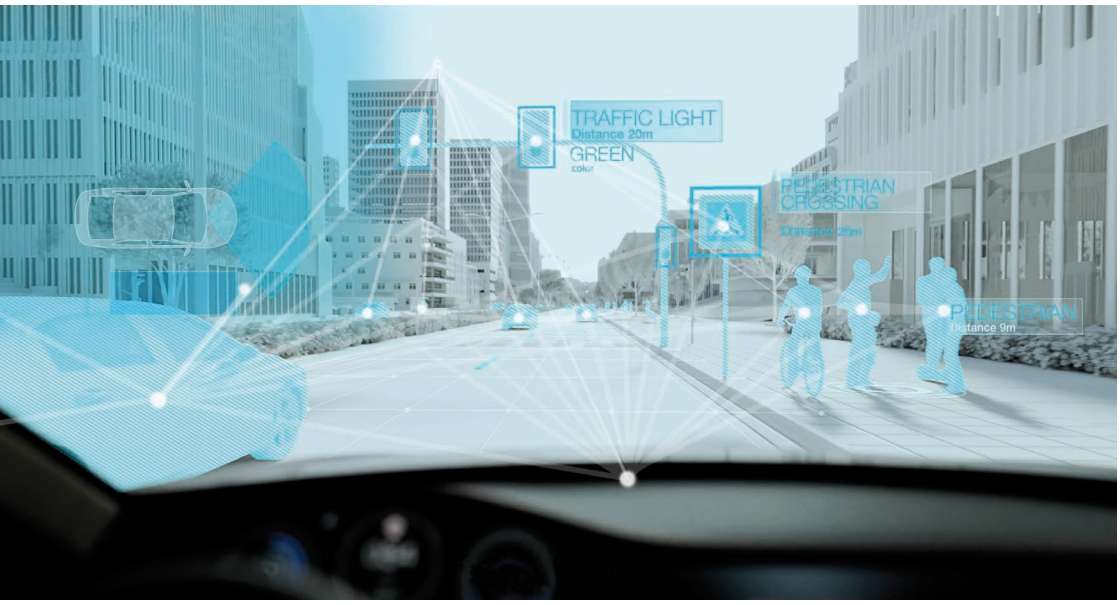
Deep Learning für IoT Anwendungen der nächsten Generation - Teil I

Das EU-Projekt VEDLIoT zeigt, wie Deep Learning und künstliche Intelligenz helfen, den Anwendungsbereich von IoT-Systemen zu erweitern.



Das Internet der Dinge (engl. Internet of Things - IoT), ein Netzwerk miteinander verbundener Geräte, die mit Sensoren und Software ausgestattet sind, hat zum Einen die Art und Weise revolutioniert, wie wir mit der Welt um uns herum interagieren, und gibt uns zum Anderen die Möglichkeit, Daten wie nie zuvor zu sammeln und zu analysieren.

Im Zuge des technologischen Fortschritts und der fortschreitenden Automatisierung werden immer mehr Gegenstände mit Konnektivität und Sensorfunktionen ausgestattet und damit Teil des IoT-Ökosystems. Es wird erwartet, dass die Zahl der aktiven IoT-Systeme bis 2027 29,7 Milliarden erreichen wird, was einen erheblichen Anstieg gegenüber den 3,6 Milliarden Geräten im Jahr 2015 bedeutet. Dieses exponentielle Wachstum erzeugt eine enorme Nachfrage nach Lösungen, um die Herausforderungen im Hinblick auf Rechenleistung, Zuverlässigkeit und Sicherheit von IoT-Anwendungen zu bewältigen. Insbesondere industrielles IoT, Automotive und Smart Homes sind drei wichtige Bereiche mit spezifischen Anforderungen, die jedoch einen gemeinsamen Bedarf an effizienten IoT-Systemen haben, um optimale Funktionalität und Leistung zu ermöglichen.

