Schlussbericht Universität zu Lübeck

Zuwendungsempfänger:	Universität zu Lübeck
Verbundprojekt:	Elektroniksysteme für künstliche Intelligenz im Edge-Computing - DAIS
Förderkennzeichen:	16MEE0126
Bewilligungszeitraum:	01.06.2021 - 31.12.2024

Kurzdarstellung

Die technologische Entwicklung im Bereich der Consumer Electronics und industriellen Anwendungen hat zur Verbreitung kleiner, vernetzter Geräte geführt, die lokal Daten erfassen und verarbeiten können. Dieses sogenannte Internet of Things (IoT) eröffnet neue Möglichkeiten in zahlreichen Bereichen wie Automobilindustrie, Energiewirtschaft, Gebäudetechnik und Industrie 4.0.

Mit der zunehmenden Verbreitung dieser Systeme steigt jedoch auch das Datenvolumen, was hohe Anforderungen an Netzwerkkapazität und Latenz mit sich bringt. Das Paradigma des *Edge Computings* begegnet dieser Herausforderung, indem es Rechenressourcen näher an die Datenquelle bringt. Dabei bleiben sensible Daten lokal, was sowohl die Netzwerklast reduziert als auch Datenschutz- und Sicherheitsanforderungen erfüllt. Auch KI-Anwendungen können so dezentralisiert ausgeführt werden. Dies führt zu geringerer Latenz, höherer Energieeffizienz und verbesserter Skalierbarkeit, stellt jedoch auch neue Anforderungen an die Leistungsfähigkeit der Edge-Geräte.

Das Projekt *Elektroniksysteme für künstliche Intelligenz im Edge-Computing – DAIS* widmete sich diesen Herausforderungen, insbesondere im industriellen Kontext des *Industrial IoT* (IIoT), und berücksichtigte dabei die zunehmende Integration von KI-gestützten Verfahren. Ziel war die Entwicklung intelligenter, sicherer und energieeffizienter Edge-Knoten, die KI-Algorithmen lokal oder verteilt verarbeiten können. Ein zentraler Aspekt war die bedarfsgerechte Ausstattung der Knoten mit Hardware (z. B. FPGAs, ASICs, SoCs) und die flexible Verteilung von KI-Verarbeitung zwischen Edge und Cloud. Zusätzlich musste parallel die Optimierung von Energieverbrauch, Rechenleistung und Kosten erfolgen.

Konkret verfolgte DAIS folgende Ziele:

- Entwicklung selbstorganisierender, energieeffizienter Hardware für Edge-KI-Anwendungen mit Fokus auf Sicherheit und Datenschutz.
- Entwicklung ergänzender Software, insbesondere um die Nutzung der entwickelten Hardware zu ermöglichen.
- Sichere Integration der Edge-Komponenten in Cloud- und Fog-Architekturen, einschließlich Koordination und Orchestrierung verteilter KI-Aufgaben und Gewährleistung von Sicherheit und Datenschutz in verteilten Systemen.
- Demonstration industrieller KI-Anwendungen auf Basis der entwickelten DAIS-Komponenten.

Insgesamt zielte DAIS auf eine durchgängige, sichere und intelligente IIoT-Infrastruktur ab, die sich flexibel in verschiedene industrielle Szenarien integrieren lässt.

Aufgabenstellung

Die Universität zu Lübeck (UzL) leistete im Rahmen des Projekts einen zentralen Beitrag im Bereich des Hardware-Designs, mit Fokus auf zwei Schlüsselentwicklungen:

- Entwicklung eines sicheren, RISC-V-basierten Rechenknotens
  Dieser Prozessor wurde speziell für energieeffizientes, sicheres Edge Computing entwickelt und mit Hardware-basierten Sicherheitskomponenten ausgestattet.
- *Entwicklung neuartiger Beschleuniger für sensornahe Datenverarbeitung*
  Ziel war die direkte Verarbeitung von Sensordaten nahe der oder in der Sensoreinheit selbst. Dadurch wurden Latenzzeiten reduziert und der Energieverbrauch gesenkt.

Gesamtheitliches Ziel war die Bereitstellung einer Hardwareplattform mit niedrigem Energiebedarf, geringer Latenz und hoher Sicherheit für Edge-Anwendungen.

Technische Ziele im Detail:

- Der RISC-V-Prozessor erhält eine integrierte, leichtgewichtige Sicherheitseinheit als Trust Anchor, die als *Secure Device Identity* (SDI) fungiert. Dieses SDI schützt vor Angriffen wie Cloning und Schadsoftware und garantiert die eindeutige Identifikation des Geräts sowie eine vertrauenswürdige Ausführungsumgebung.
- Die Hardware basiert auf der offenen RISC-V-Architektur, erweitert um spezialisierte KI-Beschleuniger. Diese Beschleuniger kombinieren digitale und analoge, RRAM-basierte Komponenten, um maximale Energieeffizienz bei der Ausführung von KI-Algorithmen zu erreichen. Dieser Beschleuniger agiert dabei als Koprozessor, um auch komplexe Signalverarbeitungsaufgaben effizient zu bewältigen. Durch die Möglichkeit zur Rekonfiguration soll eine höhere Flexibilität und Energieeffizienz als mit herkömmlichen Lösungen erzielt werden.

UzL war maßgeblich an den Arbeitspaketen AP2, AP3 und AP4 beteiligt und arbeitete eng mit den Projektpartnern zusammen. Grundlage war eine gemeinsame Anforderungsanalyse (AP1), auf deren Basis die Integration und Validierung der entwickelten Komponenten erfolgte. Durch die Entwicklung und Umsetzung spezialisierter Hardwarelösungen trug UzL substanziell zur gesamten Projektwertschöpfung bei.

Voraussetzungen, unter denen das Vorhaben durchgeführt wurde

Das Projekt erfüllte zentrale förderpolitische Zielsetzungen sowohl im Bereich Hardware als auch Software. Damit unterstützte das Projekt maßgeblich die Ziele deutscher Förderprogramme, insbesondere des Rahmenprogramm *Innovation 2021–2024 (Mikroelektronik. Vertrauenswürdig und nachhaltig. Für Deutschland und Europa*.).

Adressiert wurden insbesondere die dort hervorgehobenen Zukunftsthemen wie Künstliche Intelligenz, Hochleistungsrechnen, Kommunikationstechnologien und Industrie 4.0. Auch die in der Richtlinie Mikroelektronik benannten Schwerpunkte RISC-V und Edge-Processing waren zentrale Bestandteile des Projekts.

Auch die spezifischen deutschen Fördervorgaben werden erfüllt: Die Projektmittelverteilung innerhalb des deutschen Projektanteils war konform, ebenso wie der erforderliche Mindestanteil

deutscher Beteiligung (mind. 10 %). Alle beteiligten deutschen Partner leisteten unverzichtbare Beiträge für den Projekterfolg.

Planung und Ablauf des Vorhabens

UzL war fokussiert auf die Entwicklung und Umsetzung von Schlüsseltechnologien für sichere Rechenknoten und effiziente KI-Beschleunigung für das Umfeld des Internet of Things. Diese Schlüsseltechnologien waren hierbei im Speziellen:

- *Leichtgewichtiger Vertrauensanker*
  Neben den Teilnehmern selbst kann auch direkt deren Kommunikation Ziel böswilliger Angriffe sein. Gerade in verteilten Systemen ist das Kommunikationsaufkommen hoch und die ausgetauschten Informationen sind entscheidend für das Verhalten jedes einzelnen Teilnehmers. Einerseits muss sichergestellt werden, dass keine fremden Informationen in diese Kommunikation eingeschleust werden, andererseits dürfen die potenziell sensiblen Daten nicht unfreiwillig nach außen kommuniziert werden. Eine Kombination aus Ver- bzw. Entschlüsselung und Signierung ist ein effektives Vorgehen gegen beide Problemstellungen. Ziel muss es sein diese Sicherheitsfunktionen leichtgewichtig umzusetzen, um Overhead und damit Effizienzeinbußen der Kommunikation zu vermeiden.

