Kurzbericht

Titel des Projekts: **NeuSPIN – Entwurf eines zuverlässigen neuromorphen Edge-Systems basierend auf Spintronik für Green AI**

Zuwendungsempfänger: Karlsruher Institut für Technologie (KIT),

Projektleitung: Prof. Dr. Mehdi Tahoori

Förderkennzeichen: 01IS21090

Laufzeit des Vorhabens: 01.10.2021 - 30.09.2024

1. Ursprüngliche Aufgabenstellung und wissenschaftlich-technischer Hintergrund

Das Projekt *NeuSPIN* befasste sich mit einer der zentralen Herausforderungen der modernen Informationstechnologie: der energieeffizienten Ausführung von Künstliche Intelligenz (KI) (-Algorithmen. Klassische Hardwarearchitekturen, die auf dem Von-Neumann-Prinzip basieren, sind zunehmend ungeeignet für KI-Anwendungen, da sie Rechen- und Speichereinheiten räumlich und funktional trennen. Dies führt zu ineffizienten Datenbewegungen („Von-Neumann Bottleneck“) und hohem Energieverbrauch – ein kritischer Faktor insbesondere bei Edge-Geräten mit begrenzter Energieversorgung.

Ein vielversprechender Ansatz zur Überwindung dieser Limitationen ist das Konzept des *Neuromorphic Computings*, das sich an biologischen Strukturen des Gehirns orientiert. In *Computing-in-Memory*-Architekturen (CIM) werden Rechen- und Speicheroperationen in einheitlichen Bauelementen ausgeführt, was eine drastische Reduktion der Datenbewegung und Energieaufnahme erlaubt. Insbesondere Spintronik-basierte *Non-Volatile Memories* (NVM) eignen sich aufgrund ihrer nichtflüchtigen, stochastischen und skalierbaren Eigenschaften für solche Architekturen.

Ziel von *NeuSPIN* war es, eine robuste Plattform für KI-Anwendungen zu entwickeln, die diese Technologien nutzt. Dabei sollte die inhärente Stochastizität von Spintronic-Bauelementen nicht als Störgröße, sondern als Vorteil für bayessche, probabilistische KI-Modelle verwendet werden. Solche Modelle sind besonders geeignet für Anwendungen, bei denen Unsicherheiten abgeschätzt und Fehlerrobustheit garantiert werden müssen – etwa im medizinischen Umfeld oder im autonomen Fahren.

2. Ablauf des Vorhabens

Das Projekt wurde in enger Zusammenarbeit zwischen dem *Chair of Dependable Nano Computing* (CDNC) am Karlsruher Institut für Technologie (KIT) und der *Université Grenoble-Alpes* durchgeführt. Es umfasste folgende Arbeitsschritte:

- **Technologieauswahl und Bauelement-Modellierung:** Entwicklung und Charakterisierung spintronischer Speicherzellen, die als künstliche Neuronen und Synapsen dienen.
- **Architektur-Design:** Entwurf von neuromorphen CIM-Architekturen unter Nutzung der entwickelten Zellen.
- **Algorithmus-Hardware-Co-Design:** Anpassung von KI-Algorithmen (insbesondere bayessche neuronale Netze) an die stochastischen Eigenschaften der Bauelemente.
- **Implementierung und Evaluation:** Aufbau von Demonstratoren, Durchführung von Simulationen, und Validierung durch prototypische Anwendungen.

Besonderes Augenmerk lag auf dem Training und der Inferenz unter Berücksichtigung nicht-idealer Effekte wie Prozessvariabilität und Schaltungsrauschen. Hier wurden neue Methoden des unsicherheitsbewussten maschinellen Lernens entwickelt und in Hardware umgesetzt.

3. Wesentliche Ergebnisse

Die zentrale Innovation von *NeuSPIN* liegt in der systematischen Nutzung von spintronischer Hardware zur Umsetzung von bayesschen neuronalen Netzen, die Unsicherheiten explizit modellieren und verarbeiten können. Die wesentlichen Ergebnisse umfassen:

- **SpinDrop:** Ein neuartiger dropout-basierter Trainingsansatz, der die stochastischen Eigenschaften von Spintronik ausnutzt. (IEEE Journal On Emerging and Selected Topics In Circuits and Systems (JETCAS), Best Paper Award)
- **SpinBayes:** Eine Co-Design-Methodik, die Algorithmen und Hardware gemeinsam optimiert, um Unsicherheitsabschätzungen effizient zu ermöglichen.
- **NeuSPIN-Edge-System:** Ein energiesparsames, zuverlässiges Edge-KI-System, das auf CIM-Architektur basiert und Bayessche Modelle hardware-effizient implementiert.
- **Inverted Normalization:** Ein innovatives Konzept zur Normalisierung neuronaler Aktivierungen unter hardwarebedingten Schwankungen.

Die Forschungsergebnisse wurden in zahlreichen hochrangigen internationalen Publikationen und Konferenzen präsentiert – darunter IEEE Design and Test Automation in Europe (DATE), European Test Conference (ETS), ACM Transactions on Embedded Computing Systems (TECS) und IEEE Transactions on Nanotechnology (TNANO). Sie erhielten mehrere Auszeichnungen und wurden von der Fachwelt breit rezipiert.

4. Internationale Zusammenarbeit

Ein entscheidender Erfolgsfaktor war die enge Kooperation mit der *Université Grenoble-Alpes*, insbesondere mit den Gruppen von Prof. Guillaume Prénat und Prof. Lorena Anghel. Durch die Kombination der Kompetenzen in Schaltungstechnik, Spintronik, Systementwurf und KI-Algorithmen konnte ein einzigartiger interdisziplinärer Forschungsansatz realisiert werden. Die partnerschaftliche Zusammenarbeit ermöglichte zudem die Ausbildung junger Wissenschaftler:innen im Grenzbereich zwischen Hardware und KI.

5. Fazit und Ausblick

Das Projekt *NeuSPIN* hat gezeigt, dass stochastische Eigenschaften moderner NVM-Technologien – bislang als Herausforderung betrachtet – gezielt genutzt werden können, um zuverlässige, energieeffiziente und skalierbare KI-Systeme zu realisieren. Damit leistet *NeuSPIN* einen wichtigen Beitrag zur Entwicklung von „Green KI“, insbesondere für ressourcenbeschränkte Edge-Geräte.

Die im Projekt entwickelten Methoden und Architekturen bilden die Grundlage für weiterführende Forschungsvorhaben, u. a. im Rahmen europäischer Initiativen zur KI-Hardware sowie im Kontext sicherheitskritischer Anwendungen. Das Projekt stärkte zudem die europäische Forschungskooperation im Bereich neuromorpher Systeme und verdeutlichte das Innovationspotenzial interdisziplinärer Zusammenarbeit.

Schlussbericht

Titel des Projekts: **NeuSPIN – Entwurf eines zuverlässigen neuromorphen Edge-Systems basierend auf Spintronik für Green AI**

Zuwendungsempfänger: Karlsruher Institut für Technologie (KIT),

Projektleitung: Prof. Dr. Mehdi Tahoori

Förderkennzeichen: 01IS21090

Laufzeit des Vorhabens: 01.10.2021 - 30.09.2024

1. Einleitung und Ausgangslage

Projektzusammenfassung

Der Einsatz von Künstlicher Intelligenz (KI) in Verbindung mit Cloud-Computing verursacht derzeit rund 76 % des gesamten Datenzentrumsverkehrs. Prognosen zufolge wird sich dieser Anteil zwischen 2016 und 2021 verfünffachen. Bis 2025 dürfte er nahezu 20 % des weltweiten Stromverbrauchs ausmachen. Gleichzeitig erfordert das Aufkommen neuer Anwendungen im Bereich der personalisierten Gesundheitsversorgung – insbesondere durch das Internet der Dinge (IoT) und intelligente Wearables – die Speicherung und Verarbeitung immer größerer Datenmengen in Systemen mit extrem niedrigem Energieverbrauch.