AIoT-Architekturen

Die Steigerung der Effizienz von IoT-Systemen und die Freisetzung ihres Potenzials kann durch künstliche Intelligenz (KI) und die Schaffung von AIoT-Architekturen (Artificial Intelligence of Things) erreicht werden. Durch den Einsatz hochentwickelter Algorithmen und Methoden des maschinellen Lernens befähigt KI IoT-Systeme, intelligente Entscheidungen zu treffen, große Datenmengen zu verarbeiten und wertvolle Erkenntnisse zu gewinnen. Diese Integration treibt beispielsweise die operative Optimierung in industriellen IoT Systemen voran, ermöglicht fortschrittliche autonome Fahrzeuge und bietet intelligentes Energiemanagement sowie personalisierte Erfahrungen in intelligenten Häusern.

Deep Learning

Unter den verschiedenen KI-Algorithmen ist Deep Learning, das künstliche neuronale Netze nutzt, aus mehreren Gründen sehr gut für IoT-Systeme geeignet. Einer der Hauptgründe ist seine Fähigkeit, automatisch aus Sensor-Rohdaten zu lernen und Merkmale zu extrahieren. Dies ist besonders wertvoll bei IoT-Anwendungen, bei denen die Daten unstrukturiert und verrauscht sein können oder komplexe Beziehungen aufweisen. Deep Learning ermöglicht es IoT-Systemen außerdem, Echtzeit- und Streaming-Daten effizient zu verarbeiten. Diese Fähigkeit ermöglicht eine kontinuierliche Analyse und Entscheidungsfindung, was bei zeitkritischen Anwendungen wie Echtzeitüberwachung, vorausschauender Wartung oder autonomen Steuersystemen entscheidend ist.

Trotz der zahlreichen Vorteile, die Deep Learning für IoT-Systeme mit sich bringt, gibt es bei der Implementierung inhärente Herausforderungen, beispielsweise hinsichtlich Energieeffizienz, Zuverlässigkeit und Sicherheit, die angegangen werden müssen, um das Potenzial voll auszuschöpfen. Das Projekt Very Efficient Deep Learning in IoT (VEDLIoT) stellt Lösungen für diese Herausforderungen bereit.

VEDLIoT: Verbessertes IoT in Kombination mit effizientem Deep Learning

Eine Übersicht über die verschiedenen VEDLIoT-Komponenten ist in Bild 1 dargestellt.

Im Rahmen des VEDLIoT-Projekts wird das IoT mit Deep Learning integriert, um Anwendungen zu beschleunigen und die Energieeffizienz des IoT zu optimieren. VEDLIoT erreicht diese Ziele durch den Einsatz mehrerer Schlüsselkomponenten:

- **Spezialisierte KI-Beschleuniger:** Diese Beschleuniger werden zur Optimierung der Rechenleistung eingesetzt und ermöglichen eine erhebliche Senkung des Energieverbrauchs ohne Leistungseinbußen. Darüber hinaus verbessern sie die Gesamteffizienz von Deep-Learning-Modellen, was schnellere Inferenzen und eine bessere Skalierbarkeit für IoT-Anwendungen ermöglicht.

- **Hardware-sensitives Pruning und Quantisierung:** Durch den Einsatz von hardware-sensitiven Pruning- und Quantisierungsverfahren beschleunigt VEDLIoT Deep-Learning-Modelle und reduziert den Speicherbedarf bei gleichbleibend hoher Genauigkeit.

Autor:
VEDLIoT Konsortium
Jens Hagemeyer
Koordinator

https://www.uni-bielefeld.de/fakultaeten/technische-fakultaet/arbeitsgruppen/kognitronik-sensorik