- *RRAM-basierter KI-Beschleuniger*
  RRAM als neuartige Speichertechnologie ermöglicht die Umsetzung vollkommen neuartiger Rechenarchitekturen. Insbesondere die analogen Eigenschaften ermöglichen die Umsetzung von ungefähren oder kontinuierlichen Berechnungen, sowie Berechnungen direkt im Speicher. Gepaart mit dem generell niedrigeren Verbrauch von Energie und Platz von RRAM, bietet dieser sich an zur Umsetzung von Anwendungen in ressourcenbeschränkten Umfeldern. Insbesondere KI-Anwendungen weisen heute eine hohe Komplexität auf, die in verteilten Systemen nur schwer durch konventionelle Architekturen umzusetzen ist. Ein dedizierter KI-Beschleuniger basierend auf RRAM kann ebendiese Problemstellung effektiv adressieren.

Diese Arbeiten wurden in mehreren dedizierten Arbeitspaketen adressiert (Beschreibungen laut Teilvorhabensbeschreibung):

AP1: Anforderungsanalyse

In diesem WP werden die Anforderungen für die sensornahe Datenverarbeitung ermittelt und eine korrespondierende Architektur abgeleitet. Orthogonal hierzu erfolgt die Erarbeitung der Sicherheitsanforderungen zwecks Ableitung der Architektur der Prozessorsicherungseinheit.

AP2: Systementwurf

WP2 beinhaltet den Aufbau der Hardwarekomponenten für die digitale und analoge Datenverarbeitung in intelligenten Edge-Systemen. Hierzu gehört auch die Entwicklung sicherer und verlässlicher Kommunikationseinheiten. Besonderes Augenmerk liegt hierbei auf der Minimierung der Leistungsaufnahme bei gleichzeitiger Erfüllung der hohen Anforderungen an die Rechenleistung.

AP3: Sichere Softwareumgebung

Dieser AP adressiert die Software-Seite der integrierten HW/SW-Komponenten, d.h. Firmware, Middleware und Anwendungssoftware für die betrachteten Anwendungsszenarios (SC); hierbei liegt ein besonderes Augenmerk auf der Robustheit der Lösungen, welche auch

einen hinreichend korrekten Betrieb in Fehlersituationen, sowohl bezogen auf die einzelnen Komponenten wie das Gesamtsystem, sicherstellen

AP4: Systemintegration

Im Fokus dieses APs steht der Aufbau geeigneter industrieller Demonstratoren basierend auf dem DAIS-Framework. Adressiert werden Anwendungsbeispiele aus den Anwendungsbereichen SC6, SC7 und SC8. Mithilfe der komplexen Anwendungsbeispiele erfolgt eine Überprüfung der korrekten Funktionsweise der mithilfe des DAIS-Frameworks abgebildeten Prozesse sowie insbesondere der KI-gestützten Optimierung. Die Ergebnisse dieses APs fließen zusammen mit den Ergebnissen von AP2 und AP3 in den anschließenden AP5.

AP5: Integration und Validierung der Demonstratoren

Im Fokus dieses AP steht der Aufbau integrierter Demonstratoren auf Basis der Ergebnisse von AP1 bis AP4. Der Ausgang dieses AP treibt die Aktivitäten in AP6 (Verwertung und Standardisierung).

AP6: Bekanntmachung, Verwertung und Standardisierung

Im Fokus dieses AP steht die projektbegleitende Präsentation des Projektes über allfällige digitale and Printmedien, die Planung der Projektergebnisverwertung und Standardisierungsbestrebungen sowie korrollar aus der Projektpräsentation die Quervernetzung mit anderen industriellen und Forschungsaktivitäten.

AP7: Projektmanagement

Ziele: Das allgemeine Projektmanagement wird über diesen AP abgewickelt. Dies geschieht vorwiegend durch den ausgewiesenen Projektkoordinator, welcher durch die im Konsortialvertrag zu regelnde technische Administration bzw. dem sogenannten Quality Board, typischerweise bestehend aus den AP-Leitern bzw. den nationalen Koordinatoren, unterstützt wird

Zusammenarbeit mit anderen Stellen

Es fand innerhalb des Projektes eine intensive Zusammenarbeit der Partner statt. Insbesondere innerhalb der Demonstratoranwendungsfälle konnte eine wertvolle gemeinsame Wissensbasis geschaffen werden.

Soweit urheberrechtliche Belange nicht berührend, fand außerdem ein Austausch mit dem BMBF-geförderten Projekt VE-Jupiter sowie dem EU-geförderten Projekt A-IQ Ready statt, welche ebenfalls Aspekte der Komponentensicherheit bzw. KI-Beschleunigung betrachten.

Verwendung der Zuwendung und erzielte Ergebnisse

Der Schwerpunkt der Arbeit von UzL im Projekt DAIS lag auf der Entwicklung eines sicheren RISC-V basierten Rechenknotens mit dediziertem KI-Beschleuniger. Die Entwicklung der Komponenten erfolgte parallel und die Betrachtung daher separat. Folgende Aktivitäten wurden durchgeführt:

- Der sicherer RISC-V Rechenknoten wurde unter Berücksichtigung der spezifischen Ansprüche entworfen und die entsprechenden Hardwarekomponenten umgesetzt. Anschließend wurden diese in der Evaluationsplattform integriert und als Gesamtsystem validiert. Dies erfolgte als FPGA-basierte Implementierung.

- Der RRAM-basierte rekonfigurierbare Beschleuniger wurde entsprechend des festgelegten Anwendungsfalls entworfen. Anschließend wurden einzelne Komponenten feingranulär simuliert, das Gesamtsystem auf einer höheren Abstraktionsebene. Die Sensordaten wurden durch den Beschleuniger als Model-in-the-Loop verarbeitet und die Funktion zu verifizieren.

AP1: Anforderungsanalyse

In Arbeitspaket 1 und insbesondere Teilarbeitspaket AP1.1 *Anwendungsanalyse der Beispielanwendung "Smart Industry"* galt es die Anforderungen zu analysieren die sich aus den konkreten Anwendungsfällen ergaben. Hieraus wurden geeignete Architekturen abgeleitet die sowohl der Verarbeitung der entstehenden Sensordaten als auch den Ansprüchen an die Sicherheit gerecht werden konnten.

UzL hat dabei die folgenden Anforderungen ermittelt:

Anforderung	Beschreibung	Begründung
ML-Framework für RRAM-Beschleuniger	Ein Framework muss festgelegt werden, das die Inferenz vortrainierter Modelle auf dem RRAM-Beschleuniger ermöglicht.	Die Wahl des Frameworks beeinflusst das Design der Beschleunigerarchitektur.
ML-Algorithmus für RRAM-Beschleuniger	Ein ML-Algorithmus muss entsprechend den Fähigkeiten und Grenzen des RRAM-Beschleunigers sowie den Anforderungen des Anwendungsfalls ausgewählt werden.	Die Wahl des Algorithmus hat Einfluss auf das Design der Beschleunigerarchitektur.
Memristor-Modell für RRAM-Beschleuniger	Ein Modell für das Verhalten der Zellen des RRAM-Beschleunigers muss definiert werden, um geeignete Leistungskennzahlen durch Simulation und Emulation zu erhalten.	Da kein Tape-out des RRAM erwartet wird, muss die Modellierung ausreichend genau sein.
Schnittstelle für den RRAM-Beschleuniger	Eine Schnittstelle für den RRAM-Beschleuniger muss definiert werden, um die Kommunikation mit anderen Komponenten des SoC zu ermöglichen.	Da der RRAM-Beschleuniger in verschiedenen Implementierungen eingesetzt werden kann, muss die Schnittstelle universell gestaltet sein.
Softwareumgebung für den RRAM-Beschleuniger	Zusätzliche Softwarekomponenten wie Bibliotheken, Treiber oder Compiler-Erweiterungen könnten erforderlich sein, um den RRAM-Beschleuniger aus	Die Nutzung des RRAM sollte auf abstrakter Ebene möglich sein, ohne die internen Details kennen zu müssen.