Neuronale Netzwerke (NNs), eine Schlüsseltechnologie der KI, haben viele Anwendungsbereiche revolutioniert – insbesondere in der Mustererkennung, Entscheidungsfindung und prädiktiven Analytik. Ihre Vorhersagen sind jedoch häufig übermäßig selbstsicher, selbst bei fehlerhaften oder unbekannten Eingabedaten (Out-of-Distribution, OOD). Bayesianische Neuronale Netzwerke (BayNNs), eine spezielle Klasse von NNs, adressieren diese Schwäche, indem sie Unsicherheiten in ihren Vorhersagen quantifizieren können. Diese Fähigkeit ist entscheidend für sicherheitskritische Anwendungsfelder wie autonomes Fahren, medizinische Bildverarbeitung oder industrielle Wartung. Die Umsetzung von BayNNs stellt jedoch erhebliche Herausforderungen an die Hardware – insbesondere in klassischen von-Neumann-Architekturen –, da dort Recheneinheiten und Speicher getrennt sind (Memory Wall). Diese Trennung begrenzt Effizienz und Skalierbarkeit in realen Hochlastszenarien erheblich.

Neuromorphes Rechnen – das die Funktionsweise des Gehirns durch direkte Implementierung von neuronalen Netzen und synaptischen Arrays nachbildet – bietet demgegenüber überlegene Verarbeitungsfähigkeiten für KI-Anwendungen. Die Integration neuromorpher Hardware in „Green KI“-Ansätze reduziert grundlegend den Datenverkehr zur Cloud und ermöglicht ein neues Rechenparadigma, das die gesamte Information Technology (IT)-Landschaft verändern und die Entwicklung wirklich autonomer Systeme vorantreiben kann.

„Computing-in-Memory“ (CIM) stellt eine Hardware-Umsetzung des neuromorphen Ansatzes dar, bei der Rechenoperationen wie Multiply-and-Accumulate (MAC) direkt im Speicher durchgeführt werden – also dort, wo sich die Daten befinden. Dadurch wird die

Memory Wall überwunden. Besonders vielversprechend ist die Integration von spintronikbasierten Technologien in CIM-Architekturen: Sie ermöglichen extrem niedrigen Stromverbrauch, hohe Lebensdauer und effiziente Datenverarbeitung und sind zudem kompatibel mit bestehenden Halbleiterprozessen. Dennoch bringt diese neue Technologieklasse Herausforderungen mit sich: Produktionsschwankungen, Defekte und das inhärente stochastische Verhalten spintronischer Speicher erschweren insbesondere die zuverlässige Implementierung von BayNNs auf CIM-Hardware.

Das Projekt NeuSPIN hatte sich zum Ziel gesetzt, KI-Algorithmen, die bisher in der Cloud ausgeführt wurden, auf On-Chip-Verarbeitung mit hoher Genauigkeit und minimalem Energieverbrauch zu übertragen. Hierfür wurde Spintronik als vielversprechende nichtflüchtige Speichertechnologie (Non-Volatile Memory, NVM) eingesetzt. Im Zentrum standen neuartige Neuronen- und Synapsenentwürfe in Kombination mit BayNNs als Kern-KI-Modell. Durch einen disziplinübergreifenden Cross-Layer-Ansatz wurden sowohl Materialstapel und Bauelementparameter KI-spezifisch optimiert als auch technikbewusste Trainingsalgorithmen entwickelt.

Die Zusammenarbeit zwischen dem Karlsruher Institut für Technologie (KIT) und der Université Grenoble Alpes (UGA), zwei renommierten Forschungseinrichtungen mit komplementärer Expertise in den Bereichen Zuverlässigkeit, Spintronik, neuromorphe Schaltungen und KI, zielt darauf ab, eine exzellente europäische Partnerschaft im Bereich Edge-AI zu etablieren. Das Projekt positioniert Deutschland und Frankreich – und damit Europa – an der Spitze der Hardware-gestützten KI-Entwicklung im Edge-Bereich.

Hintergrund und Motivation

Cloud-Computing ist mittlerweile für etwa 76 % des weltweiten Rechenzentrumsverkehrs im KI-Bereich verantwortlich. Prognosen zeigen, dass sich dieser Anteil durch das zunehmende Datenvolumen, getrieben durch IoT und tragbare Gesundheitstechnologien, weiterhin drastisch erhöhen wird – mit signifikanten Auswirkungen auf den globalen Energieverbrauch. Der Bedarf an dezentraler, energieeffizienter und gleichzeitig leistungsfähiger KI-Verarbeitung – insbesondere am Rand des Netzes („Edge“) – steigt rapide.

Neuromorphes Computing, das die Struktur und Funktion des menschlichen Gehirns mittels neuronaler Netzwerke und synaptischer Arrays nachbildet, eröffnet neue Wege für energieeffiziente KI-Anwendungen. Durch Integration solcher Strukturen direkt in die Hardware kann auf energieintensive Datenübertragungen zur Cloud verzichtet werden. Die Umsetzung auf Bauelementebene ist jedoch bislang limitiert durch den hohen Hardwareaufwand pro Neuron bzw. Synapse, insbesondere bei CMOS-basierten Architekturen.

Das Projekt **NeuSPIN** setzt hier an. Es adressiert diese Herausforderung durch den Einsatz neuartiger, nichtflüchtiger Speichertechnologien (Spintronik), innovativer Neuronen- und Synapsendesigns sowie technologieangepasster Trainingsalgorithmen. Ziel ist die Entwicklung eines voll integrierten, bio-inspirierten KI-Systems, das für Edge-Anwendungen geeignet ist und gleichzeitig eine signifikante Reduktion des Energieverbrauchs ermöglicht.

Mehrwert der Zusammenarbeit

Das Projekt *NeuSPIN* stellte eine bedeutende Gelegenheit für beide Partner dar, ihre jeweilige wissenschaftliche Expertise zu bündeln und gemeinsame Lösungen zu entwickeln. Damit konnten wesentliche Beiträge zur Generierung von geistigem Eigentum (Intellectual Property, IP) geleistet werden, die eine Grundlage für einen möglichen Technologietransfer in die Industrie in der Zukunft bilden.

Das französische „Multidisciplinary Institute in Artificial Intelligence“ (MIAI) in Grenoble diente als zusätzlicher Kanal zur Verbreitung der Projektergebnisse innerhalb der wissenschaftlichen Gemeinschaft. Dies ermöglichte die Anbahnung neuer interdisziplinärer Kooperationen – insbesondere mit Forschungsgruppen im Bereich *Künstliche Intelligenz (KI) für Gesundheit* – sowie mit den mehr als 70 industriellen Partnern des Instituts.