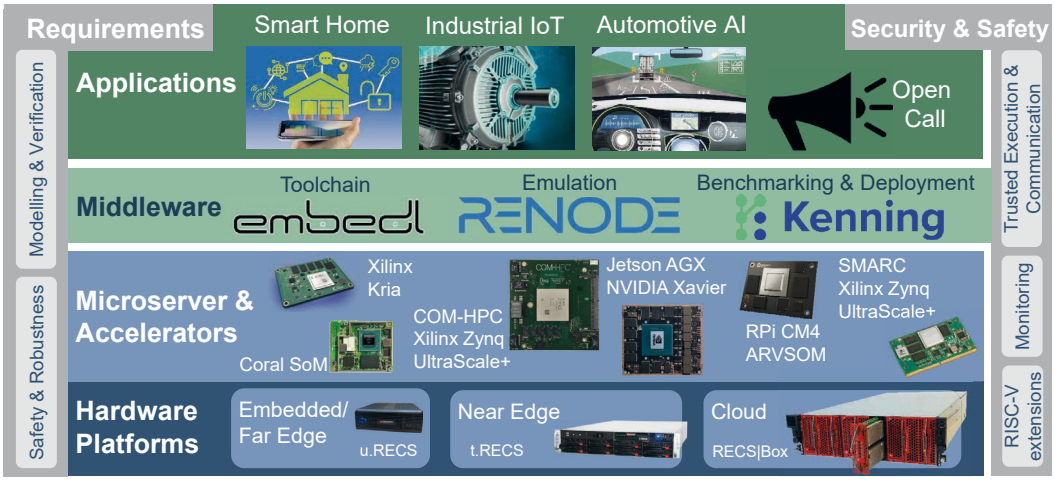


Bild 1: Überblick über die verschiedenen VEDLIoT Komponenten

- **Sicherheit und Zuverlässigkeit:** Die Verwendung von hardwarebasierten vertrauenswürdigen Arbeitsumgebungen gewährleistet die Integrität und Zuverlässigkeit der Deep-Learning-Modelle, die in IoT-Umgebungen eingesetzt werden. Darüber hinaus hilft ein spezielles Architekturframework bei der Berücksichtigung und Integration von Sicherheits- und ethischen Aspekten während des Requirements Engineering.

- **Anpassbare Hardware-Plattformen:** VEDLIoT nutzt individuell anpassbare Hardware-Plattformen und ermöglicht so maßgeschneiderte Lösungen, die spezifische IoT-Anforderungen erfüllen und Deep-Learning-Algorithmen optimieren.

Bild 2: Smart mirror Demonstrator als Teil der Smart Home Anwendung in VEDLIoT

VEDLIoT [1] konzentriert sich auf einige Anwendungsfälle, wie z. B. bedarfsorientierte Interaktionsmethoden in Smart Homes (Bild 2), industrielle IoT-Anwendungen wie Zustandsüberwachung von Elektromotoren für prädiktive Instandhaltung (vorausschauende Wartung) oder Erkennung von Lichtbögen in Gleichstromverteilungen. Im Automobilbereich (Bild 3) wird eine automatische Notbremsung für Fußgänger (PAEB)-System betrachtet, welches verteilt, sowohl lokal im Fahrzeug als auch in entfernt in der Edge implementiert wird. (Bild 3).

VEDLIoT optimiert solche Anwendungsfälle systematisch in einem Bottom-up-Ansatz durch den Einsatz von Requirements Engineering und Verifikationstechniken, wie in Bild 1 dargestellt. Das Projekt kombiniert Expertenwissen aus verschiedenen Bereichen, um eine robuste Middleware zu schaffen, die die Entwicklung durch Tests, Benchmarking und Deployment-Frameworks erleichtert und letztlich die Optimierung und Effektivität von Deep-Learning-Algorithmen in IoT-Systemen sicherstellt. In den folgenden Abschnitten stellen wir die einzelnen Komponenten des VEDLIoT-Projekts kurz vor.

Spezialisierte KI-Beschleuniger

Es gibt verschiedene Beschleuniger für ein breites Anwendungsspektrum, von kleinen eingebetteten Systemen mit einem Leistungsbudget im Milliwattbereich bis hin zu leistungsstarken Cloud-Plattformen. Diese Beschleuniger werden auf der Grundlage ihrer Spitzenleistungswerte in drei Hauptgruppen eingeteilt, wie in Bild 4 dargestellt.

Ultra-Low-Power Beschleuniger

Die erste Gruppe ist die Ultra-Low-Power-Kategorie (< 3 W), die aus energieeffizienten Mikrocontroller-ähnlichen Kernen in Kombination mit kompakten Beschleunigern für spezifische Deep-Learning-Funktionen besteht. Diese Beschleuniger sind für IoT-Anwendungen konzipiert und bieten einfache Schnittstellen für eine leichte Integration. Einige Beschleuniger dieser Kategorie verfügen über Kamera- oder Audioschnittstellen, die effiziente Bild- oder Tonverarbeitungsaufgaben ermöglichen. Sie können eine generische USB-Schnittstelle bieten, die es ihnen ermöglicht, als Beschleunigergeräte zu fungieren, die an einen Host-Prozessor angeschlossen sind. Diese Ultra-Low-Power-Beschleuniger sind ideal für IoT-Anwendungen, bei denen Energieeffizienz und Kompaktheit eine wichtige Rolle spielen, und bieten eine optimierte Leistung für Deep-Learning-Aufgaben ohne übermäßigen Stromverbrauch.