	anderen SoC-Komponenten heraus anzusteuern.	
IoT-Verarbeitungsplattform	Eine Plattform muss gewählt werden, die die gewünschte IoT-Anwendung realisieren kann und erweiterbar ist, sowohl für Sicherheitserweiterungen als auch für den RRAM-Beschleuniger.	Die Plattform sollte quelloffen sein, um Erweiterbarkeit und akademische Wiederverwendbarkeit zu gewährleisten (z. B. eine RISC-V-Plattform).
Randbedingungen für die IoT-Plattform	Die Plattform muss ggf. Randbedingungen wie Leistung, Energieverbrauch, Energieeffizienz und andere berücksichtigen.	Diese Randbedingungen beeinflussen die Definition der Plattform, die Schnittstelle zum RRAM-Beschleuniger und die Umsetzung von Sicherheitsmaßnahmen.
Sicherheitsbedrohungen	Die Sicherheitsbedrohungen für das Anwendungssystem sollten definiert werden.	Diese Bedrohungen beeinflussen die Auswahl und Entwicklung entsprechender Gegenmaßnahmen.
Sicherheitsfunktionen / Sichere Geräteidentität	Die vorgeschlagenen Sicherheitsfunktionen müssen ausgewählt, entwickelt und in die Verarbeitungsplattform integriert werden.	Die Sicherheitsfunktionen können die Auswahl der Verarbeitungsplattform und die Definition der Randbedingungen beeinflussen.
RRAM-Schnittstelle von Seiten der Verarbeitungsplattform	Eine Schnittstelle von der Verarbeitungsplattform zum RRAM-Beschleuniger muss definiert werden.	Die Schnittstelle kann sowohl die definierten Randbedingungen beeinflussen als auch von diesen beeinflusst werden.

AP 2: Systementwurf

Im Arbeitspaket zwei brachte UzL ihre Arbeiten im Teilarbeitspaket 2.1 Intelligent Embedded Electronics ein, das zudem durch UzL geleitet wurde. Hier verantwortete die UzL die Entwicklung eines Hardware-Frontends für die lokale, KI-gestützte Verarbeitung von Sensordaten im Edge-Bereich. Ziel war es, eine geeignete Hardwarearchitektur zu entwerfen, die den speziellen Anforderungen des Edge Computing gerecht wird – insbesondere in Bezug auf Energieeffizienz, Rechenleistung, Echtzeitfähigkeit, Sicherheit, Vertraulichkeit und Wirtschaftlichkeit.

Als primäre Zielanwendungen dienten die Sicherstellung der Vertrauenswürdigkeit der Rechennoten und die Beschleunigung von KI-Algorithmen. Hierfür wurden geeignete Technologien identifiziert und spezifische Hardwaremodelle entwickelt bzw. angepasst.

Technische Schwerpunkte waren:

Auswahl eines geeigneten RISC-V-Kerns:
Auf Basis existierender Open-Source-Implementierungen wurde ein geeigneter RISC-V-Prozessorkern als Ausgangspunkt gewählt, abgestimmt auf die Anforderungen der geplanten Anwendungen. Aufgrund bereits zuvor gesammelter Erfahrungen mit dem Chipyard Framework der UC Berkley, wurde auf dieses und die darin enthaltenen RISC-V Prozessoren zurückgegriffen. Konkret wurde der Rocket-Kern in der Variante RV64GC integriert. Ein Vergleich verschiedener RISC-V-Kerne und die Durchführung von Benchmarks hatten diesen als geeigneten Kandidaten ergeben.

Entwurf der Schnittstellen:
Dem RISC-V-Kern muss sowohl die Kommunikation mit dem als Co-Prozessor ausgelegtem Beschleuniger, als auch mit den Hardwarekomponenten zur Sicherheit und Vertrauenswürdigkeit möglich sein.

Die Komponenten zur Sicherheit der Prozessorumgebung wurden als externe Komponenten umgesetzt. Da das Host-System und das Sicherheits-System sogar separate Systems-on-Chip darstellen, wurden hier ausschließlich serielle Schnittstellen zur Off-Chip-Kommunikation betrachtet. Die finale Umsetzung erfolgte mittels Quad-SPI. Dieses stellt einen Kompromiss zwischen Einfachheit und Performanz dar. Die verhältnismäßig geringe Anzahl an Leiterbahnen, die dieses Busprotokoll erfordert, stellt die Leichtgewichtigkeit sicher, ohne den Durchsatz zu sehr einzuschränken. Zudem benötigte das Gesamtsystem noch einen externen Speicher in Form einer SD-Karte, welche dieses Protokoll ebenfalls unterstützt und damit die Vereinheitlichung der Datenkommunikation ermöglichte.

Um ein Maximum an Effizienz in der Ausführung von KI-Aufgaben durch den Beschleuniger zu ermöglichen, wurde davon abgesehen nur einzelne Operationen im Beschleuniger umzusetzen. Die Bewegung von Daten zwischen Speicher und CPU macht in modernen Architekturen einen entscheidenden Anteil der benötigten Energie aus. Um den Weg der Daten über den Prozessor zu vermeiden, wurde der Beschleuniger für die Anbindung mittels *Direct Memory Access* (DMA) ausgelegt. Sofern auch der Sensor DMA unterstützt, können so beispielsweise Bilder einer Kamera direkt als Ganzes vom Beschleuniger verarbeitet werden und anschließend werden nur das Klassifikationsergebnis oder andere relevante Metriken an den Prozessor weitergegeben, der dann diese Ergebnisse in die Cloud übermitteln kann oder Kontrollflussaufgaben anhand des Ergebnisses lenken kann.

SDI-basierter Secure Boot
Zur Bereitstellung einer sicheren Rechenumgebung, in der Vertrauenswürdigkeit und Integrität von Code und Daten gewahrt bleiben und Dienste wie Remote Attestation oder Data Sealing möglich sind, wurde ein quelloffenes Silicon-Root-of-Trust-Projekt (OpenTitan) mit einem RISC-V-basierten Trusted Execution Environment (TEE) kombiniert. Um die Sicherheit der des Host System-on-Chip (Host SoC) und des TEE zu gewährleisten, wurde OpenTitan (RoT SoC), um als SDI eingesetzt werden zu können.

Für den sicheren Systemstart wird die SHA-3-Einheit der RoT SoC genutzt, um mithilfe einer speziell entwickelten Firmware das Bootabbild inklusive TEE-Daten und Linux-Kernel zu hashen (1). Anschließend wird ein Abgleich mittels eines Referenz-Hash vom Referenz-Image, der im sicheren, nichtflüchtigen Speicher des RoT SoC abgelegt ist durchgeführt (2). Daraufhin initiiert das *Secure Boot Module* (SBM) den Start des Host-SoC. Der Vorgang ist schematisch in Abbildung 1 dargestellt. Nur bei erfolgreicher Prüfung wird das Hostsystem gestartet. Andernfalls bleibt der Startvorgang blockiert. Dies stellt sicher, dass keine unvertrauenswürdige Software in das System eingeschleust wird.

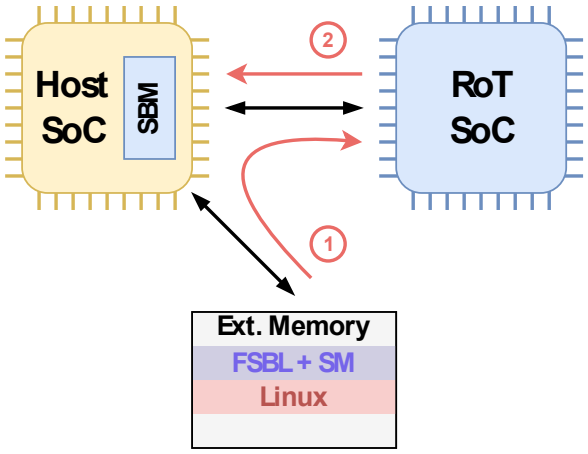

Abbildung 1: Systemübersicht bestehend aus dem Root-of-Trust System-on-Chip (RoT SoC) rechts und dem Hostsystem (Host SoC) links. Auf dem externen Speicher befindet sich das Bootabbild welches geladen und überprüft wird.

