aus der Medizintechnik untersucht. Dazu wurde ein Ansatz mit verteiltem Lernen auf Elektrokardiogramm (EKG) Daten betrachtet, der eine erhöhte Datensicherheit ermöglicht, indem die sensiblen Daten auf verschiedene Recheneinheiten verteilt werden können [3]. Zusätzlich wurden für diese EKG-Anwendung Hardwarearchitekturen implementiert und analysiert. Ein Ansatz hat sich auf die kontinuierliche EKG-Überwachung mittels Subsampling-basierter Klassifikatoren konzentriert. Hierbei lag der Fokus auf einer besonders niedrigen Leistungsaufnahme für eine effiziente Integrierbarkeit in IoT-Geräte, was eine automatisierte Erkennung von Herzrhythmusstörungen wie Vorhofflimmern im Alltag ermöglicht [6].

Eine Folgearbeit hat sich mit der Integration von Domänengeneralisierungsfähigkeiten in tragbaren Geräten befasst, die neuronale Netze mit beschränkten Ressourcen ausführen sollen. Diese Technik adressiert die Herausforderung der Domänenverschiebung, die durch Variationen zwischen Trainings- und Einsatzbedingungen entsteht, und ermöglicht eine robuste Klassifikationsqualität über verschiedene Sensoren und Patienten hinweg durch den Einsatz von Korrekturschichten [10].

Die wichtigsten Positionen des zahlenmäßigen Nachweises

Neben den gewöhnlichen Personal- und Reisekosten wurde eine NVIDIA A10 Grafikkarte angeschafft, um die algorithmische Exploration inklusive ihrer Hardwareoptimierungen auf der IT-Infrastruktur des IDS ausführen zu können.

Notwendigkeit und Angemessenheit der geleisteten Projektarbeiten

Das Projekt C hatte zum Ziel, die Hardwareseite mit ihren verschiedenen neuen Ansätzen und die Anwendungsseite und ihren zum Teil sehr komplexen Algorithmen zu verbinden. Einer Methodik für einen effizienten Co-Entwurf beider Seiten kommt damit eine Schlüsselrolle zu, da sie Designentscheidungen auf einer der Ebenen ermöglicht, die durch Charakteristiken auf der jeweils anderen Seite vorgegeben werden. Die anschließende Verwendung dieser Methodik auf verschiedene Anwendungsfälle zeigt die Effektivität und Notwendigkeit eines solchen kombinierten Ansatzes in Hinblick auf eine energieeffiziente Ausführung aktueller Netzwerkkonstrukturen.

Voraussichtlicher Nutzen, fortgeschriebener Verwertungsplan

Die verwendete Methodik ist generisch gehalten und somit für verschiedene Anwendungsvoraussetzungen adaptierbar. Sie findet bereits in weiteren Forschungsarbeiten Anwendung. Zudem wurden bei den Kooperationen mit den Projektpartnern solche Netzwerke

verwendet, die derzeit auch in der Industrie vielseitig Anwendung finden. Die aus diesem Projekt hervorgegangen Hardwarearchitekturen und neuen Optimierungstechniken sind daher ein wichtiger Schritt hin zu einer energieeffizienteren Ausführung und der europäischen Unabhängigkeit in Bezug auf solche Technologien. Die Umsetzungsphase II von NeuroSys zielt daher bereits verstärkt auf den Wissenstransfer in die Wirtschaft ab.

Während der Durchführung des Vorhabens bekannt gewordener Fortschritt auf dem Gebiet des Vorhabens bei anderen Stellen

Das Forschungsfeld der neuronalen Netze hat sich in den letzten Jahren schnell weiterentwickelt und findet auch in der Industrie zunehmend breiter gestreut Anwendung. So wurden beispielsweise große Sprachmodelle entwickelt, die ebenfalls auf Transformerarchitekturen basieren. Für manche solcher Operationen wurden ebenfalls Hardwareimplementierungen veröffentlicht, während in der Praxis dennoch häufig flexible Grafikkarten und Tensorrecheneinheiten zur Anwendung kommen, die einen hohen Energiebedarf aufweisen. Dieses Projekt mit seinem Fokus auf die Verbindung von neuen physikalischen Bauelementen mit den Algorithmen auf sehr hohen Abstraktionsebenen deckt daher ein wichtiges Feld ab.

Publikationen

- [1] Wabnitz, Malte, and Tobias Gemmeke. "Toolflow for the Algorithm-Hardware Co-Design of Memristive ANN Accelerators." *Memories-Materials, Devices, Circuits and Systems* 5, 2023.
- [2] Chen, Yi, et al. "A Unified and Energy-Efficient Depthwise Separable Convolution Accelerator." *IFIP/IEEE International Conference on Very Large Scale Integration-System on a Chip*. Cham: Springer Nature Switzerland, 2023.
- [3] Ayad, Ahmad, et al. "PEACE: Private and Energy-Efficient Algorithm for Cardiac Evaluation on the EDGE using Modified Split Learning and Model Quantization." *2023 14th International Conference on Information and Communication Systems (ICICS)*. 2023.
- [4] Chen, Yi, et al. "EDEA: Efficient Dual-Engine Accelerator for Depthwise Separable Convolution with Direct Data Transfer." *IEEE 37th International System-on-Chip Conference (SOCC)*. IEEE, 2024.
- [5] Freye, Florian, et al. "Scaling Logic Area with Multi-Tier Standard Cells." *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits* (2024).
- [6] Loh, Johnson, and Tobias Gemmeke. "Stream Processing Architectures for Continuous ECG Monitoring using Subsampling-based Classifiers." *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* 32.1. 2023.
- [7] Latotzke, Cecilia, et al. "FPGA-based Acceleration of Lidar Point Cloud Processing and Detection on the Edge." *2023 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2023.
- [8] Kauth, Kevin, et al. "neuroAlx-Framework: Design of Future Neuroscience Simulation Systems Exhibiting Execution of the Cortical Microcircuit Model 20× faster than Biological Real-time." *Frontiers in Computational Neuroscience* 17. 2023.
- [9] Chen, Yi, et al. "An Energy-Efficient and Area-Efficient Depthwise Separable Convolution Accelerator with Minimal On-Chip Memory Access." *IFIP/IEEE 31st International Conference on Very Large Scale Integration (VLSI-SoC)*. 2023.
- [10] Loh, Johnson, et al. "Towards Hardware Supported Domain Generalization in DNN-based Edge Computing Devices for Health Monitoring." *IEEE Transactions on Biomedical Circuits and Systems*. 2024.
- [11] Kauth, Kevin, et al. "neuroAlx: FPGA Cluster for Reproducible and Accelerated Neuroscience Simulations of SNNs." *EEE Nordic Circuits and Systems Conference (NorCAS)*. 2023.

- [12]** Chen, Yi, et al. "GUPA: Group-Wise Uniform Pruning Accelerator for Depthwise Separable Convolution.", *IEEE Symposium in Low-Power and High-Speed Chips (COOL CHIPS)*. IEEE, 2025.
- [13]** Lou, Jie, et al., "An All-Digital Time-Domain Compute-in-Memory Engine for Convolutional Neural Networks in 22nm.", *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2025 (Accepted).
- [14]** Lou, Jie, et al. "A 22nm 96.83-TOPS/W Time-Domain Compute-in-Memory Engine Utilizing Mixed-Fidelity for Edge-AI Applications." *Great Lakes Symposium on VLSI 2025*. 2025. Vorläufiger Link: https://camps.aptaracorp.com/ACM_PMS/PMS/ACM/GLSVLSI25/15/26ecb61b-2be5-11f0-ada9-16bb50361d1f/OUT/glsvlsi25-15.html

AP C2: Entwurf und Entwurfsmethodik neuromorpher memristiver Schaltkreise

Prof. Dr. Stefan Heinen / RWTH Aachen, IAS

Wissenschaftlich-technische Ergebnisse

In Arbeitspaket C2 wurde zunächst mit den Partnern ELMOS und PGI-10 abgestimmt, wie die letzten Lagen im CMOS-Prozess für eine erfolgreiche Postprozessierung der Memristoren geschaffen sein mussten. Damit wurden entsprechende Lagen in der Entwurfssoftware für die Memristoren eingefügt und das Layout der Memristor-Standardzelle definiert. Ebenfalls mussten durch den speziellen Abschluss des CMOS-Prozesses einige Verbindungszellen von ELMOS im Layout modifiziert werden.