Der VEDLIoT-Anwendungsfall der vorausschauenden Wartung ist ein gutes Beispiel und nutzt einen Ultra-Low-Power-Beschleuniger. Eines der wichtigsten Designkriterien ist der niedrige Stromverbrauch, da es sich um ein batteriebetriebenes kleines Gerät handelt, das extern an jedem Elektromotor installiert werden kann und den Motor mindestens drei Jahre lang ohne Batteriewechsel überwachen soll.

Low-Power Beschleuniger

Die nächste Kategorie ist die Low-Power-Gruppe (3 W bis 35 W), die auf eine breite Palette von Automatisierungs- und Automobilanwendungen abzielt. Diese Beschleuniger verfügen über Hochgeschwindigkeitsschnittstellen für externe Speicher und Peripheriegeräte sowie eine effiziente Kommunikation mit anderen Verarbeitungsgeräten oder Hostsystemen wie PCIe. Sie unterstützen modulare und auf Mikroservern basierende Ansätze und bieten Kompatibilität mit verschiedenen Plattformen. Darüber hinaus sind viele Beschleuniger dieser Kategorie mit leistungsstarken Anwendungsprozessoren ausgestattet, auf denen vollständige Linux-Betriebssysteme laufen und die eine flexible Softwareentwicklung und -integration ermöglichen. Einige Geräte dieser Kategorie enthalten spezielle anwendungsspezifische integrierte Schaltkreise (ASICs), während andere NVIDIAs eingebettete Grafikprozessoren (GPUs) integrieren. Diese Beschleuniger

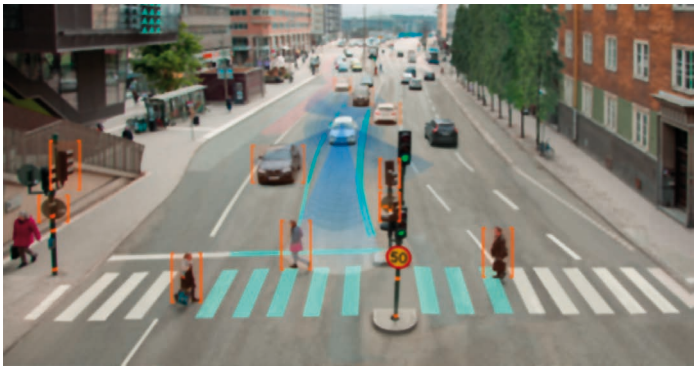


Bild 3: Automatische Notbremsung für Fußgänger, entwickelt als Teil der Automotive AI-Anwendung in VEDLIoT