Ein Demonstrator bestehend aus zwei physisch verbundenen Entwicklungsboards wurde umgesetzt: Ein Board übernimmt die Rolle des RoT SoC, das andere stellt das RISC-V-basierte Host SoC unter Linux dar. Das Host-SoC beinhaltet den bereits erwähnten Rocket-Kern, der ebenfalls umfänglich adaptiert wurde. Insbesondere wurde die Hardware des Secure Boot Modules hinzugefügt. Der Aufbau der Validierungsplattform ist in Abbildung 2 zu sehen.

Zur Demonstration eines Secure-Boot-Ansatzes mit handelsüblichen Komponenten wurde zusätzlich das NXP Secure Element SE051 integriert. Dieses bietet die für Secure Boot benötigten kryptografischen Funktionen und erlaubt eine flexible Anpassung an unterschiedliche Zielsysteme. Die erfolgreiche Integration bestätigt die Praxistauglichkeit des Konzepts mit marktverfügbaren Komponenten. Das resultierende Gesamtsystem ist ebenfalls in Abbildung 2 illustriert.

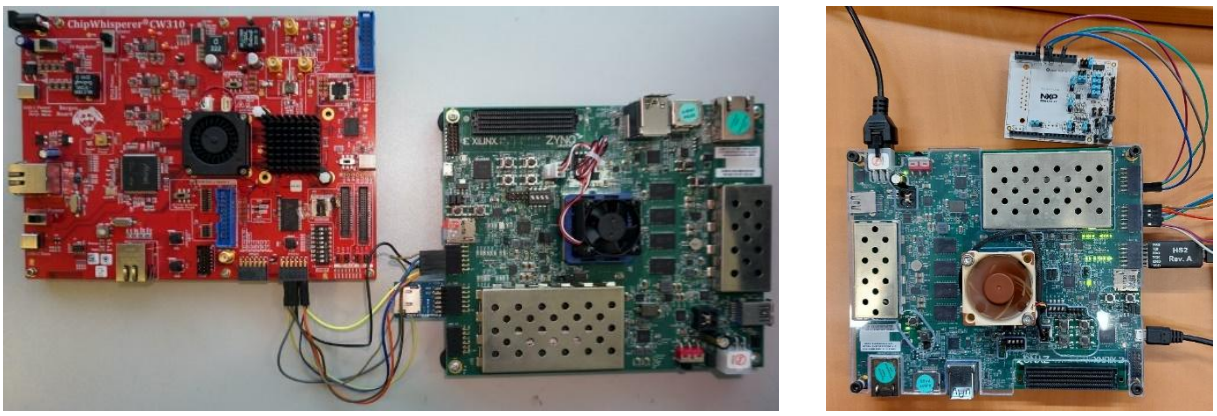

Abbildung 2: Links: Host-SoC mit Root-of-Trust-SoC, Rechts: Host-SoC mit NXP Secure Element

ASCON-basierter Vertrauensanker für ein leichtgewichtiges SDI (ASCON-SDI)

Neben der Integration sicherer Prozessorumgebungen auf bestehenden Vertrauensankern wurden auch ressourcenschonende Alternativen zur Umsetzung solcher Anker untersucht. Ziel war es, vergleichbare Sicherheitsfunktionen bei geringerem Hardwareaufwand bereitzustellen.

Ein zentraler Bestandteil ist der Algorithmus Ascon-128, der als Gewinner der CAESAR-Wettbewerbskategorie für leichtgewichtige Anwendungen hervorging und inzwischen vom NIST für kryptografische Einsatzzwecke empfohlen wird. Ascon-128 bietet vier Betriebsmodi, darunter authentifizierte Verschlüsselung/Entschlüsselung und kryptografische Hashfunktionen. Diese wurden erfolgreich implementiert. Der Funktionsumfang wurde zudem durch eine Physically Unclonable Function (PUF) erweitert, die zusätzliche Möglichkeiten zur Signierung und Zufallszahlenerzeugung bietet.

Ein ASCON-basierter Vertrauensanker inklusive PUF (ASCON-SDI) als Erweiterung des RISC-V-Prozessor wurde innerhalb des Projektes umgesetzt. Der Ablauf einer Verschlüsselung mittels ASCON ist in Abbildung 3 schematisch dargestellt.

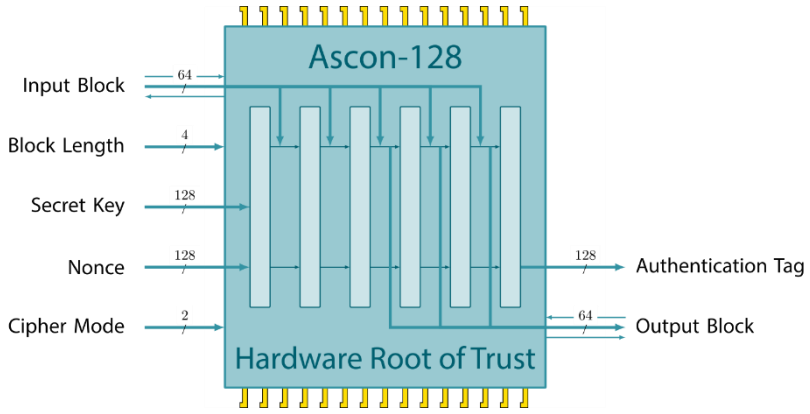

Abbildung 3.1: Skizze zur Anbindung von der Ascon-128 Einheit.

Erweiterung der SDI mit sicherem Mechanismus für In-Feld Tests von SoC

Zusätzlich wurde ein hybrider Software-/Hardware-Ansatz mit geringem Overhead in das SDI integriert, um ein sicheres Testen des Host SoC und dessen Komponenten zu ermöglichen. Der Test basiert auf (a) dem auf einem Keyed-Hash Message Authentication Code (KMAC), der gerätespezifische, sichere und gültige Signaturen ohne Aliasing bereitstellt, und (b) dem SoC-Prozessor für die Testplanung, wodurch die DUT-Verfügbarkeit erhöht wird. Der vorgeschlagene Ansatz bietet sowohl On-Chip- als auch Remote-Testfunktionen.

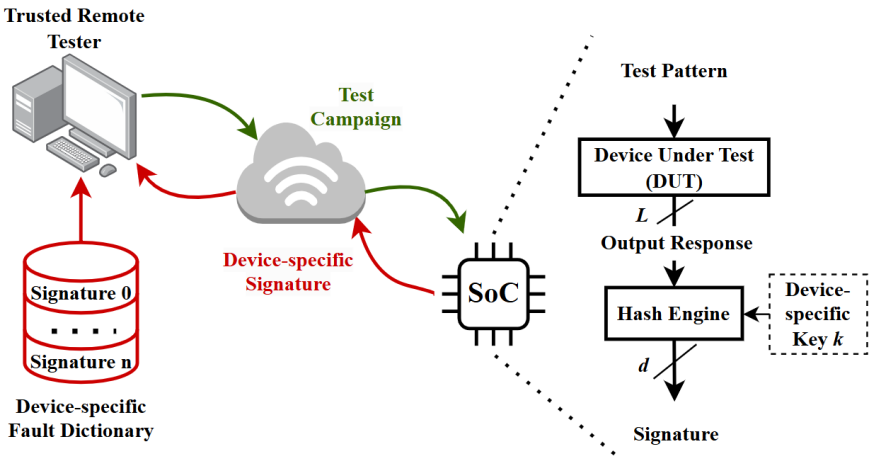

Abbildung 3.2: Erweiterung des SDI für einen KMAC-basierten In-Feld Test

RRAM-basierter rekonfigurierbarer Beschleuniger

Die Basis des memristiven Beschleunigers, der von UzL innerhalb des DAIS-Projektes entwickelt wurde, ist eine Lookup-Tabelle (LUT), die die analogen Eigenschaften von RRAM-Crossbars nutzt, um beliebige Boolsche Logik zu implementieren.