Zu den Standardzellen der Memristoren wurden Simulationsmodelle vom Partner IWE-2 hinterlegt, womit die Eigenschaften dieser Bauelemente in der Softwareumgebung für das Analogdesign untersucht wurden. Damit wurden die Spezifikationen und Funktionalität für das System und die Analogschaltungsblöcke definiert, mit denen man die Memristoren ansteuert. Hierbei standen die Flexibilität und der Integrationsgrad im Vordergrund, mit denen man die einzelnen Operationen in einer Matrix aus Memristoren durchführt. Im Anschluss wurden die Architektur der Matrix aus Memristoren und Transistoren entworfen sowie die einzelnen Analogblöcke der Peripherie. Ebenfalls musste ein digitaler Steuerungsblock entworfen werden, um den zeitlichen Ablauf der Operationen mit den Analogblöcken zu steuern, wozu auch eine digitale Schnittstelle gehört, um die Konfigurationsregister auf dem Chip mit einem externen Rechner zu setzen. Mit dem entworfenen System aus Treibern für die Zeilen und Spalten, inklusive der Digital-Analog-Wandler, kann jeder Memristor in der Matrix einzeln automatisiert geformt und auf einen fein justierbaren Wert programmiert werden. Im Anschluss kann eine Matrix-Vektor-Multiplikation durchgeführt werden.

Die Designs auf Transistorebene wurden nach erfolgreicher Validierung mit Simulationen auf ein Layout zusammen mit den Memristoren abgebildet. Dabei wurden zwei Versionen des Layouts erstellt, eine erste nur mit dem CMOS-Teil die den vollständigen CMOS-Prozess bei ELMOS durchlaufen ist und eine zweite mit den Memristoren, wo der CMOS-Prozess bei ELMOS gestoppt wurde, um anschließend die Memristoren vom PGI-10 prozessieren zu lassen.

Mit dem bei ELMOS vollständig prozessierten CMOS-Teil konnten die Funktionalität und die Performanz des Analog-Mixed-Signal Systems ohne den Einfluss der Postprozessierung bestimmt werden. Hierzu wurden die bei ELMOS gepackten Chips auf ein PCB gelötet und vermessen, wobei alle Schaltungsblöcke das erwünschte Verhalten zeigten. Währenddessen wurden die bei ELMOS gestoppten Wafer ans PGI-10 geliefert und dort die Memristoren prozessiert. Anschließend mussten diese Chips bei einem externen Partner in einem Package verbunden werden. Zurzeit werden diese Chips vermessen, wobei kein Einfluss der Postprozessierung auf die Analogblöcke erkannt wurde. Es stehen noch Messungen aus, um die Zuverlässigkeit der Programmierung der Memristoren zu bestimmen.

Parallel zu den Messungen wurde das System mit den Analogblöcken zum Auslesen der Ströme der Matrix-Spalten erweitert. Hierbei lag der Fokus auf einer parallelen Digitalisierung der Spaltenströme, um den zeitlichen Vorteil der Matrix-Vektor-Multiplikation mit Memristoren nicht zu verlieren, wie es bei den herkömmlichen seriellen Ausleseverfahren der Fall ist. Mit dieser Erweiterung der Bibliothek können hochskalierte Versionen des ersten Chips gefertigt werden.

Die wichtigsten Positionen des zahlenmäßigen Nachweises

Zur Charakterisierung der Operationen auf den Memristoren auf dem entwickelten Demonstrator-Chip wurde ein Oszilloskop mit besonders hoher Auflösung und Bandbreite beschafft.

Notwendigkeit und Angemessenheit der geleisteten Projektarbeiten

Um die Vorteile von Inferenzoperationen mit Memristoren auszunutzen, ist es unabdingbar diese mit integrierten Analogschaltungen zu betreiben die auf eine maximale Effizienz abgestimmt sind. Dafür bedarf es einer soliden Entwurfsinfrastruktur und eines zuverlässigen Technologiezugangs. Mit einer Bibliothek von Analogschaltungsblöcken können diese Systeme hochskaliert werden. Die Daten der Charakterisierung liefern die Grundlage für das Algorithmen-Hardware Co-Design.

Voraussichtlicher Nutzen, fortgeschriebener Verwertungsplan

Die Entwicklung und Fertigung eines Demonstrator-Chips bietet die Grundlage für weitere, verbesserte und hochskalierte Entwürfe, da viele Analogblöcke wiederverwendet werden können sowie die gewonnene Expertise beim Entwurf weiterer Blöcke von hoher Bedeutung ist. Die Messungen des Demonstrator-Chips dienen nicht nur dem Analogdesign größerer Systeme, sondern erlauben es, Modelle für einen hardware-nahen Algorithmen-Entwurf zu erstellen. Besonders bei hochskalierten Systemen ist die korrekte Modellierung der Hardware unumgänglich.

Während der Durchführung des Vorhabens bekannt gewordener Fortschritt auf dem Gebiet des Vorhabens bei anderen Stellen

Es wurden Systeme mit Memristor-Matrizen von deutlich größeren Dimensionen veröffentlicht, die auch schon verwendet werden, um Inferenzoperationen durchzuführen. Diese Systeme nutzen jedoch die analogen Eigenschaften der Memristoren kaum aus und leiden unter einem geringen Integrationsgrad der Peripherie.

Publikationen

[1] J. Zoche, R. Heinen, J. Grobe, R. Wunderlich and S. Heinen, "A Current-Mirroring Voltage Buffer for Analog Read-Out in Memristor Crossbar Arrays," 2024 19th Conference on Ph.D Research in Microelectronics and Electronics (PRIME), Larnaca, Cyprus, 2024, pp. 1-4, doi: 10.1109/PRIME61930.2024.10559676.

AP C3: Entwurf und Optimierung der analogen neuromorphen Schaltungen und elektro-optischen Schnittstellen

Prof. Dr. Renato Negra / RWTH Aachen, HFE

Wissenschaftlich-technische Ergebnisse

Im ersten Schritt wurden die Anforderungen an den elektronischen Chip definiert. Ein von einem Photodetektor stammender Eingangsstrom sollte anhand eines Transimpedanzverstärkers in ein Spannungssignal gewandelt werden, welches wiederum einen photonischen Modulator treibt. Dieses Ausgangssignal sollte zur akkuraten Signalverarbeitung bei bis zu 10 GHz getaktet sein. Außerdem sollte eine große Bandweite von DC bis 5 GHz im Signalpfad abgedeckt werden, und eine 1 pF Last mit 2 V Spitzenspannung bei dieser Frequenz modulierbar sein. Um diese Funktionalitäten zu gewährleisten, wurde ein System, wie in Abbildung 1 dargestellt, entworfen.

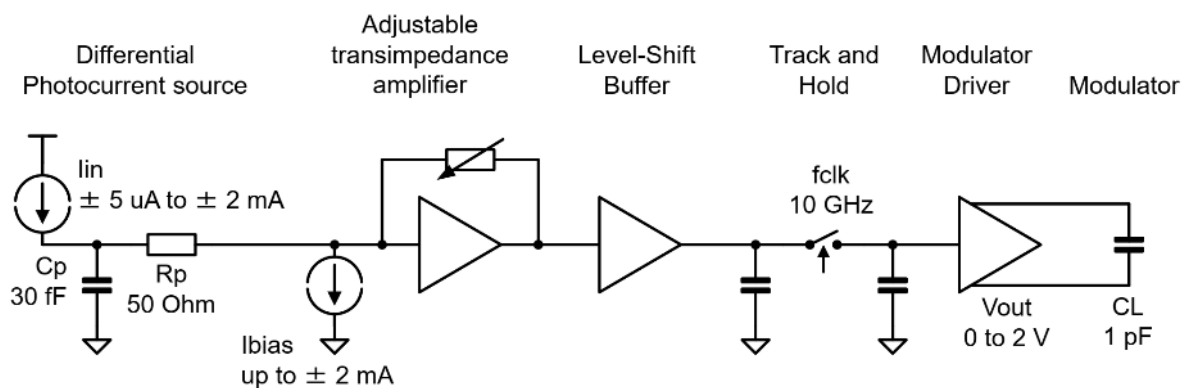


Abbildung 1 Anforderungen an den elektronischen Chip

Die neuronale Aktivierungsfunktion wurde als Sigmoid-Funktion inhärent effizient über die Sättigung der in der Schaltung vorhandenen Verstärker implementiert. Um diese Aktivierungsfunktion außerdem variabel zu gestalten, wurden ein Stromoffset und eine variable Verstärkung eingeführt.

Dieses System wurde auf Schaltungsebene in einem Chip in einer 65 nm CMOS-Technologie realisiert. Dabei wurde ein neuartiger Inverter-basierter Transimpedanzverstärker mit einstellbarer Verstärkung entwickelt. Die Taktung des Signals geschieht durch ein Transmission-Gate und der hohe Ausgangsspannungswert bei der gegebenen Last und Frequenz wird durch eine Modulator-Treiberschaltung sichergestellt.