Darüber hinaus brachten die Konsortialpartner ihre Expertise in die akademische Lehre ein. Durch die Beteiligung an Lehrveranstaltungen auf Master-, Ingenieur- und Doktorandenniveau wurde direkt zur Ausbildung einer neuen Generation hochqualifizierter Fachkräfte beigetragen. Ein erheblicher Teil der Projektmittel wurde gezielt für die Förderung von Doktorandinnen und Doktoranden sowie Postdoktorandinnen und Postdoktoranden eingesetzt, wodurch nachhaltiger wissenschaftlicher Nachwuchs aufgebaut werden konnte.

Zudem wurde die institutionelle Zusammenarbeit zwischen dem Karlsruher Institut für Technologie (KIT) und der Université Grenoble Alpes (UGA) – Grenoble Institut National Polytechnique (Grenoble INP) gestärkt. Beide Universitäten sind Mitglieder des europäischen Universitätskonsortiums *CLUSTER* (Consortium Linking Universities of Science and Technology for Education and Research; https://cluster.org). In diesem Rahmen wurden der Studierendenaustausch, die Beteiligung an Forschungsprojekten auf Bachelor- und Masterniveau sowie die Entwicklung von Doppelabschlussprogrammen weiter ausgebaut. Perspektivisch wird auch die Einrichtung gemeinsamer, thematisch mit *Künstlicher Intelligenz (KI)* verknüpfter Promotionsprogramme für die an dem Projekt beteiligten Doktorandinnen und Doktoranden in Erwägung gezogen.

Hauptziele

Ziel 1: Entwicklung eines disziplinübergreifenden neuromorphen KI-Frameworks für energieeffizientes Edge Computing

NeuSPIN verfolgt das Ziel, ein robustes Hardware-Software-Co-Design-Framework zu entwickeln, das auf Spintronik-Technologie basiert. Im Fokus steht die Co-Optimierung aller Schichten – von Materialien und Bauelementen bis hin zu Algorithmen und Systemarchitektur. Die Innovationen erfolgen auf mehreren Ebenen:

- **Technologieebene**: Optimierung von Spintronic-Stack-Parametern für geringeren Energieverbrauch bei gleichzeitig stabiler Funktionalität. Dazu gehört u. a. die Erhöhung des Gesamtwiderstands zur Stromreduktion und Entwicklung statistischer Gerätemodelle.
- **Algorithmenebene**: Entwicklung statistischer Trainingsverfahren auf Basis von Bayesianischen neuronalen Netzen, um inhärente Bauelementvariabilität zu kompensieren und die Inferenzgenauigkeit zu maximieren.
- **Systemebene**: Integration der Hardwarekomponenten zu einem leistungsfähigen Edge-KI-System, das datenlokale Verarbeitung („In-Memory Computing") realisiert und eine hohe Energieeffizienz bei gleichzeitig hoher Genauigkeit erzielt.

Ziel 2: Aufbau einer bilateralen Spitzenpartnerschaft zwischen Deutschland und Frankreich

Durch die enge Zusammenarbeit zwischen KIT und SPINTEC wird eine führende Forschungsallianz im Bereich neuromorpher KI-Hardware für Edge-Anwendungen etabliert. Beide Partner bringen komplementäre Expertisen ein:

- **SPINTEC (France)**: Materialentwicklung, Spintronic-Technologie, Bauelemente, Schaltungsdesign
- **KIT (Deutschland)**: Zuverlässiger Entwurf, neuromorphes Computing, KI-Algorithmik und Trainingsmethoden

Diese interdisziplinäre Partnerschaft wird durch gemeinsame Publikationen, Personal- und Studierendenaustausch sowie durch Vorarbeiten in Europäische Union (EU)- und Agence Nationale de la Recherche – Deutsche Forschungsgemeinschaft (ANR-DFG)-Projekten gestützt. Die Resultate des Projekts werden anhand realistischer medizinischer Anwendungsfälle validiert.

Bezug zu förderpolitischen Zielen

NeuSPIN ist eng mit den nationalen und europäischen Strategien zur Förderung von KI und Mikroelektronik verknüpft:

- **Stärkung der europäischen Innovationskraft**: NeuSPIN schafft ein vollständiges Hardware-Software-Ökosystem für neuromorphe KI auf Basis von Spintronik, wodurch Deutschland und Frankreich technologisch führend positioniert werden.
- **Clusterbildung und Technologietransfer**: Durch die Verbindung exzellenter nationaler Forschungscluster in den Bereichen zuverlässiges Rechnen (KIT) und Spintronik (SPINTEC) entsteht ein starker Nukleus für europäische Innovationsnetzwerke.
- **Energieeffiziente KI**: Durch In-Memory-Verarbeitung und nichtflüchtige Gewichtsspeicherung soll eine 100- bis 1000-fache Energieeinsparung im Vergleich zu heutigen Lösungen (Graphics Processing Unit (GPU), Complementary Metal-Oxide-Semiconductor – Application-Specific Integrated Circuit (CMOS-ASICs) wie TrueNorth oder Tianjic) erreicht werden.
- **Technologische Souveränität**: NeuSPIN leistet einen Beitrag zur Unabhängigkeit Europas im Bereich KI-Hardware – insbesondere angesichts geopolitischer

Herausforderungen und technologischer Abhängigkeiten von außereuropäischen Anbietern.

- **Gesellschaftliche Verantwortung und Nachhaltigkeit**: Die entwickelten Technologien adressieren ethische Fragen zur Vertrauenswürdigkeit und ökologischen Auswirkungen eingebetteter KI-Systeme. Damit werden zentrale Zukunftsfragen rund um Green KI und verantwortungsvolle Technologieentwicklung adressiert.

Kontext und Stand der Technik

Neuromorphes Rechnen

Deep Learning stellt den Stand der Technik in der künstlichen Intelligenz dar, von der Bildverarbeitung bis hin zur Übersetzung und Spracherkennung. Derzeit werden die meisten KI-Berechnungen in der Cloud durchgeführt und nur sehr wenig am Rande. Daher können eingebettete KI-Lösungen mit extrem niedrigem Stromverbrauch, hoher Verarbeitungskapazität und Autonomie nicht einfach durch die Übersetzung und Optimierung bestehender Cloud-basierter KI-Anwendungen in eingebettete Lösungen erreicht werden. Bei tiefen neuronalen Netzen (DNN) und rekurrenten Netzen wie Long Short-Term Memory, den beiden populären Ansätzen, sind die dominierenden Rechenoperationen niedrigpräzise Matrix-Vektor-Multiplikationen (MVMs) einer Gewichtsmatrix mit einem Eingangsvektor (Abb. 1a).