Das Herzstück der Architektur bildet die RRAM-Crossbar selbst, an deren Zeilen (Wordlines) Spannungen angelegt werden. Diese Spannungen repräsentieren einen Vektor logischer Variablen. An den Spaltenausgängen (Bitlines) lassen sich daraufhin Spannungen messen, die einzelnen Mintermen einer logischen Formel entsprechen. Entgegen typischen Vector-Matrix-Multiplikationseinheiten basierend auf RRAM-Crossbars, erfolgt hierbei keine strombasierte Berechnung. Stattdessen entsteht aus dem Widerstandsnetzwerk, das durch die Eingangsbelegung definiert wird, ein komplexer Spannungsteiler.

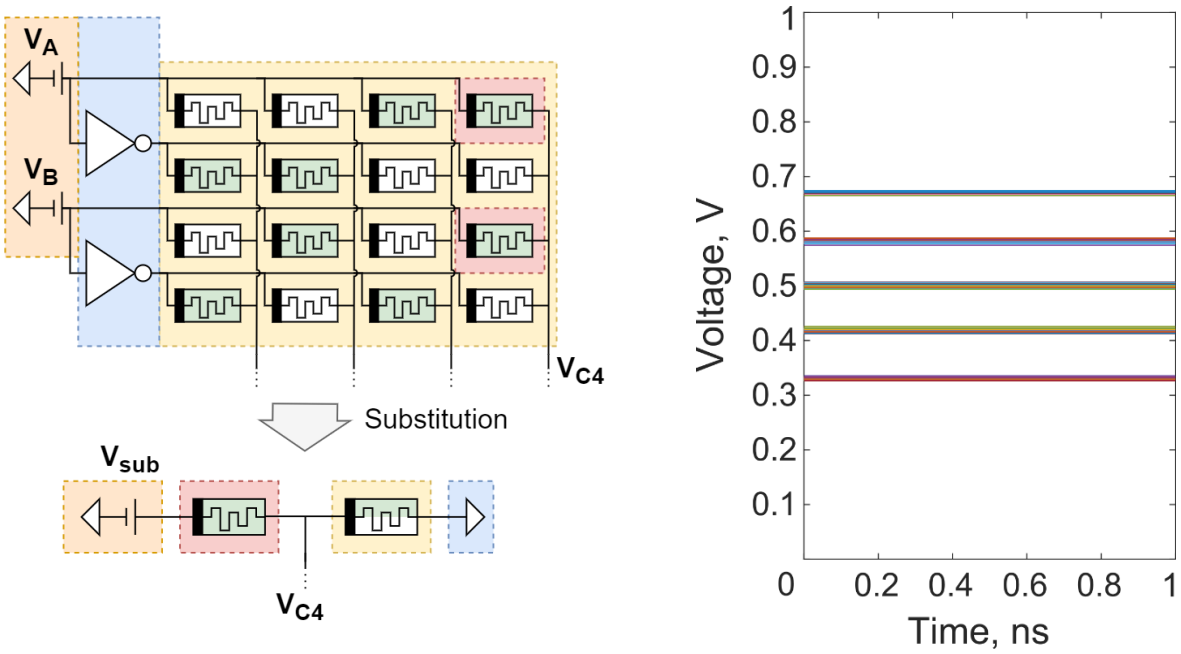

Abbildung 4: Links: Reduktion des Spannungsteilers in der Crossbar,
Rechts: Mögliche Ausgangsspannungen für alle Kombinationen von Eingangsspannungen.

Die logische Funktion wird dabei durch gezielte Programmierung der Memristoren realisiert: Je nach Konfiguration besitzen diese einen hohen oder niedrigen Widerstand. Ein niedriger Wert entspricht dabei einem nicht negierten Eingang als Teil der Konjunktion. Durch die freie Programmierbarkeit der einzelnen Zellen lässt sich beliebige Boolesche Logik realisieren. Abbildung 5 zeigt beispielhaft die Abbildung logischer Funktionen auf der memristive Struktur.

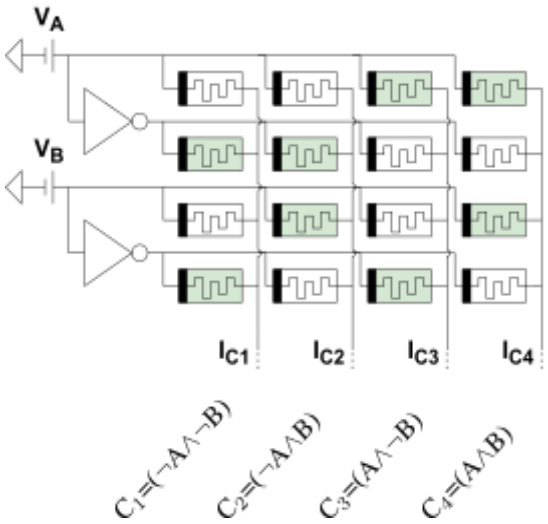

Abbildung 5: Für eine Boolsche Logik mit den Eingängen A und B werden alle möglichen Minterme auf die Spalten einer Crossbar abgebildet. Grüne Zellen sind dabei niederohmig, weiße hochohmig.

Die Ausgangssignale der Crossbar werden anschließend durch eine nachgelagerte Schaltung verarbeitet. Diese kombiniert die gemessenen Spannungen, um die Disjunktion der Spalten durchzuführen. Zudem erfolgt eine Verstärkung, sodass die analogen Ausgangssignale der LUT mit der umgebenden digitalen Peripherie kompatibel sind.

Zur Validierung wurde die gesamte Look-Up Tabelle (LUT) in Spice simuliert, funktional verifiziert und evaluiert. Ein Full-Custom-Layout wurde mithilfe von Cadence EDA-Tools unter Verwendung einer generischen 45-nm-Prozessbibliothek implementiert. Das Layout ist in Abbildung 6 dargestellt, nebst relevanter Kenndaten in Tabelle 1.

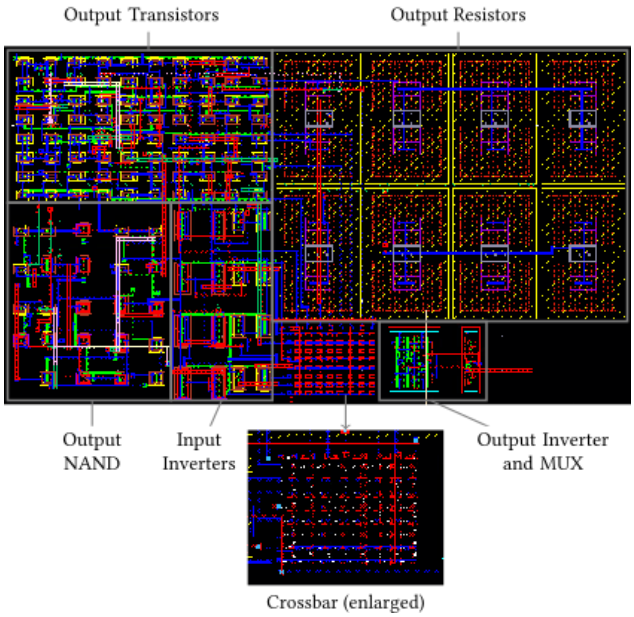

Abbildung 6: Layout der LUT in 45nm Strukturgröße mit Kennzeichnung der funktionalen Einheiten.

Stabilität	Die LUT ist stabil für alle zu erwartenden Inter- und Intra-Zellen-Variabilitäten der verwendeten Memristoren
Zeit	Die längstmögliche Verzögerung bei einem Lesezugriff der LUT beträgt 390 ps. Dies resultiert in einer Arbeitsfrequenz von 2.5 GHz.
Fläche	Ein LUT in 45nm-Technologie beansprucht 189 μm^2.
Energie	Ein Lesezugriff der LUT kostet maximal 0.46 pJ.