Es wurden vier verschiedene Schaltungsversionen auf zwei elektronischen Chips produziert. Der erste Chip ist für das Flip-Chip Packaging mit dem photonischen Chip abgestimmt. Der zweite Chip ist elektronisch im Labor messbar, da die Kontaktierungen hier so angeordnet sind, dass er anhand eines Wafer-Probers kontaktiert und validiert werden kann. Abbildung 2 zeigt den Chip zur elektronischen Validierung.

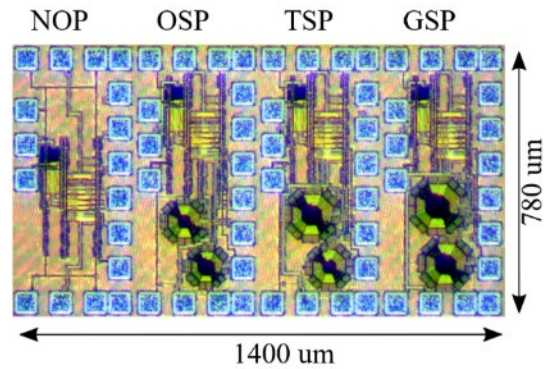


Abbildung 2 Chip zur elektronischen Validierung

Die vier Schaltungsvarianten sind hier markiert als NOP, OSP, TSP und GSP. Sie unterscheiden sich durch den Einsatz von Induktoren an unterschiedlichen Stellen in der Verstärkerkette. Dieser Chip wurde im Labor des HFE anhand eines Wafer-Probers und eines Netzwerkanalysators vermessen. Die gemessenen Transimpedanzen der verschiedenen Schaltungsvarianten, sowie mit unterschiedlichen Einstellungen der Verstärkung des Inverter-basierten Transimpedanzverstärkers sind in Abbildung 3 dargestellt.

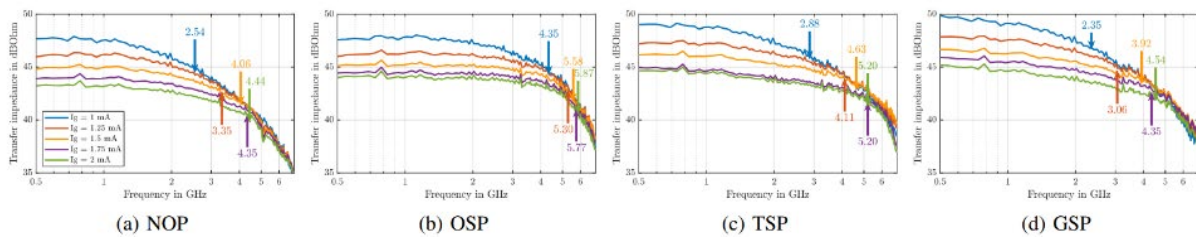


Abbildung 3 Gemessene Transimpedanzen der Schaltungsvarianten

Dieser Chip wurde auf der Konferenz MOCAS (International Conference on Modern Circuits and Systems Technologies) vorgestellt und veröffentlicht. Dort wurde die Arbeit ausgezeichnet als „Best Paper on Electronics, Circuits and Systems“.

Zur gleichen Zeit wurden am IPH neue Erkenntnisse bezüglich der photonischen Schaltung gewonnen, auf deren Grundlage neue Spezifikationen für die Entwicklung einer zweiten Iteration des elektronischen Chips definiert wurden. Nun war das Ziel das Stromeingangssignal gezielt über definierte Zeitbereiche zu integrieren und in ein gesättigtes Ausgangssignal, welches weiterhin die zuvor definierten Parametern erfüllen sollte, zu transformieren. Abbildung 4 zeigt das zur Realisierung dieser Anforderungen entworfene System.

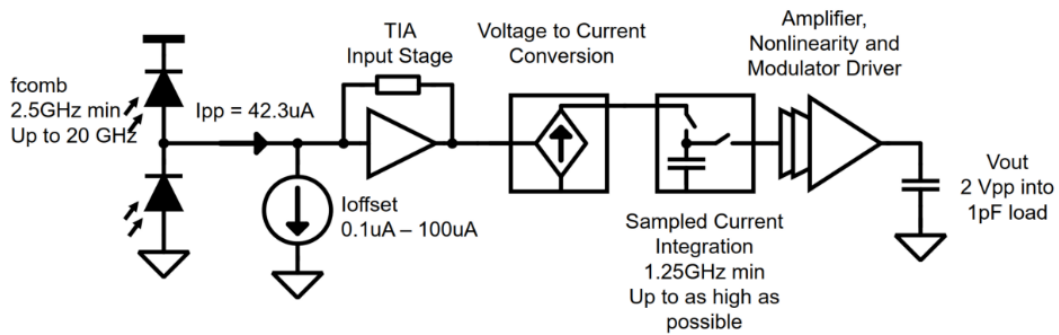


Abbildung 4 Entworfenes System zur Realisierung der Anforderungen in der zweiten Iteration

Einige Komponenten, wie der Transimpedanzverstärker als Eingangsstufe, die Stromoffset Stufe, die Transmission Gates oder der Modulator-Treiber, konnten basierend auf den Erfahrungen aus der ersten Iteration ausgelegt werden, während die Kernfunktion der diskreten Stromintegration über die Zeit neu entwickelt werden musste. Die neuronale Aktivierungsfunktion wird auch hier über die Sättigung der Verstärkerkette realisiert.

Die Schaltung wurde in derselben Technologie, wie die erste Iteration, entworfen und produziert. Auch wurden erneut zwei Chipversionen, eine für das Flip-Chip-Packaging und eine für die elektronische Validierung, produziert. Eine Aufsicht auf den Chip ist in Abbildung 5 dargestellt.

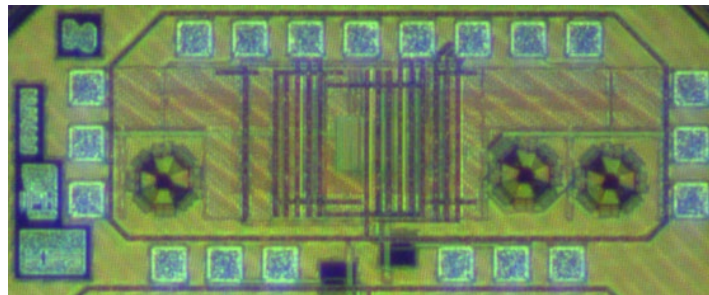


Abbildung 5 Aufsicht auf den elektronischen Chip der zweiten Iteration

Der Chip der zweiten Iteration wurde bereits vermessen und eine Publikation der Messergebnisse ist zeitnah geplant.

Publikationen

[1] Hüssen, Lukas, et al. "A DC to 5.8 GHz CMOS Variable-Gain Transimpedance Amplifier for Photonic Neuromorphic Hardware." *2024 13th International Conference on Modern Circuits and Systems Technologies (MOCAS)*. IEEE, 2024.

AP C4: Software-Mapping und Neuromorphic Security

Prof. Dr. Rainer Leupers / RWTH Aachen, SSS

Wissenschaftlich-technische Ergebnisse

Im Rahmen des Vorhabens wurde ein Software Development Kit (SDK) für Systems-on-Chip (SoCs) mit neuromorphen Hardware-Beschleunigern konzipiert und prototypisch umgesetzt. Die Arbeiten orientierten sich an den in der Vorhabenbeschreibung definierten Teilpaketen (C4.1.1 bis C4.1.6), wurden jedoch im Verlauf konkretisiert und erweitert, insbesondere im Hinblick auf eine realistische Evaluation auf unterschiedlichen Hardwarekonzepten. Im gesamten Arbeitspaket wurde der Fokus auf RRAM-basiertes Computing-in-Memory gelegt.

Die durchgeführten Schritte umfassten:

- Eine detaillierte Analyse bestehender ML-Frameworks und Definition einer Software-Zwischenrepräsentation (C4.1.1) und die Spezifikation eines anpassbaren SDKs, das zukünftige Erweiterungen erlaubt (C4.1.2).
- Die schrittweise Implementierung von drei SDK-Prototypen; von der Abbildung fester ML-Workloads auf spezifische Architekturen bis zur allgemeinen Kompatibilität auf beliebige Crossbars (C4.1.3–C4.1.5).
- Verifikation und Benchmark-basierte Evaluation des finalen SDKs auf unterschiedlichen Crossbars (C4.1.6). Da uns keine Crossbars vorlagen, die der Last eines ML-Workloads standgehalten hätten, wurde mit Crossbar-Simulatoren gearbeitet.