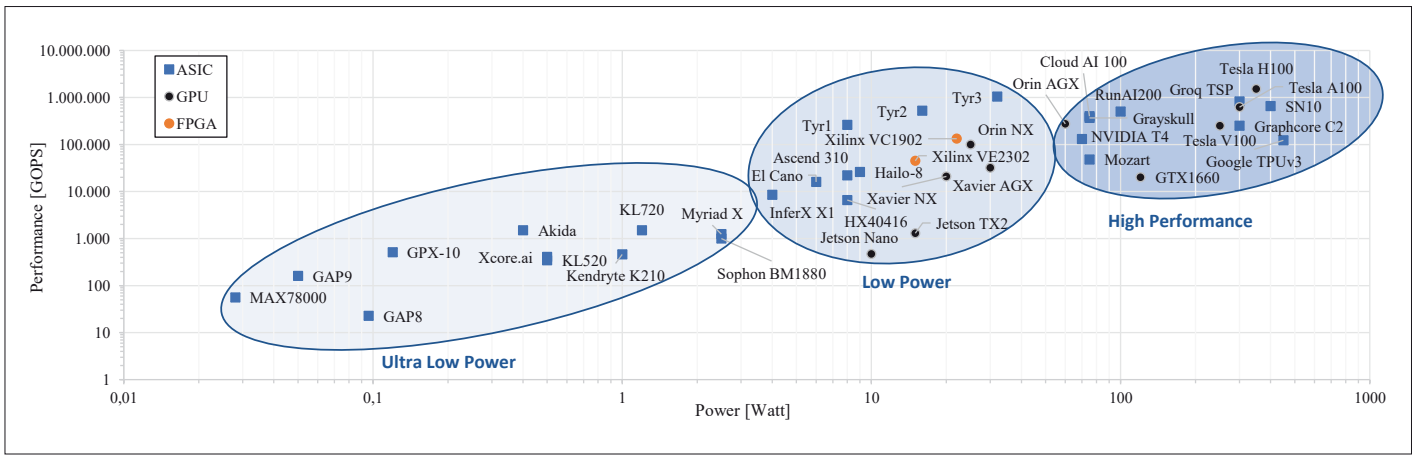


Bild 4: Überblick über verschiedene KI-Beschleuniger

bieten ein ausgewogenes Verhältnis zwischen Stromverbrauch und Verarbeitungsleistung und eignen sich daher gut für verschiedene rechenintensive Aufgaben in der Automatisierung und im Automobilbereich.

High-Performance Beschleuniger

Die Hochleistungskategorie (> 35 W) von Beschleunigern ist für anspruchsvolle Inferenz- und Trainingsszenarien in Edge- und Cloud-Servern konzipiert. Diese Beschleuniger bieten eine außergewöhnliche Verarbeitungsleistung und eignen sich daher für rechenintensive Aufgaben. Sie werden in der Regel als PCIe-Erweiterungskarten eingesetzt und bieten Hochgeschwindigkeitsschnittstellen für eine effiziente Datenübertragung. Die Geräte dieser Kategorie haben eine hohe Rechenleistung, aber auch eine recht hohe thermische Entwurfsleistung (TDP). Zu diesen Beschleunigern gehören dedizierte ASICs, die für ihre spezielle Leistung bei Deep Learning-Aufgaben bekannt sind. Sie bieten beschleunigte Verarbeitungskapazitäten und ermöglichen schnellere Inferenz- und Trainingszeiten. Einige GPUs der Verbraucherklasse können ebenfalls in Benchmarking-Vergleiche einbezogen werden, um eine breitere Perspektive zu bieten.

Unterstützung bei der Auswahl der Beschleuniger

Die Auswahl des richtigen Beschleunigers aus der oben erwähnten breiten Palette der verfügbaren Optionen ist nicht einfach. VEDLIoT erleichtert dies jedoch, indem gründliche Bewertungen und Evaluierungen verschiedener Architekturen, einschließlich GPUs, Field-Programmable Gate Arrays (FPGAs) und ASICs durchführt wurden. Im Rahmen des Projekts [2] wurden die Leistung und der Energieverbrauch dieser Beschleuniger sorgfältig untersucht, um ihre Eignung für bestimmte Anwendungsfälle sicherzustellen.

Hardware-gesteuertes Pruning und Quantisierung

Trainierte Deep-Learning-Modelle weisen Redundanzen auf, die manchmal auf das 49-fache ihrer ursprünglichen Größe komprimiert werden können, ohne dass die Genauigkeit darunter leidet. Obwohl sich viele Arbeiten mit einer solchen Komprimierung befassen, zeigen die meisten Ergebnisse theoretische Geschwindigkeitssteigerungen, die sich nur manchmal in einer effizienteren Hardwareausführung niederschlagen, da sie die Zielhardware nicht berücksichtigen. Obwohl bereits verschiedene Frameworks für diese Schritte zur Verfügung stehen, variiert ihre Interoperabilität, was zu unterschiedlichen Ergebnissen führt. VEDLIoT adressiert diese Herausforderungen durch eine hardwarenahe Modelloptimierung unter Verwendung von ONNX, einem offenen Format zur Darstellung von Machine-Learning-Modellen, dass die Kompatibilität mit dem aktuellen offenen Ökosystem gewährleistet. Darüber hinaus dient Renode (Antmicro Ltd, „Renode IoT development Framework“, www.renode.io), ein Open-Source-Simulationsframework, als funktionaler Simulator für komplexe heterogene Systeme, der die Simulation kompletter System-on-Chips (SoCs) und die Ausführung derselben Software auf der Hardware ermöglicht.