Tabelle 1 Relevante Kenndaten zur Performanz der entwickelten LUT-Architektur.

In Kombination mit einer klassischen Routing-Architektur bildet diese memristive LUT die Grundlage für einen embedded FPGA (eFPGA), der sich als anwendungsspezifischer Beschleuniger vielseitig einsetzen lässt.

AP3: Sichere Softwareumgebung

Arbeitspaket 3 adressierte die Software-Seite der integrierten HW/SW-Komponenten, d.h. Firmware, Middleware und Anwendungssoftware für die betrachteten Anwendungsszenarien. Hierbei beteiligte sich UzL an den Teilarbeitspaketen AP 3.2 Enabling Software und AP3.4: Security and Privacy.

Für die Sicherheitskomponenten des Rechenknotens musste Software entwickelt werden, die insbesondere die Orchestrierung der Komponenten und der einzelnen Schritte im sicheren Bootvorgang übernehmen konnte. So wurden die kryptographischen Primitive wie bereits in Arbeitspaket zwei geschildert zwar auf Hardware ausgeführt und durch diese beschleunigt, die Flusskontrolle erfolgte jedoch auf der höheren Abstraktionsebene der Software. Zusätzlich hervorzuheben ist außerdem die softwareseitige Generierung der Schlüssel. Während diese in Hardware zu generieren zwar möglich ist, limitiert dies entscheidend die Flexibilität. Weitere der Softwareebene zuzuordnende Aufgaben waren beispielsweise die Vorbereitung des Bootabbildes für den Secure Boot und die Entwicklung von Software zu Ansteuerung des NXP Secure Elements durch den Host-Prozessor.

Die Software die später auf dem Rechenknoten ausgeführt wurde ist nicht als Teil von AP3 zu betrachten, sondern als Teil der Demonstratoren und wird entsprechend später gesondert erörtert.

Auch der Beschleuniger benötigte einige Softwarekomponenten um seine Nutzung zu ermöglichen. Diese fallen entsprechend in die Kategorie der Enabling Software. Um KI-Anwendungen adressieren zu können, müssen gängige Frameworks für die Umsetzung entsprechender Modelle unterstützt werden. Da die Architektur des RRAM-basierten Beschleunigers sich an konventionellen FPGA-Architekturen orientiert, wurde das quelloffene HLS4ML Framework genutzt, das es ermöglicht neuronale Netze auf Boolesche Logik abzubilden. Dieses wurde unter Berücksichtigung der besonderen Eigenschaften von RRAM und der Architektur des Beschleunigers so konfiguriert, dass der Beschleuniger die meisten gängigen KI-Modelle auf Basis von Pytorch ausführen kann. Da der Beschleuniger weiterhin auf seine Simulation begrenzt ist, dient das Mapping der KI-Anwendung primär der Ableitung von Kennzahlen zur generellen Machbarkeit und Performanz und beinhaltet nicht die explizite Konfiguration der Hardware.

Die KI-Anwendung, die im Zuge der Demonstratoren umgesetzt wurde, wird an entsprechender Stelle genauer beleuchtet.

AP4: Systemintegration

Dieses Arbeitspaket konzentrierte sich auf den Aufbau der industriellen Demonstratoren. Dabei wurden von UzL praxisnahe Anwendungsfälle aus SC8 adressiert. Die komplexen Beispiele dienen der Validierung der in DAIS entwickelten Komponenten und Prozesse.

Die in SC8 definierten Anwendungsfälle fanden Eingang in das Teilarbeitspaket AP4.3 Intelligent Transport and Mobility mit Fokus auf Drohnenschwärmen. Ziel war die Demonstration KI-gestützter Verfahren auf Edge-Ebene zur Steigerung der Transportsicherheit, Verfügbarkeit und Leistungsfähigkeit. Hierfür wurden eine dezentrale, redundante Sensordatenverarbeitung sowie eine Inter-Drohnen-Kommunikation realisiert. Darauf aufbauend kamen Methoden zum Einsatz

wie sicheres Drohnenmanagement im Fehlerfall, hierarchische Missionsplanung, autonome zielbasierte Steuerung, automatische Flugroutengenerierung sowie Kollisionserkennung mit Ausweichverhalten.

UzL trug zur Erfüllung dieser an Anforderung in zweifacher Hinsicht bei:

- Durch den sicheren Rechenknoten wurde das Problem der Vertraulichkeit im Drohnenschwarm adressiert. Die durch den Secure Boot gewährleistete Integrität des Knotens und die durch ASCON-SDI bereitgestellte Signierung können garantieren, dass keine böswilligen Aktoren die Organisation im Drohnenschwarm ausnutzen können, um Fehlverhalten herbeizuführen.
- Der Beschleuniger kann eingesetzt werden um eine energiesparende Echtzeitverarbeitung der durch die Drohne aufgenommenen Daten, insbesondere Infrarotkameraaufnahmen, zu ermöglichen. Durch den dedizierten Beschleuniger wird hierbei eine Balance erreicht zwischen der Nutzung des CPUs in der Drohne, welche für Bildverarbeitungs- und KI-Aufgaben ungeeignet ist und dem nicht praktikablem Ansatz die Drohne um eine leistungshungrige GPU zu erweitern.

Die Validierung der eingebrachten Komponenten anhand konkreter Anwendungen erfolgte in Arbeitspaket fünf.

AP5: Integration und Validierung der Demonstratoren

In Arbeitspaket fünf erfolgte die finale Integration der Demonstratoren und ihr Einsatz anhand der vordefinierten Anwendungsfälle. UzL war dabei in den Anwendungsfällen von Teilarbeitspaket *AP5.3 Smart City Neighbourhood* beteiligt. Ursprünglich beteiligt in *AP 5.1 Ai-Powered Smart Industry,* hat UzL ihre Tätigkeit innerhalb eines Amendments des Projektes zu AP5.3 gewechselt, da dieses besser zum Profil passte und die dort umgesetzten Demonstratoren sich aus SC8 ableiteten, an der UzL ebenfalls beteiligt war. Konkret war UzL an drei Anwendungsfällen direkt beteiligt:

1. Detect and Avoid and 3D Surface Scanning Demonstrations
2. Self-Provisioning of Drone Fleet for the Transportation of Goods
3. Fire Monitoring and Firefighting System

Die Arbeiten in Arbeitspaketen zwei und drei bezüglich der Sicherheitskomponenten waren explizit darauf ausgerichtet die Anforderungen der Demonstratoren eins und zwei zu erfüllen. In diesen sind die Drohnen nicht einzeln und unabhängig beteiligt, sondern agieren als Schwarm welches als Gesamtsystem gesteuert und koordiniert werden muss. Entsprechend ist hier eine Sicherung der Kommunikation, sowie die Garantie der Vertrauenswürdigkeit der einzelnen Schwarmteilnehmer notwendig.

Ersteres wurde ermöglicht durch die Signierung und Verschlüsselung, die durch die ASCON-Erweiterung des Prozessors bereitgestellt werden. Durch die leichtgewichtige Generierung der Signaturen ist es möglich, sämtliche Kommunikation im Drohnenschwarm dahingehend abzusichern, dass keine Nachrichten von nicht vertrauenswürdigen Teilnehmern eingeschleust werden können, die zu einer nachteiligen Verhaltensänderung des Drohnenschwarms führen könnten. Außerdem kann durch die Verschlüsselung der Nachrichten ausgeschlossen werden, dass versendete Nachrichten außerhalb des Schwarms empfangen und verarbeitet werden können, was wiederum ermöglichen würde extern auf ihn einzuwirken. Der Aufbau des Host-

SoC mit ASCON-SDI -Erweiterung ist in Abbildung 7 zu sehen. Die Betrachtung die Anbindung an den RoT-SoC erfolge bereits anhand Abbildung 2.