Im Verlauf des Projekts wurden alle definierten Arbeitspakete erfolgreich abgeschlossen. Die Ergebnisse wurden in fünf wissenschaftlichen Publikationen veröffentlicht, die wesentliche Aspekte der Entwicklung adressieren:

- Scheduling mit CLSA-CIM [1]
In dieser Arbeit wird ein Compiler-Algorithmus vorgestellt, der die Ausführungszeit von ML-Netzen auf RRAM-basierten CIM-Architekturen durch Cross-Layer-Scheduling signifikant reduziert. Beschleunigungen bis zu 29.2x konnten erreicht werden.
- Nutzung einer Messplattform für RRAM-Crossbars [2,3]
Ziel war das automatisierte Erheben von RRAM-Device Daten für unsere Crossbar Simulatoren. Diese konnten mit der entwickelten Plattform erfolgreich erhoben und im Simulator verwendet werden.
- Entwicklung eines Compilers für Computing-in-Memory Architekturen [4,5]
Diese Arbeit behandelt Synchronisationstechniken und Compilerkonzepte für Multi-Core-CIM-Systeme.
- Eine weitere Publikation befindet sich im Review-Prozess und behandelt das Hardware-Software Co-Design von beliebigen ML-Workloads auf beliebige Crossbars (C4.1.5-C4.1.6). Verschiedene Crossbargrößen, Mapping-Strategien und ML-Workloads können ausgewählt werden. Die entwickelten Tools schätzen dann die zu erwartende Genauigkeit des ML-Workloads ab. Die Tools können online getestet werden. [6,7]

Sicherheitsaspekte wurden im Rahmen des Arbeitspakets C4.2 detailliert untersucht.

Die definierten Arbeitspakete C4.2.1 bis C4.2.6 sahen sowohl die Erzeugung von Angriffsvektoren als auch eine Verteidigung gegen diese vor. Im Rahmen der Bearbeitung des Projektes wurde der

Fokus auf die Angriffsvektoren gelegt, da diese bisher sehr wenig erforscht wurden und eine grundlegende Analyse benötigten, bevor Verteidigungen entwickelt werden können.

Es wurden hauptsächlich neuartige Angriffsvektoren überprüft, die in klassischen CMOS-Systemen in dieser Art nicht vorhanden sein können und damit exklusiv für neuromorphe Systeme sind. Darüber hinaus wurde der Effekt bereits bekannter Gegenmaßnahmen für CMOS-Systeme beachtet.

Zunächst wurden Analog-zu-Digital-Wandler als mögliches Angriffsziel betrachtet. Diese sind Teil der kritischen Infrastruktur um die eigentlichen Memristoren der neuromorphen Hardware herum. Als analoge Schaltungen beinhalten sie auffällige Strukturen, die nicht durch existierende Techniken zur Verschlüsselung wie Logic Locking versteckt werden können. Insbesondere kann die Vergleichsspannung durch Trojaner leicht modifiziert werden. Der Effekt wurde auf einer repräsentativen Anwendung in Form des neuronalen Netzes VGG16 überprüft. Simulationsergebnisse zeigen, dass Angriffe auf conv2d-Layer durchaus erfolgreich sind und zu falschen Klassifizierungen und überhöhter Selbstsicherheit des Netzes führen. Wenn jedoch auch voll vernetzte Layer angegriffen werden, werden alle anderen Klassen stark unterdrückt. Dies macht den Angriff leicht erkennbar, und würde so zur Entdeckung des Trojaners führen. Ein erfolgreicher Trojaner müsste also verschiedene Layer auseinanderhalten können, was detailliertes räumliches und zeitliches Wissen über den Chip, den verwendeten Compiler und die laufende Anwendung benötigt. Dies erschwert eine derartige Attacke erheblich und macht neuromorphe Systeme resistenter und damit sicherer.

Neben Analog-zu-Digital-Wandlern besitzen Memristoren eine weitere Eigenschaft, die neue Angriffsfläche einführt: Ihr Energieverbrauch hängt von der Berechnung ab. Angreifer können dies ausnutzen, indem sie Stromverbrauch oder Temperatur messen und anhand der Messdaten auf die Berechnungen zurückschließen können.

Zunächst wurde die grundsätzliche Machbarkeit einer derartigen Seitenkanalattacke mithilfe eigens entwickelter Simulatoren überprüft. Dazu wurden Inferenzoperationen von neuronalen Netzen auf Memristor-basierter Hardware mithilfe von SystemC AMS TDF simuliert. Es wurden Leistungswerte ermittelt und anschließend in dem Wärmesimulator 3D-ICE übergeben. Anhand der so entstandenen Temperaturwerte ist es möglich, geistiges Eigentum in Form der Gewichte des neuronalen Netzes zu extrahieren [8].

Zusätzliche Komponenten zukünftiger neuromorpher Chips wie klassische CPU-Kerne oder auch die parallele Ausführung mehrerer Memristor-Crossbars könnten die Messdaten jedoch verfälschen. Hier könnte ein Angreifer einen Hardware-Trojaner einsetzen, um die zusätzlichen Komponenten selektiv abzuschalten.

Der Effekt dieses Ansatzes wurde für Memristor-Aided LOGIC (MAGIC) als besonders komplexes Ziel untersucht. Bei MAGIC werden boolesche Funktionen direkt in der neuromorphen Hardware realisiert. Memristoren bilden sowohl Eingabe als auch Ausgabe der einzelnen Logikgatter. Bei der Berechnung von MAGIC Gattern wird der jeweilige Ausgabe-Memristor je nach Ausgabewert umgeschrieben, was Energie benötigt. Im Rahmen des Projekts wurde mit SPICE-Simulation gezeigt, dass dieser Energieverbrauch durch Angreifer messbar ist. Werden störende Faktoren mithilfe eines Hardware-Trojaners deaktiviert, so ergibt dies sehr klare Daten. Auf dieser Basis wurde ein Algorithmus entworfen, implementiert und verifiziert, der die Struktur und Logik der booleschen Funktion aus den Daten extrahieren kann. Die Größe der notwendigen Datenbasis ist sehr klein. Ein solcher Angriff ist aufgrund des Funktionsprinzips nicht einfach zu verhindern. Bekannte Gegenmaßnahmen wie Logic Locking sind nicht effektiv, da Memristor-Crossbars

grundsätzlich immer sichtbar und damit angreifbar für Hardware-Trojaner sind. Auch der bedingte Energieverbrauch der Berechnungen ist unumgänglich bei Verwendung von Memristoren [9].

Die wichtigsten Positionen des zahlenmäßigen Nachweises

Die Mittel wurden gemäß Zuwendungsbescheid für die Personalaufwendungen im Bereich Softwareentwicklung, Evaluierung und Publikation verwendet.

Notwendigkeit und Angemessenheit der geleisteten Projektarbeiten

Die im Projekt erarbeiteten Konzepte und Umsetzungen waren notwendig, um die Lücke zwischen vorhandenen ML-Frameworks und neuartigen CIM-Architekturen zu schließen. Die erarbeiteten Lösungen adressieren bestehende Defizite, etwa bei der Scheduling-Effizienz oder bei der Hardware-Kompatibilität, und wurden in realitätsnahen Evaluierungen validiert.

Publikationen

[1] Pelke, Rebecca, et al. "CLSA-cim: a cross-layer scheduling approach for computing-in-memory architectures." 2024 Design, Automation & Test in Europe Conference & Exhibition (DATE). IEEE, 2024.

[2] Pelke, Rebecca, et al. "A Fully Automated Platform for Evaluating ReRAM Crossbars." 2024 IEEE 25th Latin American Test Symposium (LATS). IEEE, 2024.

[3] Pelke, Rebecca, et al. "The show must go on: a reliability assessment platform for resistive random access memory crossbars." Philosophical Transactions A 383.2288 (2025): 20230387.

[4] Pelke, Rebecca, et al. "Architecture-Compiler Co-design for ReRAM-Based Multi-core CIM Architectures." IFIP/IEEE International Conference on Very Large Scale Integration-System on a Chip. Cham: Springer Nature Switzerland, 2023.

[5] Pelke, Rebecca, et al. "Mapping of CNNs on multi-core RRAM-based CIM architectures." 2023 IFIP/IEEE 31st International Conference on Very Large Scale Integration (VLSI-SoC). IEEE, 2023.