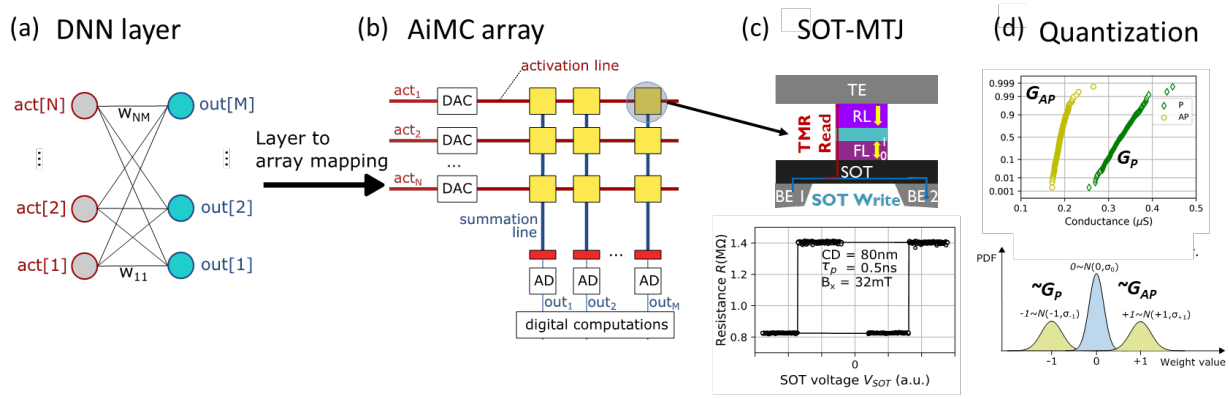

Abb. 1: a) Deconvolutional Neural Network (DNN)-Schicht, implementiert als b) ein analoges Vektor-Matrix-Multiplikator-Mapping-Array, c) SOT-MRAM-Zelle, die wir als hochohmigen Gewichtsspeicher vorschlagen, der für AiMC benötigt wird, d) P- und AP-MTJ-Leitwertverteilungen, transformiert in quantisierte Gewichte (einschließlich statistischer Variation, die für eine genaue Inferenzauswertung minimiert werden muss).

Daher besteht die primäre Idee der neuromorphen Hardware darin, MVMs innerhalb von resistiven Crossbar-Speicherarrays (Synapse) von Rechenelementen zu implementieren, die aus neuartigen Speicherbausteinen (auch als Gewichtsspeicherbausteine bezeichnet) bestehen, während die Neuronenmimik durch CMOS betrieben wird. Die Rolle der Neuronen besteht darin, die Beiträge der gewichteten Signale zu summieren, die von der vorhergehenden Schicht von Neuronen kommen (Akkumulationsoperation), und dann einen Reiz an die folgende Schicht zu senden, wenn die Summe einen Schwellenwert überschreitet ("Feuer"-Operation). Die übliche Strategie ist in der Regel ein Offline-Training

(Quantisierung), so dass die Gewichte dauerhaft in der Hardware gespeichert werden können, mit begrenzten Aktualisierungen im Feld und einem erneuten Training, was den Verschleiß an Schreibenergie verringert. Obwohl eine Vielzahl von Materialsystemen für den Aufbau synaptischer Geräte erforscht wurde, haben die meisten von ihnen relativ große Betriebsströme (siehe [RAJ19] für eine Übersicht).

Zukünftige Technologien auf Basis der Spintronik

Das Bausteinelement von Magnetoresistive Random Access Memory (MRAM) ist der Magnetic Tunnel Junction (MTJ), der in seiner einfachsten Form aus zwei magnetischen Schichten besteht, die durch eine sehr dünne dielektrische Schicht (etwa 1nm dick) getrennt sind [BRATA12]. Je nach relativer Ausrichtung der Magnetisierungen der beiden Schichten (Parallel oder Antiparallel) ändert sich der elektrische Widerstand, wodurch die binäre Information nichtflüchtig kodiert wird. Das Lesen besteht in der Messung des Widerstands durch den Tunneling Magnetoresistance Effect (TMR), wie in Abb. 2.a gezeigt. Das Schreiben kann auf verschiedenen Effekten beruhen (Abb. 2.b,c): Spin-Transfer-Torque (STT) [KENT15] und Spin-Orbit-Torque (SOT) [MANC19], die jeweils unterschiedliche Leistungswerte in Bezug auf Schreibgeschwindigkeit, Ausdauer und Zellgröße haben. Während STT das heute am weitesten verbreitete Schreibschema ist, ist es durch die gemeinsamen Schreib- und Lesepfade begrenzt, was eine freie Abstimmung des Widerstands der MTJ verhindert. SOT ist ein aufstrebenderes Schreibschema, das allmählich industrielle Reife erlangt und sich besonders für den schnellen Static Random Access Memory (SRAM)-Cache-Ersatz eignet. Dank seiner getrennten Lese- und Schreibpfade ermöglicht es eine starke Erhöhung des elektrischen Widerstands, wodurch die Leseeigenschaften des MTJ für die neuromorphe Nutzung und nicht für die Speichernutzung angepasst werden können, ohne die Schreibfähigkeiten zu beeinträchtigen, was ihn zu einem geeigneteren Kandidaten für neuromorphe Systeme mit niedrigem Energieverbrauch macht.

MRAM wird heute hauptsächlich für Speichermärkte entwickelt [DIENY20] und gilt als die Lösung der Wahl für eingebettete NVM an fortgeschrittenen Technologieknoten, da es CMOS-Spannungs- und Stromkompatibilität, hohe Retentionszeit (10 Jahre), große Ausdauer ($>10^{12}$ Zyklen) und relativ schnelle Schreib-/Lesezeit (~10ns) kombiniert. Aktuell haben "große" Halbleiterunternehmen (Foundries, Integrated Device Manufacturer (IDM) und Zulieferer) wie Samsung, IBM, Intel, GlobalFoundries, Intel und TSMC, Advance Materials umfangreiche Entwicklungsprogramme in Angriff genommen und kündigen alle an, für die Massenproduktion für eingebetteten Flash und langsamen SRAM-Ersatz bereit zu sein. Auch wenn die Herausforderungen in Bezug auf den hohen Schreibstrom (Verringerung der Zellgrundfläche aufgrund des übergroßen Transistors), die Erhöhung der TMR auf über 150 % oder die Kompatibilität mit dem BEOL-Wärmebudget noch bestehen, weist SOT-MRAM bereits alle erforderlichen Eigenschaften auf, um effizientes Inferenz-Edge-Computing mit geringem Stromverbrauch zu ermöglichen.

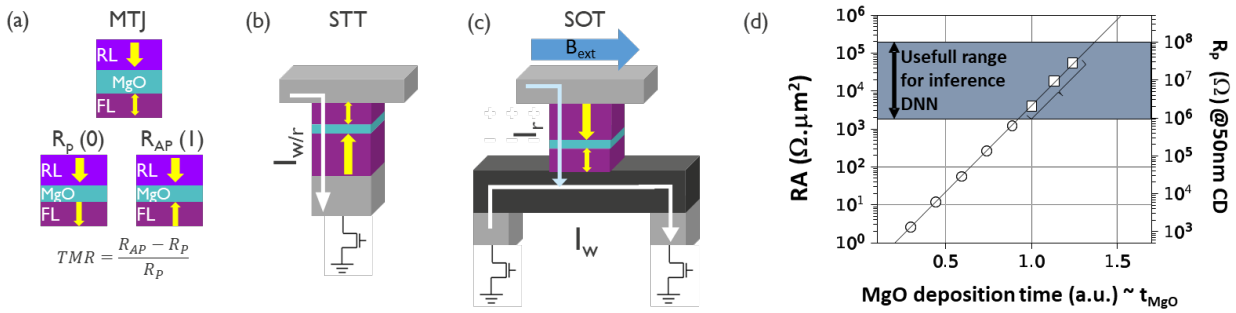

Abb. 2: a) MTJ-Lesekonzept unter Ausnutzung des TMR-Effekts, b) STT, c) SOT mit verschiedenen Schreibmechanismen, die in der MRAM-Technologie verwendet werden können, wobei der Lese- (Ir) und der Schreibpfad (Iw) angegeben sind. d) Widerstandsdurchstimmbarkeit (RP) von MTJ als Funktion der MgO-Tunnelbarrieren-Abscheidungszeit

NeuSPIN Konzept: Spin Orbit Torque (SOT) Technologie zum Aufbau neuromorpher Hardware

Nichtflüchtige Speichertechnologien (NVM) drängten sich als die natürlichen Kandidaten für synaptische Gewichte auf [IELM18], da sie schnelles Instant-On-Off-Computing mit dichter Architektur ermöglichen, während Neuronen noch aus CMOS-Transistoren bestehen, daher die Bezeichnung In-Memory-Computing (IMC).