Zusätzlich verwendet VEDLIoT das EmbeDL-Toolkit zur Optimierung von Deep-Learning-Modellen (EmbeDL AB, „EmbeDL Model Optimization SDK“, https://www.embedl.com/product). Das EmbeDL-Toolkit bietet umfassende Werkzeuge und Techniken zur Optimierung von Deep-Learning-Modellen für einen effizienten Einsatz auf ressourcenbeschränkten Geräten. Durch die Berücksichtigung hardwarespezifischer Einschränkungen und Merkmale ermöglicht das Toolkit Entwicklern zu komprimieren, zu quantifizieren, zu beschneiden und Modelle zu optimieren, während gleichzeitig die Ressourcenauslastung minimiert und eine hohe Inferenzgenauigkeit beibehalten wird. EmbeDL konzentriert sich auf die hardwarenahe Optimierung und stellt sicher, dass Deep-Learning-Modelle effektiv auf Edge- und IoT-Geräten eingesetzt werden können, um das Potenzial für intelligente Anwendungen in verschiedenen Bereichen zu erschließen. Mit EmbeDL können Entwickler eine überragende Leistung, schnellere Inferenzen und eine verbesserte Energieeffizienz erzielen, was es zu einer unverzichtbaren Ressource für alle macht, die das Potenzial von Deep Learning in realen Anwendungen ausschöpfen wollen.

Sicherheit und Zuverlässigkeit

Da VEDLIoT darauf abzielt, Deep Learning mit IoT-Systemen zu kombinieren, wird die Gewährleistung von Sicherheit und Zuverlässigkeit entscheidend. Um diese Aspekte in den Mittelpunkt zu stellen, nutzt das Projekt vertrauenswürdige Ausfüh-