[6] Simulator and mapping strategies: <https://github.com/rpelke/analog-cim-sim>

[7] Design space exploration tool and pre-compiled neural networks: <https://github.com/rpelke/CIM-E>

[8] Pfeifer, Lorenzo, et al. "Analysis of Thermal Side-Channel Attacks on Analog/Digital Computing-in-Memory Accelerators" 2024 IEEE 25th Latin American Test Symposium (LATS). IEEE, 2024.

[9] Pfeifer, Lorenzo, et al. "EXAMINER: IP Extraction Algorithm from MAGIC Logic-in-Memory" 2025 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)

AP C5: Anforderungen an Optimierungsmöglichkeiten durch Lokalität und Spärlichkeit

Prof. Dr. Markus Diesmann / FZJ, INM-6

Prof. Dr. John Paul Strachan / FZJ, PGI-14

Prof. Dr. Emre Neftci / FZJ, PGI-15

Wissenschaftlich-technische Ergebnisse

C5.1: Lernalgorithmen, die Spärlichkeit in Raum und Zeit ausnutzen

C16: Erstellung von Konzepten zur Ausnutzung von Spärlichkeit in Raum und Zeit und Anwendung von lokalem Lernen zur Echtzeitüberwachung und -steuerung (D4) (M36)

Im Zuge dieses Projekts wurde ein strukturbewusstes Neuronenverteilungsschema in Kombination mit einem neuen Spike-Kommunikationsschema für große Multi-Areal-Netzwerk Simulationen entwickelt. Die Implementierung der strukturbewussten Lastverteilung und Kommunikation basierend auf unserem Referenzcode NEST wurde umfassend evaluiert, wofür weiterhin der Hochleistungsrechner JURECA zum Einsatz kam. Die Ergebnisse zeigen eine deutliche Beschleunigung der Simulation und eine bessere Skalierbarkeit bei Verwendung der neuen Methode. Insbesondere die Zeit, die während der Spike-Kommunikation zur Synchronisation zwischen den Rechenknoten in Anspruch genommen wird, wurde stark reduziert.

Das eigens für Skalierungsexperimente entwickelte, gut kontrollierbare Benchmark-Netzwerkmodell wurde adaptiert, sodass es bezüglich Neuronenzahl, Konnektivität und Aktivität mit dem Multi-Areal Modell aus Schmidt et al. (2018)¹ übereinstimmt. Das Benchmarkmodell kann somit genutzt werden, um über das Multi-Areal Modell von Schmidt et al. (2018) hinaus die Effizienz der neuen Technik unter unterschiedlichen Randbedingungen umfassend zu beurteilen. Insbesondere Unterschiede in Arealgrößen und Verzögerung der Signalübertragung innerhalb und zwischen Arealen wurden damit systematisch untersucht.

Anhand der Simulationsdaten konnten theoretische Modelle zur Beschreibung der Effizienz von Kommunikation und Datentransfer im Netzwerk hergeleitet werden.

Eine automatisierte Benchmarking-Pipeline ermöglicht die systematische Erhebung umfangreicher Messdaten zur Beurteilung der Effizienz der neuen Simulationstechnologie. Dieses Projekt hat dabei aufgrund seiner komplexen Ansprüche maßgeblich zur Weiterentwicklung der Pipeline beigetragen.

C5.2: Lokales Lernen durch Gradient-basierte Methoden

C13: Technische Umsetzung der entworfenen lokalen Lernregeln (M12)

Die Anforderungen an die Hardware-Implementierung von Lernregeln wurden als Teil der Graphcore IPU-Implementierung des Spiking Neural Network evaluiert.

¹ Schmidt M., Bakker R., Shen K., Bezgin G., Diesmann M., van Albada S.J. (2018) A multi-scale layer-resolved spiking network model of resting-state dynamics in macaque visual cortical areas. PLoS Computational Biology 14:e1006359. DOI: 10.1371/journal.pcbi.1006359

Die Arbeit im Zusammenhang mit der SNN-Simulation auf der Graphcore IPU wurde auf der AAAI veröffentlicht und zeigt eine beeindruckende 20-fache Beschleunigung im Vergleich zu GPUs beim Training von SNNs mit spärlichen Aktivitäten. Diese Arbeit war das Ergebnis einer Zusammenarbeit mit Graphcore.

Darüber hinaus haben wir an einem Spiking Neural Network Simulator gearbeitet, der auf dem JAX-Framework für maschinelles Lernen basiert und SNNAX heißt. Die Vorteile von JAX liegen in der Geschwindigkeit dank der Just-in-Time-Kompilierung und in der großen Flexibilität bei der Gradientenverarbeitung. Letzteres ermöglicht im Prinzip die Berechnung lokaler synaptischer Plastizitätsaktualisierungen (z.B. e-prop) aus fundamentalen Prinzipien. Die Software enthält mehrere vollständig trainierbare Modelle für spikende Neuronen, darunter das adaptive exponentielle LIF, sowie Standard-Benchmarks wie die Heidelberg spoken digits und die DVS Gestures-Datensätze. Die Bibliothek wurde der Öffentlichkeit zugänglich gemacht².

C14: Konzepte für die technische Implementierung von Backpropagation durch die Zeit und neurosynaptisches Trace-Based Lernen mit lokalen Lernregeln (M24)

Ein Konzept für lokales Gradientenabstiegslernen auf Basis von Drei-Faktor-Regeln sowie rekurrentem Lernen in Echtzeit wurde durch Übersetzung zeitbasierter Gewichtsadjustierungen in ereignisbasierte entwickelt (C5.2.1a). Konzepte zur Implementierung einer lokalen Approximation der Rückpropagation durch die Zeit, der Eligibility-Propagation-Algorithmus (e-prop), eine Form des Surrogat-Gradienten-Lernens wurden ausgearbeitet (C5.2.1b) und in NEST implementiert mit mehreren Surrogat-Gradienten-Funktionen³ (C5.2.1c). Validierung erfolgte durch Reproduktion von Ergebnissen aus der Originalpublikation⁴. Der Algorithmus wurde mit Hinblick auf Hardware-Tauglichkeit entwickelt (C5.2.2a). Die Nützlichkeit spärlicher N-Schritt-Approximationen⁵ wurde untersucht (C5.2.2b). Die Effizienz wurde durch die Verwendung eines Mean-Squared-Error-Verlustes anstelle von Softmax⁶ (C5.2.2c) optimiert. Entkopplung von Zeitkonstanten ergab eine lokalisiertere Lernregel mit gleichbleibender Lernperformanz (C5.2.3a). Das Modell wurde in großen spärlichen Netzen eingesetzt (C5.2.3b). Die Implementierung wurde dokumentiert und ein Manuskript ist in Arbeit (C5.2.4). Wie Lernalgorithmen, die dendritische Dynamik von Neuronen ausnutzen können (C5.2.5), wurde nicht untersucht, da sich dieser Themenkomplex als eigenständiges Projekt herausstellte.

² Lohhoff, Finkbeiner, Neftci. SNNAX - Spiking Neural Networks in JAX, <https://arxiv.org/pdf/2409.02842>

³ Neftci, E. O., Mostafa, H., & Zenke, F. (2019). Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. *IEEE Signal Processing Magazine*, 36 (6), 51–63. <https://doi.org/10.1109/msp.2019.2931595>

⁴ Bellec, G., Scherr, F., Subramoney, A., Hajek, E., Salaj, D., Legenstein, R., & Maass, W. (2020). A solution to the learning dilemma for recurrent networks of spiking neurons. *Nature Communications*, 11 (1), 3625. <https://doi.org/10.1038/s41467-020-17236-y>

⁵ Menick, J., Elsen, E., Evci, U., Osindero, S., Simonyan, K., & Graves, A. (2020). A Practical Sparse Approximation for Real Time Recurrent Learning. *arXiv*. <https://doi.org/10.48550/arXiv.2006.07232>

⁶ Hui, L., & Belkin, M. (2021). Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks. *arXiv*. <https://doi.org/10.48550/arXiv.2006.07322>

C5.3: Memristor-basierte Modelle des gradientenbasierten Lernens

C15: Demonstration der lokalen Memristor-Dynamiken für lokales Lernen in simulierten Geräten bei Klassifikationsaufgaben (M24)

Unter Verwendung eines kompakten Modells von Oxid-basierten filamentär-schaltenden Memristor-Updates (VCM) haben wir die Ursachen der Schreibstochastizität und ihre Auswirkungen auf Klassifizierungsprobleme untersucht⁷. Dabei haben wir festgestellt, dass das Rauschen, das den SET/RESET-Abgleich stört, die Leistung des Netzwerks am stärksten beeinträchtigt. Damit ist der Meilenstein C15 abgeschlossen. Das Modell und seine Bewertung wurden auf dem IBM AIHWKIT-Simulator implementiert.