Der Summationsansatz der analogen Inferenzberechnung im Speicher erfordert niedrige Leseströme und führt zu Speichergewichten mit >MΩ Widerstand [COSE19]. Darüber hinaus sind geringe Schwankungen von Zelle zu Zelle und keine zeitliche Drift des Lesewiderstands und des Lesefensters erforderlich, um einen Verlust der Inferenzlesegenauigkeit zu vermeiden. Um eingebettete Edge-Green-KI-Anwendungen zu ermöglichen und neuromorphe Architekturen mit geringem Stromverbrauch zu erreichen, muss man sich daher auf NVM-Technologien konzentrieren, die geringe Leseströme (Hauptpriorität) und enge Parameterverteilungen kombinieren. Interessanterweise ist ein wesentlicher Nachteil traditioneller NVMs, nämlich die Schreiblatenz und die Schreibenergie, für neuromorphe Anwendungen kein Thema mehr, da sie hauptsächlich auf Lese- und sehr seltenen Aktualisierungen (Schreiboperationen) basieren. Die Hauptanforderung beim Schreiben sollte sein, dass die Zugriffstransistoren nicht den Zellen-Footprint diktieren und mit den Schreibspannungen kompatibel sind.

Tabelle 1: Leistungsmerkmale von NVMs für DNN-Inferenzanwendungen (* Werte abhängig von der endgültigen Speicherfeldgröße)

	OxRAM, CB-RAM	PCM	STT-MRAM	**SOT-MRAM**	Inference specs.
TMR, Reading windows (%)	>10k	>10k	150-200	100-200	150
$\sigma_{weights}$	$\sigma_{weights}$ >> 10% at R_{ON} > 1MΩ	<10%	<5%	**<10%**	<10%*
R_{ON}		<10MΩ	<20k	**Process-tunable up to 100M**	1-100M*

Technology restrictions	Cell to cell variability	Resistance drift, density	R_{on}	TMR, density	

Unter den verschiedenen NVMs, die für die Implementierung solcher engines in Frage kommen, wurden bereits mehrere künstliche Synapsen realisiert, z. B. mit ReRAM, Phase Change Memories und ferroelektrischen Materialien. Die Hauptvorteile von MTJs für KI-Anwendungen sind geringe Leistung, Geschwindigkeit und Ausdauer, wie in Tabelle 1 zusammengefasst. Wir argumentieren dennoch, dass MTJs diese Anforderungen erfüllen können, trotz des geringen On/Off-Verhältnisses (TMR) im Vergleich zu anderen NVMs. Noch wichtiger ist, dass sie auch eine enge Leseverteilung haben und ihr Widerstandsniveau von kΩs bis zu zehn MΩs abstimmbar ist. Darüber hinaus sollen neuromorphe Berechnungsimplementierungen auf System on Chips (SoCs) in Form eines Beschleunigers integriert werden, der mit einem herkömmlichen Prozessor und dessen Speicher verbunden ist. In Anbetracht der Tatsache, dass MRAM die Zukunft der eingebetteten Speicher darstellt, erleichtert die Verwendung derselben Technologie für die Herstellung sowohl der Speicher als auch des neuromorphen Beschleunigers die Integration erheblich und führt zu einer erheblichen Kosteneinsparung bei der Herstellung.

Da NVMs und insbesondere die SOT-Technologie einzigartigen Variationen und Unvollkommenheiten unterliegen, werden bestehende NN-Trainingsalgorithmen, wie sie für traditionelle CMOS-basierte NN-Implementierungen verwendet werden, zu einer schlechten Inferenzgenauigkeit führen, wenn sie auf spintronic-basierter neuromorpher Hardware eingesetzt werden. Aufgrund der geringen Widerstandsvariation von STT/SOT im Vergleich zu anderen NVMs können sich die Verteilungen von Ron und Roff einzelner Bausteine, die synaptische Gewichte repräsentieren, überlappen, was bei Verwendung der traditionellen Trainingsalgorithmen zu einer falschen synaptischen Operation führen kann. Gleichzeitig müssen größere Modifikationen am magnetischen Stack von SOT und an den Bauteilcharakteristika, die primär für Speicheranwendungen verwendet werden, vorgenommen werden, um im Kontext des neuromorphen Rechnens nutzbar zu sein. Daher sollten die Trainingsalgorithmen und Backpropagation-Methoden probabilistisch sein und die Verteilung der Gewichte statt fester Werte berücksichtigen. Daher sind Bayes'sche neuronale Netze, die von Natur aus mit synaptischen Gewichtsverteilungen statt mit nominalen Werten arbeiten, in diesem Zusammenhang sehr vielversprechend. Bayes'sches Lernen ist nicht nur eine Möglichkeit, die Variabilität abzuschwächen, sondern auch die Anfälligkeit von groß angelegtem Lernen zu überwinden und Unsicherheitsintervalle für die von neuronalen Netzen gelieferten Vorhersagen bereitzustellen.

---

2. Überblick über die Arbeitspakete und erzielten Ergebnisse

Zentrale Erkenntnisse des NeuSPIN-Projekts

1. **Effektive Erkennung von Out-of-Distribution-Daten**
   In mehreren Studien konnte gezeigt werden, dass Bayesianische Neuronale Netzwerke (BayNNs) in der Lage sind, bis zu 100 % der Out-of-Distribution (OOD)-Daten zuverlässig zu erkennen. Dies unterstreicht ihre Robustheit im Umgang mit unerwarteten oder verrauschten Eingabedaten.
2. **Verbesserte Inferenzgenauigkeit bei beschädigten Daten**
   Die Implementierung bayesianischer Methoden führte zu einer signifikanten Verbesserung der Inferenzgenauigkeit, insbesondere in Szenarien mit beschädigten oder unvollständigen Daten. Dies ist besonders relevant für sicherheitskritische Anwendungen, in denen die Datenintegrität nicht immer gewährleistet ist.
3. **Integration von spintronikbasierten Technologien**
   Der Einsatz von Spintronik-Speichern, insbesondere in Computing-in-Memory (CIM)-Architekturen, bietet Vorteile wie niedrigen Energieverbrauch und hohe Schaltgeschwindigkeit. Diese Eigenschaften sind essenziell für eine effiziente Hardware-Implementierung von BayNNs. Zudem kann die stochastische Natur spintronischer Elemente zur Realisierung von Zufallszahlengeneratoren (RNGs) genutzt werden – mit derselben Technologie.
4. **Modellierung von Defekten in Bauelementen**
   Das Verständnis und die Berücksichtigung von Defekten und Fertigungsvarianzen in spintronischen Bauelementen ist entscheidend. Die Integration solcher Unregelmäßigkeiten in die Modellierung erlaubt realitätsnahe Hardware-Charakterisierungen und führt zu verbesserten Strategien im Algorithmus-Hardware-Co-Design.
5. **Algorithmus-Hardware-Co-Design**
   Die gezielte gemeinsame Entwicklung von Algorithmen und Hardware – insbesondere die Anpassung bayesianischer Verfahren an moderne Hardwareplattformen wie Spintronik und CIM – ermöglicht deutlich effizientere und praxistauglichere BayNN-Implementierungen.
6. **Quantisierte BayNNs**
   Die Quantisierung oder Binarisierung von Parametern und Aktivierungen stellt eine zentrale Strategie dar, um den Speicherbedarf zu senken und die begrenzten Zustände von Spintronik-Speichern optimal zu nutzen. Dadurch lassen sich die typischen Ressourcenanforderungen bayesianischer Modelle reduzieren, was deren Einsatz in ressourcenbeschränkten, sicherheitskritischen Anwendungen ermöglicht.
7. **Ressourcenbewusste Methoden zur bayesianischen Approximation**
   Neue Dropout-Techniken wie **Scale Dropout** und **Affine Dropout** sowie speicherzentrierte Näherungsverfahren ermöglichen eine Reduktion der benötigten Zufallszahlengeneratoren und eine effizientere BayNN-Implementierung.
8. **Inhärente Robustheit und Selbstheilungseigenschaften**
   BayNNs weisen im Vergleich zu klassischen neuronalen Netzen eine höhere Fehlertoleranz gegenüber Variationen und Störungen auf. Diese Selbstheilungseigenschaften lassen sich durch speziell entwickelte Methoden – etwa die **invertierte Normalisierung mit Affine Dropout** – weiter verstärken.