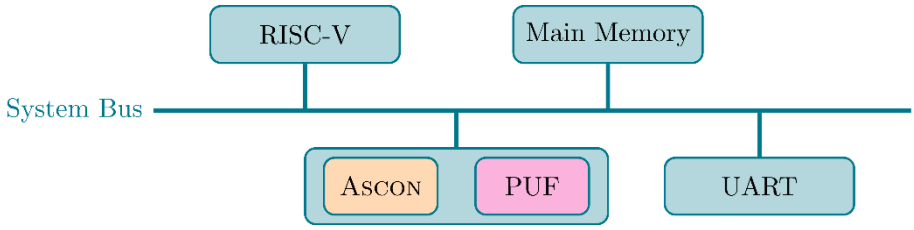

Abbildung 7: SoC mit ASCON-SDI Erweiterung

Da es wie eingangs bereits erwähnt nicht möglich war zusätzliche Hardware an Board der Drohne unterzubringen, wurden die Sicherheitskomponenten parallel zur eigentlichen Demonstration des Schwarmverhaltens validiert. Dies erfolgte in enger Rücksprache mit den Partnern, die die Drohnen bereitstellten und die Demonstration leiteten, Beyond Vision und ESC Aerospace. Innerhalb dieser Zusammenarbeit wurden bereits frühzeitig realistische Anforderungen formuliert und diese konnten durch die Validierungsplattform erfolgreich erfüllt werden. Eine Portierbarkeit in reelle Anwendungsszenarien ist damit sichergestellt.

Parallel zu den Sicherheitsaspekten wurde der Anwendungsfall drei primär durch den von UzL entwickelten RRAM-basierten Beschleuniger adressiert. Hier wurde die Drohne ausgestattet, um Menschen in verschiedenen Rettungsszenarien zu identifizieren. Insbesondere sollte während Gebäudebränden sichergestellt werden, dass sich keine Personen dem Feuer nähern, und Menschen sollten aus der Drohnenperspektive sowohl im Wald als auch in Gewässern detektiert werden. Die Drohne wurde daher mit einer Infrarotkamera ausgestattet in deren Aufnahmen sich Personen durch ihre Körperwärme vom Hintergrund abheben. Aufgrund ethischer und logistischer Bedenken, konnte der Datensatz von Aufnahmen während eines echten Feuers nicht durch die Demonstratorleitung generiert werden. Es wurde daher ein Datensatz von Menschen in wäldlicher Umgebung generiert und parallel der frei Verfügbare Datensatz *CAMEL Dataset for Visual and Thermal Infrared Multiple Object Detection and Tracking* verwendet, der Infrarotbilder von Personen und Fahrzeuge in einer städtischen Umgebung enthält.

Beim Entwurf des Modells zur Personendetektion für Drohneneinsätze auf einem RRAM-basierten FPGA mussten mehrere Einschränkungen berücksichtigt werden:

- *Kompaktheit:* Das Modell musste möglichst ressourcenschonend gestaltet werden, da große FPGAs aufgrund des begrenzten Energieangebots auf Drohnen ungeeignet sind.
- *Sparsity:* Um Speicher zu sparen, wurde ein Netzwerkdesign verwendet, bei dem viele Gewichtungen den Wert null annehmen.
- *Quantisierung:* Für die Umsetzung auf Hardware wurden die kontinuierlichen Gewichte und Bias-Werte aus dem Training in diskrete Werte umgewandelt. Unterschiedliche Netzwerkschichten nutzen dabei unterschiedliche Präzisionen, um durch kleinere Datentypen zusätzlichen Platz zu sparen – ohne die Genauigkeit stark zu beeinträchtigen.
- *Ressourcenteilung:* Faltungskerne im CNNs wurden mehrfach genutzt, um Hardware zu teilen und Eingabedaten effizient sequentiell zu verarbeiten. Dabei wurde ein Kompromiss zwischen Leistung und Ressourcenverbrauch angestrebt.

Auf Basis dieser Vorgaben wurde ein besonders kompaktes Modell entwickelt. Die Eingabebilder wurden auf 168×128 Pixel reduziert, um einen Mittelweg zwischen Genauigkeit und Effizienz zu finden. Das Netzwerk besteht aus zwei Hauptstufen: Zunächst drei Faltungsschichten mit Aktivierung und Pooling, anschließend eine Umformung in einen Vektor, dem drei vollvernetzte Schichten folgen. Die Klassifikation erfolgt zwischen „Person erkannt“ und „keine Person erkannt“. Dies wurde in Rücksprache mit den anderen Partnern als ausreichend erachtet, da die Drohne im ersten Schritt nur erkennen muss ob Personen überhaupt anwesend sind, um die Daten dann an eine leistungsfähigere Basis oder einen menschlichen Operator weiterzuleiten.

Trotz seiner geringen Größe mit nur 7.953 Gewichten erreicht das Modell eine Erkennungsgenauigkeit von über 97 %. Die Sparsity betrug dabei beinahe 70%. Die Netzwerkarchitektur ist in Abbildung 8 zu sehen.

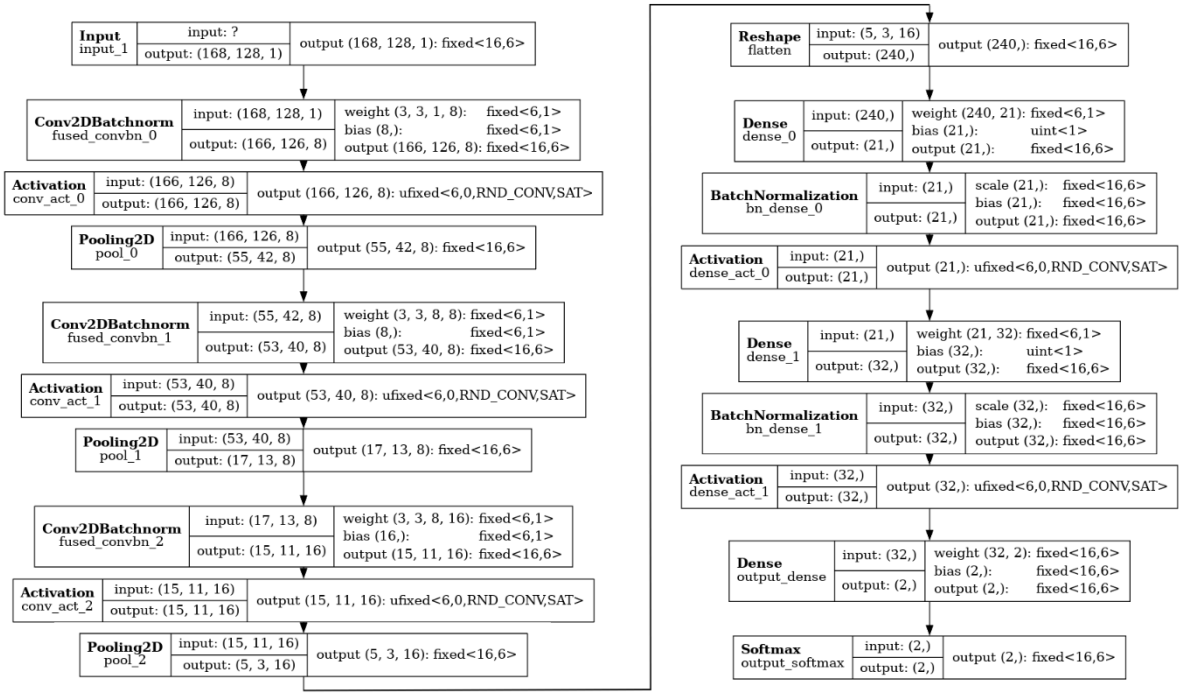

Abbildung 8: Verwendetes Model zur Detektion von Personen in langwelligen Infrarotbildern.