Als Nächstes haben wir mithilfe von Learning-to-Learn, einer Form der zweistufigen Optimierung, gezeigt, wie hochgradig nicht-ideale Memristor-Verhaltensaktualisierungen durch systematisches Lernen von Hyperparametern kompensiert werden können⁸. Ähnlich wie in unserer veröffentlichten Arbeit⁹ könnte dieser Ansatz verwendet werden, um das Online-Lernen auf Edge-Geräten weitgehend zu vereinfachen, vorausgesetzt, die Netzwerkleitwerte können präzise initialisiert werden.

Im Hinblick auf eine Implementierung in Hardware-Memristoren haben wir uns vorgenommen, die Effizienz der Gewichtsinitialisierung zu verbessern, ein Prozess, der im Allgemeinen durch teure Schreibüberprüfungsansätze erfolgt. Hier haben wir sowohl synthetische als auch experimentelle Leitwert-Aktualisierungsdaten verwendet, um ein neuronales Netz zu trainieren, das die idealen Programmierbedingungen für den Memristor ausgibt¹⁰. Unsere Arbeit zeigt, dass die Programmierung bei geringen Kosten für die Programmiergenauigkeit bis zu 50-mal schneller sein kann als bei den Standardansätzen zur Schreibüberprüfung (siehe auch C5.5).

Um diese Arbeit weiter voranzutreiben, haben wir ein Modell eines prädiktiven steuerungs-basierten Programmieransatzes für Memristoren in der Simulation untersucht. Ein mit synthetischen und experimentellen memristiven Leitfähigkeitsdaten trainiertes neuronales Netzwerk dient als prädiktive Modell für die Memristordynamik. Die resultierenden Vorhersagen wurden genutzt, um die Programmierbedingungen in Echtzeit zu optimieren. Dieser Ansatz ermöglichte eine präzisere Steuerung des Systems, sodass die simulierten Memristor-Bauelemente innerhalb von etwa 5 bis 10 Impulsen Zielwerte innerhalb von $\pm 5 \mu\text{s}$ erreichen konnten.

C5.4: Lokale Lernalgorithmen für adaptive Kantenüberwachung und Steuerungsprobleme mit niedriger Leistungsaufnahme

C16: Erstellung von Konzepten zur Ausnutzung von Spärlichkeit in Raum und Zeit und Anwendung von lokalem Lernen zur Echtzeitüberwachung und -steuerung (D4) (M36)

In diesem Arbeitspaket haben wir die Verwendung von autoregressiven Decoder-Only-Transformern für lokale Lernalgorithmen untersucht. Autoregressive Transformers wie GPT sind zu Schlüsselkomponenten der modernen Sprachverarbeitung, Bildverarbeitung und Robotersteuerung geworden.

⁷ Yu *et al.* 2022, 2022 56th Asilomar Conference on Signals, Systems, and Computers

⁸ Yu, Zhenming, Nathan Leroux, and Emre Neftci, 2022 International Electron Devices Meeting (IEDM). IEEE, 2022.

⁹ Stewart and Neftci, 2022, Neuromorphic Computing and Engineering 2.4 (2022): 044002.

¹⁰ Yu, Zhenming, et al. "The Ouroboros of Memristors: Neural Networks Facilitating Memristor Programming." arXiv preprint arXiv:2403.06712 (2024).

Der Selbstaufmerksamkeitsmechanismus des Transformers erfordert jedoch bei jedem Zeitschritt die Übertragung vorheriger Token-Projektionen aus dem Hauptspeicher (DRAM), was seine Leistung auf herkömmlichen Prozessoren stark einschränkt.

Selbstaufmerksamkeit kann als dynamische Feedforward-Schicht betrachtet werden, deren Matrix von der Eingabesequenz abhängt, ähnlich wie das Ergebnis lokaler synaptischer Plastizität.

Wir haben die Ineffizienz von Transformern auf herkömmlichen Prozessoren betrachtet, indem wir uns von den Ergebnissen von C5.2 und C5.3 inspirieren ließen. Konkret haben wir die Selbstaufmerksamkeitsoperation auf Intel Loihi übertragen, einen digitalen neuromorphen Prozessor, der Spiking Neural Networks (SNNs) und programmierbare synaptische Plastizität implementiert. Wir haben den Plastizitätsprozessor (lokales Lernen) des Loihi-Chips verwendet, um die Dynamik zu implementieren. Wir haben unseren Ansatz anhand der Lösung eines Few-Shot-Vision-Klassifizierungsproblems demonstriert. Ein Konferenzbeitrag wurde 2024 eingereicht und auf der AICAS 2025 Konferenz vorgestellt.

Diese Forschungsrichtung wich zwar leicht von unserem ursprünglichen Plan ab, wurde jedoch durch die unerwarteten Auswirkungen von Transformer-Modellen und die Möglichkeit, deren Effizienz durch neuromorphe Hardware zu steigern, motiviert. Diese Arbeit war entscheidend für die Festlegung der Forschungspläne für NeuroSys Phase 2.

Die Mechanismen, mit denen ein solches Selbstaufmerksamkeitsmodell in memristiven Systemen implementiert werden könnte, wurden parallel dazu im Neurotec2-Projekt untersucht, in dem wir einen analogen Kern für Selbstaufmerksamkeit entwerfen.

C5.5: Nutzung von stochastischem Verhalten für industriell relevante probabilistische Inferenzprobleme

C17: Evaluierung des Einsatzes von künstlichen stochastischen Neuronen/Synapsen-Systemen in einem effizienten Hardware-Systemdesign zur Berechnung probabilistischer Inferenz in Zeitreihendaten (M36)

Ziel dieses Projekts ist die Bewertung des Leistungsgewinns durch stochastische Modellierungsmethoden für Hardware-synapsen oder Gewichte neuronaler Netze. Durch die Modellierung der Schaltvorgänge von memristiven Bauelementen als Zeitreihenprozesse nutzen wir probabilistische Algorithmen (z. B. sequentielle Monte-Carlo-Verfahren) und tiefe neuronale Netze, um die Programmierung eines Arrays von memristiven Bauelementen im Nanomaßstab zu verbessern. Eine effiziente Abstimmung ist entscheidend, um maschinelles Lernen und KI-Workloads, wie z. B. die Beschleunigung von CNN-Inferenzen auf memristiver Hardware, bei deutlich geringeren Abstimmungskosten im Vergleich zu Basisverfahren, zu ermöglichen.

Aufbauend auf einem statistischen Memristor-Modell, das mit einem Partikelfilter implementiert wurde¹¹, haben wir ein tiefes neuronales Netzwerk (DNN) trainiert, um den optimalen Abstimmimpuls vorherzusagen. Dieser Ansatz reduziert die Abstimmzeit um zwei Größenordnungen im Vergleich zur Basismethode¹².

¹¹ M.-J. Yang and J. P. Strachan, "State-space modeling and tuning of memristors for neuromorphic computing applications," Proceedings of the 2023 International Conference on Neuromorphic Systems (ICONS'23).

¹² Zhenming Yu, Ming-Jay Yan¹² M.-J. Yang and J. P. Strachan, "State-space modeling and tuning of memristors for neuromorphic computing applications," Proceedings of the 2023 International Conference on Neuromorphic Systems (ICONS'23)

Zhenming Yu, Ming-Jay Yang, ...E. Neftci and J. P. Strachan, "The Ouroboros of Memristors: Neural Networks Facilitating Memristor Programming," 2024 IEEE 6th International Conference on AI Circuits and Systems (AICAS'24)

Im Jahr 2024 haben wir diese Fortschritte genutzt und das Deep Reinforcement Learning (DRL) erforscht, um die Steuerung der Memristorabstimmung weiter zu optimieren. Dazu gehört die Verbesserung der Hardware-Machbarkeit durch die Vorhersage der optimalen Impulsamplitude anstelle der Impulsbreite, die durch die Betriebsfrequenz von Mikrocontrollern eingeschränkt ist. Mit unserer DRL-basierten Abstimmungstechnik haben wir die synaptischen Gewichte eines Faltungsneuronalen Netzes (CNN) erfolgreich auf ein Memristor-Crossbar-Array abgebildet und eine hochpräzise Bildklassifizierung auf einem programmierten Memristor-Chip demonstriert, wie in Yang et al. 2024¹³ berichtet. Die detaillierte Methodik für das Training von DRL mit einem physikalisch basierten statistischen Gerätemodell wurde auf einer internationalen Konferenz vorgestellt und beim IEEE veröffentlicht¹⁴.