Das Projekt gliederte sich in vier aufeinander aufbauende Arbeitspakete (APs). Die Verknüpfung der Ergebnisse mit den jeweiligen Publikationen und Deliverables ist im jeweiligen Abschnitt aufgeführt.

AP1 – Fabrikations- und Device-Level-Optimierungen von SSZ-Zellen

Ziele:
Optimierung von Spintronic-Zellen (SOT-MTJs) hinsichtlich Variabilität, Energiebedarf, Stabilität und stochastischer Eigenschaften. Entwicklung statistischer Modelle für das Verhalten der Zellen als Grundlage für spätere Simulationen und das Hardwaredesign.

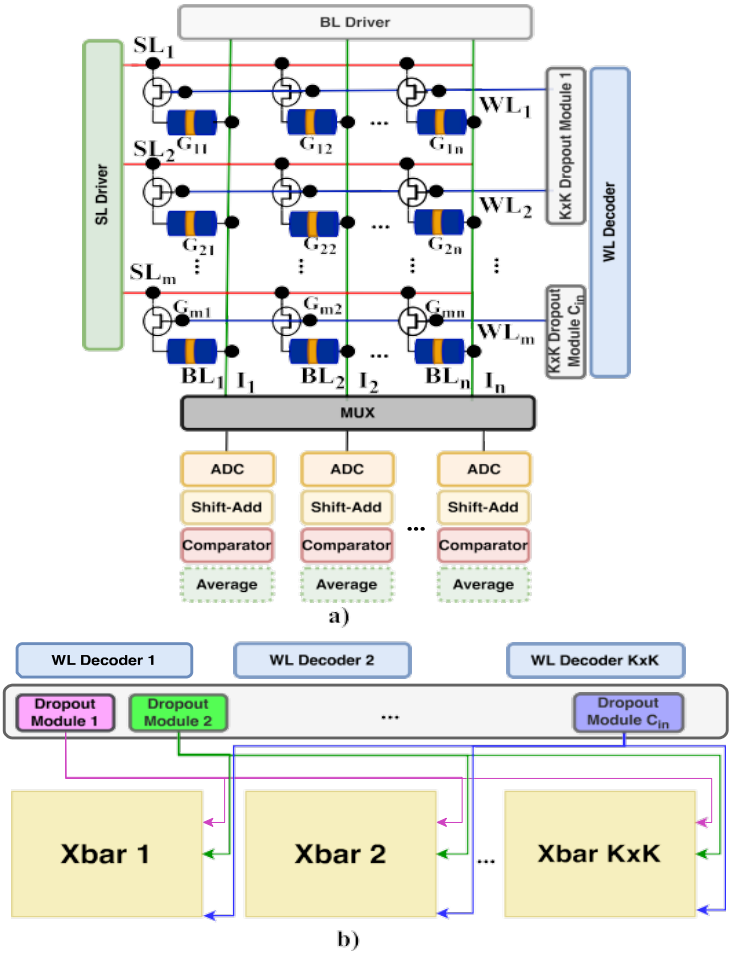

Abb. 3: Crossbar-Design für MC-SpatialDropout basierend auf der Mapping-Strategie

Arbeiten und Ergebnisse:

- Aufbau eines variabilitätsbewussten Technologie-Modells basierend auf der statistischen Charakterisierung realer SOT-MTJ-Bauelemente (z. B. Verteilungen von Schreibströmen, Anisotropie, TMR).
- Entwicklung eines PDKs mit kompaktem stochastischem Modell für SSZ-Zellen.
- Untersuchung der inhärenten Stochastizität von MTJs und deren Nutzen für bayessche Lernalgorithmen.

Veröffentlichungen:

- *SpinDrop*, *SpinBayes* und *Spatial-SpinDrop* (z. B. [Ahmed et al., IEEE TNANO 2023], [Ahmed et al., TECS 2023]) nutzen direkt die stochastische Modellierung und Validierung aus WP1.

Deliverables:

- D1.2: Modellbibliothek mit stochastischen Charakteristika der SOT-MTJ-Zellen.

AP2 – Schaltungsentwurf des neuromorphen Kerns

Ziele:
Design und Optimierung eines neuromorphen Hardware-Kerns auf SSZ-Basis. Fokus auf Energieeffizienz, Fehlertoleranz und Inferenzgenauigkeit. Co-Design der Peripherie zur effizienten synaptischen Gewichtung und Datenverarbeitung.

Arbeiten und Ergebnisse:

- Entwurf einer Matrixarchitektur mit parallelem Zugriff auf Synapsenströme.
- Entwicklung von Stromspar-Peripherie (z. B. getaktete Multiplex-Schaltungen).
- Untersuchung redundanter MTJ-Architekturen zur Reduktion von Variabilitätseinflüssen.
- Integration von Schaltungsrauschen in Designentscheidungen durch Monte-Carlo-Analysen.

Veröffentlichungen:

- *NeuSPIN* [Ahmed et al., DATE 2024] präsentiert eine vollständige Systemarchitektur.
- *Inverted Normalization* [Ahmed et al., DATE 2024] zeigt eine Methode zur Fehlerkompensation auf Schaltungsebene.

Deliverables:

- D2.1: Entwurfsrichtlinien für SSZ-basierte NN-Schaltungen.
- D2.2: Modell für Co-Design mit Lernalgorithmen (verwendet in AP3).

AP3 – Technologie-bewusster Entwurf von Trainingsalgorithmen

Ziele:
Entwicklung robuster Lernalgorithmen, die hardwarebedingte Stochastizität und Variabilität integrieren. Fokus auf bayessche Netzwerke mit Unsicherheitsabschätzung und auf rechenarme Trainingsmethoden.

Arbeiten und Ergebnisse:

- Entwicklung und Simulation von *Variational Bayes Learning Rules* (VBLR) unter stochastischen Bedingungen.
- Implementierung von *SpinDrop* als dropout-basierte Approximation von Gewichtsunsicherheiten.
- Validierung von *Bayes by Backprop* und *Monte Carlo Dropout* auf Spintronic-Modellen.