Aus den präzisen Simulationsergebnissen mittels SPICE konnte die Performanz für die Ausführung der Detektion auf dem Beschleuniger zuverlässig geschätzt werden, ohne ihn vollständig zu simulieren. Das gesamte System in SPICE zu simulieren war nicht möglich, da es die zulässige Komplexität bei weitem überstieg. Insbesondere eine Analyse von Latenz und Durchsatz zeigte, dass eine Detektion innerhalb von 5,2 Mikrosekunden pro Bild erfolgen kann. Somit kann eine Detektion ohne weiteres in Echtzeit erfolgen.

Auch die Ansprüche an die Personenerkennung wurde in Zusammenarbeit mit allen am Demonstrator beteiligten Partnern festgelegt. Die Simulation des Beschleunigers zeigte, dass diese Ansprüche erfüllt und übertroffen werden konnten. Da RRAM noch eine junge und experimentelle Technologie ist, sind die Möglichkeiten zur Umsetzung noch beschränkt, doch konnten die Eignung der entworfenen Architektur und des umgesetzten Modells eindeutig belegt werden.

AP6: Bekanntmachung, Verwertung und Standardisierung

UzL präsentierte die Arbeit und Errungenschaften innerhalb von DAIS sowohl auf ihrer Webpräsenz als auch auf dem Universitätscampus. Ebenfalls fließen die Erkenntnisse aus DAIS bereits aktiv in die universitäre Lehre an der UzL ein.

Veröffentlichungen im Rahmen des Projektes durch UzL wurden zudem in angesehenen wissenschaftlichen Journalen und auf Konferenzen öffentlichkeitswirksam präsentiert.

AP7: Projektmanagement

UzL hat keine Projektweiten Managementaufgaben übernommen, hat jedoch an den Meetings zur Steuerung des Projektes aktiv partizipiert. Im Arbeitspaket zwei hat UzL ein Teilarbeitspaket geleitet, dessen Ergebnisse erfolgreich und im Einklang mit der Projektverlängerung fristgerecht abgeliefert wurden.

Gegenüberstellung der vorgegebenen Ziele mit dem Erreichten

- Im Einklang mit der Verlängerung des Projektes, haben die Arbeiten der UzL länger angedauert als ursprünglich veranschlagt. Dies hatte jedoch keine inhaltlichen Konsequenzen.
- UzL wechselte früh im Projekt von einer Teilnahme an Teilarbeitspaket 5.1 zur Teilnahme in Teilarbeitspaket 5.3. Dies veränderte geringfügig die inhaltliche Ausrichtung der adressierten Anwendungsfälle, nicht aber die Beiträge der UzL selbst.
- Ursprünglich war die Integrationen der durch UzL entwickelten Komponenten in die als Demonstratorplattform eingesetzten Drohnen geplant. Aufgrund der eingeschränkten Zuladung der Drohnen und der Verwendung von Evaluationsboards durch die UzL, war dies nicht möglich. Es wurden jedoch äquivalente Validierungsplattformen außerhalb der Drohnen erfolgreich umgesetzt.

Notwendigkeit und Angemessenheit der geleisteten Arbeit

Auch nun nach Abschluss des Projektes, sind die Anwendung von KI im Internet of Things, sowie die Absicherung der Teilnehmer des IoT und deren Kommunikation weiterhin hochgradig relevante Themen. Unter Hinzunahme neuer Paradigmen und Technologien, ist es ein hoch anspruchsvolles, aber auch vielversprechendes Unterfangen den Stand der Technik über bestehende kommerzielle Lösungen hinaus zu erweitern.

Insbesondere angesichts einer akuten Verschiebung der Abhängigkeiten im globalen Markt, ist es von hoher Bedeutung, diese Innovationen innerhalb Europas und insbesondere innerhalb Deutschlands zu erreichen. Das teilweise nicht zu vernachlässigende Risiko radikal neue Ansätze zu verfolgen ist ohne öffentliche Förderung nur einigen wenigen Unternehmen vorbehalten, die zunehmend den Markt monopolisieren. Das Ziel der Unabhängigkeit und Eigenständigkeit ist daher nur erreichbar durch geförderte Kooperation innerhalb Europas.

Die im Rahmen des Projektes geleistete Arbeit folgte dem genehmigten, in der Gesamtvorhabensbeschreibung dokumentierten Arbeitsplan.

Nutzen und Verwertbarkeit der Ergebnisse

UzL hat die Arbeit innerhalb von DAIS in folgenden Veröffentlichungen verwertet:

- Grothe, Philipp, Saleh Mulhem, and Mladen Berekovic. “An Almost Fully RRAM-Based LUT Design for Reconfigurable Circuits." *International Symposium on Applied Reconfigurable Computing*. Cham: Springer Nature Switzerland, 2023.
- Mulhem, Saleh, et al. "Secure Software/Hardware Hybrid In-Field Testing for System-on-Chip." *2024 IFIP/IEEE 32nd International Conference on Very Large Scale Integration (VLSI-SoC)*. IEEE, 2024.
- Najork, Daniel, et al. „Single Engine Architecture for Hardware Root of Trust”, Journal of Cryptographic Engineering. Springer 2025

Forschungsprojekte

Die im Projekt gewonnenen Erkenntnisse leisten auch einen wertvollen Beitrag zur Umsetzung weiterer Forschungsvorhaben. Die vertieften Kompetenzen im Bereich der Entwicklung sicherer Systeme und KI-Beschleuniger schaffen eine solide Grundlage für die Verwertung in parallel laufenden und nachfolgenden Projekten. Insbesondere in folgende Projekte ist das gewonnene Wissen bereits eingeflossen oder wird einfließen:

- A-IQ Ready erforscht ähnlich DAIS Multiagentensysteme im Internet of Things. Insbesondere die Erkenntnisse zum effizienten Beschleunigerentwurf konnten hier bereits einfließen und haben durch die gegenüber DAIS nach hinten verlagerte Projektlaufzeit schon zu direkten Weiterentwicklungen geführt.
- VE-Jupyter zielt ebenfalls auf die Sicherstellung der Vertrauenswürdigkeit in der Mikroelektronik ab. Beispielsweise wurden dort ebenfalls Physically Unclonable Functions eingesetzt, sodass eine thematische Abstimmung erfolgen konnte.
- Gemeinsam mit einem Großteil des ursprünglichen Konsortiums gibt es bereits Bemühungen einen direkten Nachfolger von DAIS zu begründen. Der Fokus soll hierbei voraussichtlich mehr auf den Rechenknoten und Sensoren liegen, als auf der Vernetzung. Da UzL bereits innerhalb von DAIS den Fokus auf diese Aspekte gelegt hat, ist eine sehr effiziente Verwertung zu erwarten.
- SASVI zielt auf einen leichtgewichtigen, energieeffizienten Ansatz ab, um ressourcenbeschränkte Geräte des Internet-der-Dinge (IoT) abzusichern. Dafür wurden die Erkenntnisse des SDI-basierten Secure Boot genutzt und ein neuer leichtgewichtiger Vertrauensanker entworfen, welcher den Energie- und Ressourcenverbrauch des Gesamtsystems verbessert.

Universitäre Lehre

Die im Projekt entwickelte Architekturen und Komponenten, sowie die gewonnenen Erkenntnisse fließen in die Lehre ein. Auch die im Projekt identifizierten Herausforderungen und deren Lösungsansätze werden dabei berücksichtigt. So erhalten Studierende einen praxisnahen Einblick in aktuelle Fragestellungen und technologische Entwicklungen, die in den kommenden Jahren weiter an Bedeutung gewinnen werden.

Fortschritte bei anderen Stellen

Während der Projektlaufzeit wurde regelmäßig eine Überprüfung des Stands der Technik durchgeführt, um die Relevanz der getätigten Forschung sicherzustellen. Nach Nr. 2.1 BNBest-

BMBF 98 sind keine Ergebnisse Dritter bekannt geworden die für die Durchführung des Projektes relevant sind.

Erfolgte oder geplante Veröffentlichungen

Die erfolgten Veröffentlichungen sind im Abschnitt *Nutzen und Verwertbarkeit der Ergebnisse* gelistet.