Die wichtigsten Positionen des zahlenmäßigen Nachweises

In C5.1 bis C5.5 wurden entsprechend der Projektbewilligung insgesamt drei Doktoranden eingestellt, die zur Erreichung der Meilensteine C13 bis C16 beigetragen haben. In C5.5 trug ein Postdoktorand zur Erreichung des Meilensteins C17 bei. Für die angefallenen Reisen wurden teilweise Projektreisemittel verwendet. Dem Projekt standen keine Sachmittel zur Verfügung.

Notwendigkeit und Angemessenheit der geleisteten Projektarbeiten

C5.1: Für den Vergleich von Simulationsergebnissen mit experimentellen Daten sind oft großflächige Netzwerksimulationen erforderlich, die typischerweise hohe Rechenressourcen und einen erheblichen Energieeinsatz in Hochleistungsrechenzentren erfordern. Die im Projekt erzielten Fortschritte verbessern die Skalierbarkeit und Effizienz der Simulation großer Multi-Areal-Gehirnnetzwerke deutlich und stärken damit die Verbindung zwischen Modellierung und Experiment. Effizientere Simulationen solcher hierarchischen Netzwerke ermöglichen deutlich längere Beobachtungszeiträume, die insbesondere für die Untersuchung des Zusammenspiels von Netzwerkstruktur und Plastizität unerlässlich sind. Zudem können größere Parameterräume erforscht werden.

C5.2: Mit der entwickelten Implementierung steht nun die performanteste überwachte Lernregel im NEST-Simulator zur Verfügung – geeignet sowohl für Regressions- als auch für Klassifikationsaufgaben. Dabei kommen moderne Optimierungsmethoden wie der Adam-Optimizer¹⁵, Feuerraten-Regularisierung sowie verschiedene Surrogat-Gradienten-Funktionen zum Einsatz. Darüber hinaus erlaubt die neue Lernregel erstmals das effektive Training großskaliger, puls-gekoppelter rekurrenter neuronaler Netzwerke im NEST-Simulator – ein zentraler Meilenstein für realitätsnahe neuronale Modellierung. Gleichzeitig leisten die Ergebnisse einen Beitrag zur Entwicklung energieeffizienter Lernalgorithmen, die insbesondere für den Einsatz auf neuromorpher Hardware von Bedeutung sind. Schließlich schafft die Arbeit

Ming-Jay Yang, Zhenming Yu, Emre Neftci and J. P. Strachan, "Learning and Tuning Memristor Dynamics with Deep Neural Networks," 2024 - International Conference on Neuromorphic Computing and Engineering (ICNCE'24)

Ming-Jay Yang, Zhenming Yu, Giacomo Pedretti, Emre Neftci and J. P. Strachan, "Improved Memristor Control using Device Physics and Deep Reinforcement Learning", 2025 IEEE 7th International Conference on AI Circuits and Systems (AICAS'25)

¹³ Ming-Jay Yang, Zhenming Yu, Emre Neftci and J. P. Strachan, " Learning and Tuning Memristor Dynamics with Deep Neural Networks," 2024 - International Conference on Neuromorphic Computing and Engineering (ICNCE'24)

¹⁴ Ming-Jay Yang, Zhenming Yu, Giacomo Pedretti, Emre Neftci and J. P. Strachan, "Improved Memristor Control using Device Physics and Deep Reinforcement Learning", 2025 IEEE 7th International Conference on AI Circuits and Systems (AICAS'25)

¹⁵ Kingma, D. P., & Ba, J. (2017). Adam: A method for stochastic optimization. <https://doi.org/10.48550/arXiv.1412.6980>

eine wichtige Grundlage für das systematische Erfassen und Erforschen biologischer Lernmechanismen in komplexen Netzwerken.

C5.3: Diese Arbeit verbessert den Memristor-Programmierprozess gegenüber der herkömmlichen Methode des Schreibens und Überprüfens. Aktuelle memristive KI-Beschleuniger sind in erster Linie für Inferenzaufgaben konzipiert, vor allem weil die bestehenden Programmiermethoden für Memristor-Bauelemente nicht effizient genug sind. Herkömmliche Schreib- und Verifizierungsverfahren erzielen zwar eine hohe Präzision, sind jedoch zu langsam, um On-Chip-Lernen oder Feinabstimmungen zu unterstützen. Durch die Verbesserung sowohl der Geschwindigkeit als auch der Genauigkeit der Memristor-Programmierung ebnet unser Ansatz den Weg für breitere Anwendungen und ein größeres Potenzial von Memristor-basierten KI-Systemen.

C5.4: Diese Arbeit befasste sich mit der dringend notwendigen Verbesserung der Energieeffizienz von Transformer-Modellen. Durch den Nachweis, dass solche Modelle in digitaler neuromorpher Hardware implementiert werden können, ebnet wir den Weg für spezielle ereignisbasierte Schaltungen und Systeme, die solche Workloads auf Edge-Geräten, z. B. im Bereich des Internets der Dinge (IoT), ausführen können. Diese Arbeit war entscheidend dafür, dass wir den Industriepartner Mercedes Benz für die Zusammenarbeit am NeuroSys-Folgeprojekt interessieren konnten.

C5.5: Die verbesserte Steuerung von nanoskaligen memristiven Bauelementen in Crossbar-Arrays ermöglicht eine Vielzahl von In-Memory-Computing-Anwendungen und unterstreicht die vielfältigen Möglichkeiten für die Integration physikalischer Modelle in maschinelle Lernverfahren. So zeigt beispielsweise der mit unserem Surrogate Gerät Modell trainierte neuronale Netzwerkcontroller die Fähigkeit, die analoge Leitfähigkeit an Kreuzungspunkten unter Verwendung optimaler Impulsbedingungen abzustimmen, wodurch eine mindestens zweifache Verbesserung der Abstimmungsleistung (z. B. Anzahl der Impulse) im Vergleich zur grundlegenden Schreib- und Verifizierungsmethode (ISPP) erzielt wird.

Voraussichtlicher Nutzen, fortgeschriebener Verwertungsplan

C5.1: Die Arbeit diente als Pilotprojekt und zeigte, dass durch eine strukturangepasste Lastverteilung die Rechenzeit und der Kommunikationsaufwand in großskaligen Netzwerksimulationen deutlich reduziert werden können. Die entwickelte Methode ist in den Referenzsimulator NEST integriert und steht somit unmittelbar für die Anwendung und Weiterentwicklung zur Verfügung. Modellierer großer neuronaler Netzwerke können direkt von dieser Implementierung profitieren, indem sie ihre Simulationen beschleunigen und Ressourcen effizienter nutzen. Die Ansätze sind zudem auf andere Modelle übertragbar und bilden eine Grundlage für weitere Optimierungen.

C5.2: Die Ergebnisse des Projekts bieten der NEST-Community einen unmittelbaren praktischen Nutzen. Durch bereitgestellte Tutorials kann die implementierte Lernregel direkt für eigene Lernexperimente genutzt werden, was in mehreren Projekten bereits erfolgreich geschehen ist. Darüber hinaus eignet sich die Implementierung als Referenz und Schablone für die Entwicklung verwandter, biologisch plausibler Lernregeln in NEST. Dadurch wird nicht nur der Zugang zu lokalem Lernen erleichtert, sondern auch die Weiterentwicklung des Simulators im Bereich des maschinellen Lernens unterstützt.

C5.3: Diese Arbeit verbessert den Memristor-Programmierprozess und legt den Grundstein für eine breitere Anwendung von Memristor-basierten KI-Beschleunigern. Eine schnellere Memristor-Programmiermethode könnte auch den Weg für On-Chip-Training und Feinabstimmung ebnen.

C5.4: Der in diesem Arbeitspaket vorgestellte Benchmark zeigte die Anwendbarkeit ereignisbasierter Computersysteme für das Online-Lernen in einem Few-Shot-Szenario. Die Few-Shot-Lernfähigkeit ist weitreichend anwendbar in Steuerungssystemen und Überwachungssystemen (z. B. Anomalieerkennung) und gleichzeitig robust gegenüber Datenbereichsverschiebungen.

C5.5: Diese Arbeit stellt einen Ansatz für eine effiziente und wiederholbare Anpassung der Leitfähigkeit an Zielwerte vor, der breit anwendbar ist für In-Memory-Computing-Aufgaben sowohl im Offline- als auch im Online-Modus, einschließlich kontinuierlichem Lernen, Signalverarbeitungskernen und wissenschaftlichem Rechnen.