Veröffentlichungen:

- *SpinDrop* [Ahmed et al., IEEE JETCAS 2023], Best Paper.
- *SpinBayes*, *SpinDrop*, *Spatial-SpinDrop* – umfassende algorithmische Beiträge.

Deliverables:

- D3.1: Code und Dokumentation zu VBLR mit Unsicherheitsmodellierung.
- D3.2: Vergleich von STLR und VBLR auf realer Datenbasis.

AP4 – Schichtübergreifende Simulation und Anwendungsbenchmarking

Ziele:
Entwicklung eines Cross-Layer-Simulationsframeworks zur Bewertung von Hardware-algorithmischer Co-Optimierung. Anwendung auf realitätsnahe Szenarien (u. a. medizinische Signalverarbeitung).

Arbeiten und Ergebnisse:

- Aufbau eines Co-Simulationsframeworks von Gerät bis Anwendung.
- Simulation realitätsnaher Tasks mit injected noise/variability.
- Benchmark auf Aufgaben aus der Bioelektronik und Bildverarbeitung.

Veröffentlichungen:

- *Testing Spintronics-Implemented Monte Carlo Dropout* [Ahmed et al., ETS 2024].
- Präsentation von Cross-Layer-Effekten auf DATE, ETS und CODES+ISSS.

Deliverables:

- D4.1: Framework-Dokumentation.
- D4.2: Benchmarking-Analysebericht.

---

3. Verwendung der Zuwendung

Die Fördermittel wurden überwiegend eingesetzt für:

- **Personal**: Promotionsstellen, wissenschaftliche Mitarbeiter.
- **Reisekosten** für Konferenzen.
- **Sachmittel** für Simulation, Tool-Lizenzen, Datenspeicher.
- **Kooperationsmittel** für bilaterale Workshops mit SPINTEC.

---

4. Angemessenheit der Projektarbeiten

Alle vier Arbeitspakete wurden gemäß Projektplan umgesetzt, Verzögerungen wurden durch parallele Aktivitäten kompensiert. Die Verbindung von Materialwissenschaft, Schaltungsdesign und ML-Algorithmen war für die Zielerreichung zwingend notwendig. Die Aufgabenverteilung zwischen KIT und Grenoble erlaubte eine optimale Nutzung der jeweiligen Stärken.

---

5. Voraussichtlicher Nutzen und Verwertung

Die Resultate von NeuSPIN sind in mehrfacher Hinsicht verwertbar und haben bereits zur Initiierung weiterführender Forschungs- und Kooperationsaktivitäten beigetragen:
• **Wissenschaftlich:**
Das Projekt legte die methodischen und technologischen Grundlagen für die Entwicklung energieeffizienter und robuster KI-Architekturen („Green AI“) auf Basis neuartiger Speichertechnologien. Es führte zur Etablierung einer strategischen Partnerschaft zwischen dem Karlsruher Institut für Technologie (KIT) und der Université Grenoble Alpes (UGA), zwei führenden europäischen Forschungseinrichtungen im Bereich Mikroelektronik und Künstlicher Intelligenz. Auf Basis der komplementären Kompetenzen der beiden Partner wurden gemeinsam Anschlussprojekte im Rahmen der bilateralen Förderprogramme von **DFG und ANR** (Deutsche Forschungsgemeinschaft und Agence Nationale de la Recherche) geplant.
• **Technologisch:**
Das entwickelte Co-Design-Framework sowie das technologiebezogene PDK (Process Design Kit) werden in aktuellen Forschungsprojekten des KIT weiterverwendet, u. a. in Kooperation mit industriellen Partnern wie **Bosch**. Die Integration von BayNNs auf Spintronik-basierter Computing-in-Memory-Hardware ist dabei zentraler Bestandteil weiterer Technologieentwicklungen.
• **Industriell:**
Es bestehen erste Kontakte mit europäischen Halbleiterunternehmen zur Überführung der im Projekt entstandenen Konzepte in industrielle Produktlinien im Bereich Edge-KI. Hierzu zählen Anwendungen in den Bereichen Mobilität, Medizin und industrielle Zustandsüberwachung. Eine Weiterentwicklung der Konzepte erfolgt auch im Kontext europäischer Förderinitiativen, insbesondere im Rahmen der **Chips Joint Undertaking (Chips JU)**, bei denen beide Partner – KIT und UGA – aktiv eingebunden sind.
• **Infrastrukturell:**
Ein bedeutender struktureller Impuls des Projekts war die Mitinitiierung des **Karlsruhe Chipdesign House (KCH)**, das im Rahmen einer Förderung des Ministeriums für Wissenschaft, Forschung und Kunst Baden-Württemberg (MWK) gegründet wurde. KCH

fungiert als zentrale Anlaufstelle für Ausbildung, Forschung und Vernetzung im Bereich Chipdesign – auf **regionaler, nationaler und europäischer Ebene** – und stellt eine wesentliche Brücke zwischen akademischer Forschung und industrieller Anwendung dar. Die Inhalte, Zielstellungen und Expertise aus dem NeuSPIN-Projekt waren maßgeblich an der Konzeption, inhaltlichen Fokussierung und Ausrichtung von KCH beteiligt.

Geplante Verwertung

- **Einreichung eines EU-Horizon-Forschungsprojekts** mit Fokus auf den Technologietransfer von NeuSPIN-Ansätzen in sicherheitskritische Automobilanwendungen.
- **Veröffentlichung des entwickelten Simulationsframeworks** (inkl. BayNN-Modelle, Spintronik-CIM-Integration und Störmodellen) auf **GitHub** zur freien Nutzung in der Forschungscommunity.
- **Gemeinsame Anschlussprojekte von KIT und UGA** im Rahmen bilateraler Programme (DFG/ANR) und Chips JU zur weiteren Vertiefung der Kooperation im Bereich Green Edge AI.
- **Einbindung der Projektergebnisse in Curricula und Ausbildungsangebote** am KCH zur Förderung des Chipdesign-Nachwuchses in Deutschland und Europa.

---

6. Beobachteter Fortschritt bei Dritten

Während der Projektlaufzeit wurden weltweit vergleichbare Aktivitäten beobachtet, darunter:

- IBM Research Zürich – probabilistische Rechenknoten mit PCM.
- Tsinghua University – resistive In-Memory-Netzwerke für probabilistische ML.
- CEA Leti – SOT-basierte CIM-Prozessorprototypen.

Der projektinterne Vergleich zeigte, dass *NeuSPIN* durch den Fokus auf unsicherheitsbewusste KI und die reale Modellintegration stochastischer Effekte weltweit führend ist.

---

7. Veröffentlichungen

Die nachfolgend aufgeführten wissenschaftlichen Veröffentlichungen sind im Rahmen des NeuSPIN-Projekts bereits erschienen und dokumentieren die wesentlichen Forschungsergebnisse des Vorhabens. Der aktuelle Stand der Publikationstätigkeit spiegelt den erfolgreichen Abschluss der wissenschaftlich-technischen Arbeiten wider. Weitere Veröffentlichungen zu diesem Projekt sind **derzeit nicht geplant**.

Im Rahmen des Projekts *NeuSPIN* entstanden eine Vielzahl hochrangiger wissenschaftlicher Veröffentlichungen in renommierten Fachzeitschriften und auf internationalen Konferenzen. Einige Arbeiten wurden als *Best Paper* oder *Best Paper Finalist* ausgezeichnet. Die Publikationen decken das gesamte Spektrum von Bauelementmodellierung über Schaltungsdesign bis zu algorithmisch-architektonischem Co-Design ab.