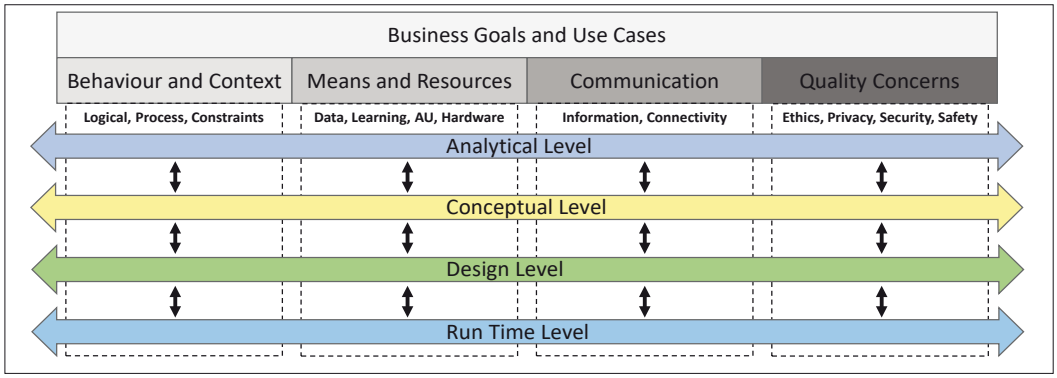


Bild 5: Architekturframework mit verschiedenen Ebenen

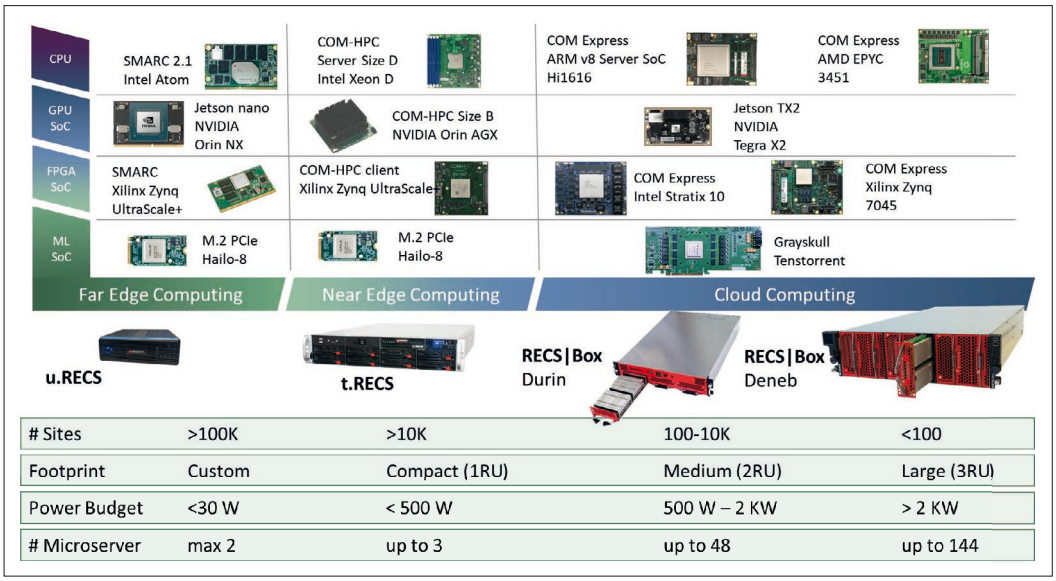
rungsumgebungen (TEEs) wie Intel SGX und ARM TrustZone zusammen mit Open-Source-Laufzeiten wie WebAssembly [3], [4]. TEEs bieten sichere Umgebungen, die kritische Softwarekomponenten isolieren und vor unbefugtem Zugriff und Manipulationen schützen. Durch die Verwendung von WebAssembly bietet VEDLIoT eine gemeinsame Umgebung für die Ausführung im gesamten Kontinuum, vom IoT über Edge bis hin zur Cloud.

Im Zusammenhang mit TEEs stellt VEDLIoT Twine [5] und WaTZ [6] als vertrauenswürdige Laufzeitumgebungen für Intels SGX bzw. ARMs TrustZone vor. Diese Laufzeitumgebungen vereinfachen die Softwareerstellung in sicheren Umgebungen durch die Nutzung von WebAssembly und seiner modularen Schnittstelle. Diese Integration überbrückt die Lücke zwischen vertrauenswürdigen Ausführungsumgebungen und AIoT und hilft Deep-Learning-Frameworks nahtlos zu integrieren. Innerhalb von TEEs, die WebAssembly verwenden, erreicht VEDLIoT einen hardwareunabhängigen, robusten Schutz vor böswilligen Eingriffen, wobei die Konfiguration sowohl der Daten als auch der Deep Learning-Modelle erhalten bleibt. Diese Integration unterstreicht das Engagement von VEDLIoT, kritische Softwarekomponenten zu sichern, eine sichere Entwicklung zu ermöglichen und datenschutzfreundliche AIoT-Anwendungen in Cloud-Edge-Umgebungen zu ermöglichen.

Spezielles Architekturframework

Darüber hinaus verwendet VEDLIoT ein spezielles Architekturframework [7], [8], wie in Bild 5 dargestellt, das dabei hilft, die Anforderungen und Spezifikationen von KI-Komponenten und traditionellen IoT-Systemelementen zu definieren, zu synchronisieren und zu koordinieren.

Dieses Rahmenwerk besteht aus verschiedenen architektonischen Ebenen, die sich mit den spezifischen Designbelangen und Qualitätsaspekten des Systems befassen, einschließlich Sicherheits- und ethischer Überlegungen. Durch die Verwendung dieser Architekturansichten als Vorlagen und deren ausfüllen, können Korrespondenzen und Abhängigkeiten zwischen den qualitätsbestimmenden Architekturansichten und anderen Entwurfsentscheidungen, wie z. B. der Konstruktion von KI-Modellen, der Datenauswahl und der Kommunikationsarchitektur identifiziert werden. Dieser ganzheitliche Ansatz gewährleistet, dass Sicherheits- und Ethikaspekte nahtlos in das Gesamtsystemdesign integriert werden. Er unterstreicht das Engagement von VEDLIoT für Robustheit und die Bewältigung neuer Herausforderungen in KI-gestützten IoT-Systemen.



# Sites	>100K	>10K	100-10K	<100
Footprint	Custom	Compact (1RU)	Medium (2RU)	Large (3RU)
Power Budget	<30 W	< 500 W	500 W – 2 KW	> 2 KW
# Microserver	max 2	up to 3	up to 48	up to 144

Bild 6: Überblick über die verschiedenen RECS-Plattformen, welche die VEDLIoT Hardwareplattform bilden.