Während der Durchführung des Vorhabens bekannt gewordener Fortschritt auf dem Gebiet des Vorhabens bei anderen Stellen

C5.1: Die Ansätze, die in diesem Projekt entwickelt wurden, fanden direkte Anwendung in der parallelen Entwicklung der GPU-basierten Version des Simulators NEST. Insbesondere das Konzept der strukturgepassten Lastverteilung wurde übernommen. Die Ergebnisse des Projekts leisten damit einen wichtigen Beitrag zur Vorbereitung auf die Nutzung kommender Exascale-Systeme wie JUPITER.

C5.2: Während der Projektlaufzeit wurden auch Fortschritte ähnlicher Arbeiten in anderen Simulationsumgebungen und auf Hardwareplattformen bekannt. So wurde e-prop beispielsweise in den Simulator GeNN¹⁶ sowie auf neuromorphen Plattformen wie SpiNNaker^{17 18} und ReckOn¹⁹ implementiert. Diese parallelen Entwicklungen unterstreichen die Relevanz des Themas und zeigen das wachsende Interesse an effizienten, biologisch motivierten Lernverfahren in der internationalen Forschungslandschaft. Die Resultate aus unserer Arbeit wurden von diversen Forschungsgruppen (Bonn, Sussex in Großbritannien, Aas in Norwegen, und Jülich) aufgegriffen und dienen als Grundlage für neue Projekte.

C5.3: Diese Arbeit stellt eine effiziente Memristor-Programmiermethode vor, bei der ein Warteschritt („Reverse Forming“) eingeführt wird, der es den internen Filamenten des Memristors ermöglicht, sich in einen stabileren Zustand zu setzen. Durch Multiplexing über verschiedene Geräte innerhalb der Programmierpipeline kann die durch diesen Warteschritt verursachte

¹⁶ Knight, J. C., & Nowotny, T. (2022). Efficient GPU training of LSNNs using eprop. *Neuro-Inspired Computational Elements Conference*, 8–10. <https://doi.org/10.1145/3517343.3517346>

¹⁷ Perrett, A., Summerton, S., Gait, A., & Rhodes, O. (2022). Online learning in SNNs with e-prop and neuromorphic hardware. *Proceedings of the Annual Neuro-Inspired Computational Elements Conference (NICE)*, 32–39. <https://doi.org/10.1145/3517343.3517352>

¹⁸ Rostami, A., Vogginger, B., Yan, Y., & Mayr, C. G. (2022). E-prop on SpiNNaker 2: Exploring online learning in spiking RNNs on neuromorphic hardware. *Frontiers in Neuroscience*, 16, 1018006. <https://doi.org/10.3389/fnins.2022.1018006>

¹⁹ Frenkel, C., & Indiveri, G. (2022). Reckon: A 28nm sub-mm² task-agnostic spiking recurrent neural network processor enabling on-chip learning over second-long timescales. *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, 65, 1–3. <https://doi.org/10.1109/isscc42614.2022.9731734>

Verzögerung effektiv ausgeblendet werden, was zu minimalen Auswirkungen auf die Gesamtschreiblatenz führt²⁰.

C5.4: In dieser Arbeit wurde eine Variante des Zustandsraummodells auf dem Intel Loihi implementiert, die eine schnelle und energieeffiziente Inferenz im sequenziellen Betrieb demonstriert. Unser Ansatz unterscheidet sich in seiner Implementierung durch die Verwendung des Plastizitätsprozessors sowie durch die Anwendung (Few-Shot-Lernen)²¹.

C5.5: Diese Arbeit identifiziert die Ursachen für Leitfähigkeitsschwankungen in Memristoren durch experimentelle und theoretische Untersuchungen und stellt ein elektrisches Betriebsprotokoll zur Unterdrückung von Rauschen für einen hochpräzisen Betrieb vor. Sie zeigt, dass in einem einzelnen Memristor 2.048 unterschiedliche Leitfähigkeitsstufen erreicht werden können²².

Aufbauend auf den obigen Arbeiten wurde eine innovative Schaltungsarchitektur und ein Programmierprotokoll demonstriert, die eine effiziente Programmierung von inhärent ungenauen analogen Geräten mit beliebig hoher Präzision ermöglichen - begrenzt nur durch die Grenzen der digitalen Periphereschaltungen²³.

Publikationen

C5.1

[1] M. Lober, M. Diesmann, S. Kunkel, "Optimizing communication in brain-scale multi-area model simulations", 2024 - International Conference on Neuromorphic Computing and Engineering (ICNCE'24)

[2] M. Lober, M. Diesmann, S. Kunkel, "Exploiting network topology in brain-scale multi-area model simulations", 2024 - Federation of European Neuroscience Societies (FENS'24)

C5.2

[3] Korcsak-Gorzo, A., Linssen, C., Albers, J., Dasbach, S., Duarte, R., Kunkel, S., Morrison, A., Senk, J., Stapmanns, J., Tetzlaff, T., Diesmann, M., & van Albada, S. J. (2024). Phenomenological modeling of diverse and heterogeneous synaptic dynamics at natural density. In J. H. R. Lübke & A. Rollenhagen (Eds.), *New aspects in analyzing the synaptic organization of the brain* (pp. 277–321). Springer US. <https://doi.org/10.1007/978-1-0716-4019-7>

C5.3

[4] Yu, Z., Menzel, S., Strachan, J. P., & Neftci, E. (2022, October). Integration of Physics-Derived Memristor Models with Machine Learning Frameworks. In *2022 56th Asilomar Conference on Signals, Systems, and Computers* (pp. 1142-1146). IEEE.

[5] Yu, Z., Leroux, N., & Neftci, E. (2022, December). Training-to-learn with memristive devices. In *2022 International Electron Devices Meeting (IEDM)* (pp. 21-1). IEEE.

C5.4 & C5.5

[6] M.-J. Yang and J. P. Strachan, "State-space modeling and tuning of memristors for

²⁰ Z. Jiang et al., "COPS: An Efficient and Reliability-Enhanced Programming Scheme for Analog RRAM and On-Chip Implementation of Denoising Diffusion Probabilistic Model," 2023 International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2023, pp. 1-4, doi: 10.1109/IEDM45741.2023.10413764.

²¹ Meyer, Svea Marie, et al. "A Diagonal Structured State Space Model on Loihi 2 for Efficient Streaming Sequence Processing." arXiv preprint arXiv:2409.15022 (2024).

²² Rao, M., Tang, H., Wu, J. et al. Thousands of conductance levels in memristors integrated on CMOS. *Nature* 615, 823–829 (2023). <https://doi.org/10.1038/s41586-023-05759-5>

²³ Wenhao Song et al., Programming memristor arrays with arbitrarily high precision for analog computing. *Science* 383, 903-910 (2024). DOI:10.1126/science.adi9405

neuromorphic computing applications,” Proceedings of the 2023 International Conference on Neuromorphic Systems (ICONS’23).

[7] Zhenming Yu, Ming-Jay Yang, ...E. Neftci and J. P. Strachan, “The Ouroboros of Memristors: Neural Networks Facilitating Memristor Programming,” 2024 IEEE 6th International Conference on AI Circuits and Systems (AICAS’24)

[8] Ming-Jay Yang, Zhenming Yu, Emre Neftci and J. P. Strachan, ” Learning and Tuning Memristor Dynamics with Deep Neural Networks,” 2024 - International Conference on Neuromorphic Computing and Engineering (ICNCE’24)

[9] Ming-Jay Yang, Zhenming Yu, Giacomo Pedretti, Emre Neftci and J. P. Strachan, “Improved Memristor Control using Device Physics and Deep Reinforcement Learning”, 2025 IEEE 7th International Conference on AI Circuits and Systems (AICAS’25)

Zusammenfassung

Das Projekt NeuroSys war ein umfassender Erfolg in der Entwicklung neuer technologischer Ansätze, die sowohl wissenschaftlich als auch praktisch relevant sind. Eine Vielzahl von innovativen Methoden wurde entwickelt, darunter effiziente Co-Entwurfsstrategien für Hardwarebeschleuniger und neuromorphe Schaltkreise sowie die Implementierung eines Software Development Kits für Systems-on-Chip mit neuromorphen Beschleunigern und die Implementierung eines e-prop-Algorithmus im NEST-Simulator. Die Ergebnisse haben nicht nur zur wissenschaftlichen Weiterentwicklung beigetragen, sondern auch den Grundstein für wirtschaftliche Anwendungen gelegt, wie etwa durch die Gründung des Spin-offs RooflineAI GmbH. Zahlreiche Publikationen unterstreichen die Bedeutung der erzielten Fortschritte auf internationaler Ebene. Das Projekt hat erfolgreich gezeigt, wie moderne KI-Technologien mit energieeffizienten Lösungen kombiniert werden können, um den Strukturwandel in Europa voranzutreiben und gleichzeitig den Schutz sensibler Daten sicherzustellen.