Konferenzveröffentlichungen:

- Tuhin, S., Danouchi, K., Münch, C., Prenat, G., Anghel, L., & Tahoori, M. B.
  Binary Bayesian Neural Networks for Efficient Uncertainty Estimation Leveraging Inherent Stochasticity of Spintronic Devices.
  17th ACM International Symposium on Nanoscale Architectures (NANOARCH), Dezember 2022
  (Best Paper Finalist)
- Ahmed, S. T., Danouchi, K., Hefenbrock, M., Prenat, G., Anghel, L., & Tahoori, M. B.
  Enhancing Reliability of Neural Networks at the Edge: Inverted Normalization with Stochastic Affine Transformations.
  Design, Automation & Test in Europe (DATE), 2024, pp. 1–6.
  (Best Paper Finalist)
- Ahmed, S. T., Danouchi, K., Hefenbrock, M., Prenat, G., Anghel, L., & Tahoori, M. B.
  NeuSpin: Design of a Reliable Edge Neuromorphic System Based on Spintronics for Green AI.
  DATE 2024, pp. 1–6.
- Ahmed, S. T., Danouchi, K., Hefenbrock, M., Prenat, G., Anghel, L., & Tahoori, M. B.
  Testing Spintronics Implemented Monte Carlo Dropout-Based Bayesian Neural Networks.
  IEEE European Test Symposium (ETS), 2024, pp. 1–6.
- Ahmed, S. T., Danouchi, K., Hefenbrock, M., Prenat, G., Anghel, L., & Tahoori, M. B.
  Scalable Spintronics-Based Bayesian Neural Network for Uncertainty Estimation.
  DATE 2023, pp. 1–6.
- Ahmed, S. T., Hefenbrock, M., & Tahoori, M. B.
  Tiny Deep Ensemble: Uncertainty Estimation in Edge AI Accelerators via Ensembling Normalization Layers with Shared Weights.
  IEEE/ACM International Conference on Computer-Aided Design (ICCAD), 2024, pp. 1–9.

Journalveröffentlichungen:

- Ahmed, S. T., Danouchi, K., Münch, C., Prenat, G., Anghel, L., & Tahoori, M. B.
  SpinDrop: Dropout-Based Bayesian Binary Neural Networks With Spintronic Implementation.
  IEEE Journal on Emerging and Selected Topics in Circuits and Systems, Vol. 13, No. 1, 2023, pp. 150–164.
  (Best Paper Award)
- Ahmed, S. T., Danouchi, K., Hefenbrock, M., Prenat, G., Anghel, L., & Tahoori, M. B.
  SpinBayes: Algorithm-Hardware Co-Design for Uncertainty Estimation Using Bayesian In-Memory Approximation on Spintronic-Based Architectures.
  ACM Transactions on Embedded Computing Systems, Vol. 22, No. 5s, 2023, pp. 1–25.

- Ahmed, S. T., Danouchi, K., Hefenbrock, M., Prenat, G., Anghel, L., & Tahoori, M. B.
  Spatial-SpinDrop: Spatial Dropout-Based Binary Bayesian Neural Network with Spintronics Implementation.
  IEEE Transactions on Nanotechnology (TNANO), 2023.

Zusätzliche Hinweise zur Verwertung:

Viele dieser Publikationen sind Open Access verfügbar und dienen bereits als Referenz für weitere Arbeiten in verwandten EU-Projekten und bei Industriepartnern. Besonders hervorzuheben ist die Veröffentlichung *SpinDrop* (IEEE JETCAS), die im Review-Prozess mit dem *Best Paper Award* ausgezeichnet wurde und einen zentralen Beitrag zur Co-Design-Strategie von *NeuSPIN* leistet.

---

8. Zusammenfassung

Bayesianische Neuronale Netzwerke (BayNNs) besitzen die inhärente Fähigkeit, Unsicherheiten in ihren Vorhersagen abzuschätzen – ein entscheidender Vorteil gegenüber konventionellen neuronalen Netzen. Gleichzeitig sind BayNNs jedoch sehr ressourcenintensiv: Sie erfordern hohen Energieverbrauch, große Speicherkapazitäten, geringe Latenzen und eine Vielzahl an Zufallszahlengeneratoren (RNGs) pro Modellinstanz. Ein vielversprechender Ansatz zur Reduktion dieser Kosten besteht in der Implementierung innerhalb einer **spintronikbasierten Computing-in-Memory (CIM)**-Architektur. Diese Umsetzung ist jedoch technisch anspruchsvoll und mit mehreren Herausforderungen verbunden.

Das **NeuSPIN-Projekt**, ein gemeinsames deutsch-französisches Vorhaben, verfolgte das Ziel einer durchgängigen Optimierung entlang des gesamten Technologie-Stacks – von der Algorithmusebene über die BayNN-Topologie bis hin zum Schaltungsdesign. Dabei stand insbesondere die Implementierung und Optimierung bayesianischer Trainings- und Inferenzprozesse in CIM-Architekturen mit **nichtflüchtiger Spintronik-Technologie** im Fokus. Durch diesen integrierten Ansatz sollte sowohl der Hardware-Footprint signifikant reduziert als auch energieeffiziente Unsicherheitsabschätzungen für sicherheitskritische Edge-AI-Anwendungen ermöglicht werden.

Zu den wichtigsten Projekterfolgen zählen:

- Die Fähigkeit zur **Erkennung von bis zu 100 % Out-of-Distribution-Daten** (OOD),
- Eine **kontinuierliche Steigerung der Inferenzgenauigkeit**, insbesondere bei fehlerhaften oder verrauschten Eingaben,
- Eine **erhöhte Robustheit gegenüber Nichtidealitäten** und Prozessvariationen in spintronischen Bauelementen.

Ein zentraler Durchbruch wurde durch die enge Verzahnung von **algorithmischen Innovationen** mit **fortschrittlichen Hardwaretechnologien** wie Spintronik und CIM erzielt. Dies führte zu einer drastischen **Reduktion des Energieverbrauchs (bis zu 100×)** und des **Speicherbedarfs (bis zu 158,7×)**. Somit wurde ein wesentlicher Grundstein für die **skalierbare, zuverlässige und ressourceneffiziente Implementierung** von BayNNs in realen und sicherheitskritischen Anwendungen gelegt.

Die **Zusammenarbeit zwischen den Projektpartnern**, dem Karlsruher Institut für Technologie (KIT) und der Université Grenoble Alpes (UGA), erwies sich als überaus erfolgreich – nicht nur aufgrund der sich ergänzenden Kompetenzen beider Teams, sondern auch durch den Zugang zu realitätsnahen Gerätedaten und Prozessparametern aus der **SPINTEC-Fertigung und -Charakterisierung**.

Abschließend lässt sich festhalten, dass das Projekt **NeuSPIN** ein wertvolles **Know-how an der Schnittstelle von Chipdesign, neuartigen Speichertechnologien und Edge-KI** aufgebaut hat. Diese deutsch-französische Zusammenarbeit – zwischen den beiden führenden europäischen Nationen im Bereich **Künstliche Intelligenz und Mikroelektronik** – hat bereits zu weiterführenden Aktivitäten und Kooperationen im Kontext des **EU Chip Acts**, des **AI Acts** sowie der **„Chipdesign Germany“** geführt.