Adaptierbare Hardware-Plattformen für IoT-Systeme

Traditionelle Hardware-Plattformen unterstützen nur homogene IoT-Systeme. RECS [9] hingegen, eine KI-fähige Mikroserver-Hardwareplattform ermöglicht jedoch die nahtlose Integration verschiedener Technologien. So ermöglicht sie eine Feinabstimmung der Plattform auf spezifische Anwendungen und bietet eine umfassende Cloud-to-Edge-Plattform. Alle RECS-Varianten haben das gleiche Design-Paradigma: eine dicht gekoppelte, hochintegrierte Kommunikationsinfrastruktur. Für die verschiedenen RECS-Varianten werden unterschiedliche Mikroservergrößen verwendet, von der Kreditkartengröße bis zur Tablet-Größe. Dies ermöglicht es den Kunden, für jeden Anwendungsfall und jedes Szenario die beste Variante zu wählen. Bild 6 gibt einen Überblick über die RECS-Varianten. Die drei verschiedenen RECS-Plattformen eignen sich für Cloud/Datencenter (RECS|Box), Edge (t.RECS) und IoT-Nutzung (u.RECS). Alle RECS-Server verwenden Mikroserver nach Industriestandard, die austauschbar sind und die Nutzung der neuesten Technologie durch den einfachen Austausch eines Mikroservers ermöglichen. Hardware-Anbieter dieser Mikroserver bieten ein breites Spektrum unterschiedlicher Rechenarchitekturen wie Intel-, AMD- und ARM-CPUs, FPGAs und Kombinationen aus einer CPU mit einer eingebetteten GPU oder einem KI-Beschleuniger.

Teil II beschäftigt sich mit verschiedenen IoT-Anwendungsfällen für effiziente KI.

Das VEDLIoT-Projekt wurde durch das Forschungs- und Innovationsprogramm Horizont 2020 der Europäischen Union unter Nr. 957197 gefördert.

Referenzen

[1] K. Mika, R. Griessl, J. Hagemeyer, P. Trancoso und M. Pasin, „VEDLIoT — Next generation accelerated AIoT systems and applications," 20th ACM International Conference on Computing Frontiers, 2023.

[2] R. Griessl, J. Hagemeyer, M. Porrmann und P. Trancoso, „Evaluation of heterogeneous AIoT Accelerators within VEDLIoT," DATE Conference 2023, 2023.

[3] J. Ménétrey, A. Grüter, P. Yuhala, J. Oeftiger, P. Felber, M. Pasin und V. Schiavoni, „A Holistic Approach for Trustworthy Distributed Systems with WebAssembly and TEEs," 27th International Conference on Principles of Distributed Systems (OPODIS 2023), 2023.

[4] J. Ménétrey, M. Pasin, P. Felber, V. Schiavoni, G. Mazzeo, A. Hollum und D. Vaydia, „A Comprehensive Trusted Runtime for WebAssembly with Intel SGX," IEEE Transactions on Dependable and Secure Computing, 2023.

[5] J. Ménétrey, M. Pasin, P. Felber und V. Schiavoni, „TWINE: An Embedded Trusted Runtime for WebAssembly," 37th IEEE International Conference on Data Engineering (ICDE'21), 2021.

[6] J. Ménétrey, M. Pasin, P. Felber und V. Schiavoni, „WaTZ: A Trusted WebAssembly Runtime Environment with Remote Attestation for TrustZone," 42nd IEEE International Conference on Distributed Computing Systems (ICDCS'22), 2022.

[7] H.-M. Heyn, E. Knauss und P. Pelliccione, „A compositional approach to creating architecture frameworks with an application to distributed AI systems," Journal of Systems and Software, 2023.

[8] S. K. Pradhan, E. Knauss und H.-M. Heyn , „Identifying and managing data quality requirements: a design science study in the field of automated driving," Software Quality Journal, 2023.

[9] K. Mika, F. Porrmann, K. Nils, R. Griessl und J. Hagemeyer, „RECS: A Scalable Platform for Heterogeneous Computing".36th IEEE International System-On-Chip Conference (SOCC). ◄