Testumgebung für Edge Computing in Intelligenten Transportsystemen (ITS)

Mathias Schneider, M.Eng.
Seifeddine Saadani, M.Eng.
Ruben Prokscha, B.Eng.
Prof. Dr.-Ing. Alfred Höß

Zusammenfassung

Die Verschiebung der Berechnung von Algorithmen aus dem Bereich des maschinellen Lernens von zentralen Serverstrukturen (Cloud), hin zu energieeffizienten Rechenplattformen in Steuergeräten etwa in Maschinen und Fahrzeugen (Edge) ist ein derzeitiger Trend der Industrie. Der Vorteil ist hierbei vielversprechend: niedrigere Verarbeitungszeiten ohne dauerhafte Verbindung zum Server. Dieses sogenannte *Edge Computing* und die damit zusammenhängenden Herausforderungen sind ein Hauptaugenmerk der Untersuchungen des europäischen Förderprojektes „*Artificial Intelligence for Digitalizing Industry*“ (AI4DI), in welchem Partner aus den verschiedensten Industriezweigen kollaborieren.

Gefördert wird das Projekt von der Europäischen Union im Rahmen des H2020 ECSEL Joint Undertaking sowie den involvierten nationalen Förderstellen, in Deutschland dem Bundesministerium für Bildung und Forschung (BMBF).

Die Automotive-Forschungsgruppe der Ostbayerischen Technischen Hochschule (OTH) Amberg-Weiden erforscht in diesem Projekt über einen Zeitraum von insgesamt drei Jahren die Optimierungsmöglichkeiten von KI-basierten Datenverarbeitungsprozessen in Edge-Netzwerken, welche als Teil des Anwendungsbereiches von Intelligenten Transportsystemen (ITS) einhergehen. Im abgeschlossenem ersten Projektjahr widmete sich die OTH Amberg-Weiden der Anforderungsanalyse, der Erarbeitung einer Systemarchitektur und der Konzipierung einer Testumgebung für die weiteren Untersuchungen zu Mobility-as-a-Service (MaaS).

Abstract

Moving machine learning inference from backbone servers closer to the actual process at the edge is a recent trend in the industry. Advantages are promising: lower latency for the processing pipeline without the requirement for a permanent connection to the cloud. This so-called *Edge Computing* and its related challenges are a major research objective of the investigation in the European project “*Artificial Intelligence for Digitalizing Industry*” (AI4DI), which assembles partners from various industrial sectors.

The project is funded in the program H2020 ECSEL JU by the European Union as well as the involved national authorities, including the Federal Ministry of Education and Research (BMBF) in Germany.

Within three years of project duration, the Automotive Research Team of the Ostbayerische Technische Hochschule (OTH) Amberg-Weiden investigates concepts for the optimization of data processing pipelines in Edge networks applied in the field of Intelligent Transportation Systems (ITS). In the first project year, OTH Amberg-Weiden’s activities comprised requirements analysis, system design and the setup of an ITS testbed for further investigations towards Mobility-as-a-Service (MaaS).

1 Einleitung

Bereits 2016 betitelte Bundeskanzlerin Angela Merkel Daten als „*Rohstoffe des 21. Jahrhunderts*“ [1]. Für die Veredelung der ungeheuren Datenmengen werden jedoch neue Methoden und Konzepte benötigt. Hierbei spielen Algorithmen aus dem Bereich der Künstlichen Intelligenz (KI) eine tragende Rolle in den kommenden Jahren. Um die Kommunikationslast zwischen Datenerfassung und Verarbeitung zu reduzieren und geringe Latenzzeiten zu gewährleisten, arbeitet das Konsortium des Forschungsprojektes *Artificial Intelligence for Digitalizing Industry* (AI4DI) unter anderem an der Entwicklung und Einsatz von neuen Edge Computing Plattformen, auf welchen die KI-Algorithmen gerechnet werden sollen. Das Projekt fällt unter dem Schirm *Electronic Components and Systems for European Leadership Joint Undertaking* (ECSEL JU) und umfasst ein Gesamtbudget von etwa 30 Millionen Euro für den Zeitraum von drei Jahren (2019 – 2022) [2].

Die OTH Amberg-Weiden arbeitet zusammen mit den Partnern VTT, Vaisto (beide Finnland) und ITML (Griechenland) in der Wertschöpfungskette für das Transportwesen an einem „Last-Mile“ Anwendungsfall [3]. Dieser ordnet sich in das zukunftsträchtige Konzept des Mobility-as-a-Service (MaaS) ein, welches sich der Umsetzung von individuellem, multimodalem und betreiberübergreifendem Personenverkehr widmet. Innerhalb von „Last-Mile“ soll die Umsteigezeit eines Fahrgasts zwischen öffentlichen Nahverkehr und einem automatisierten Taxi optimiert werden. Die Problemstellung umfasst dabei mehrere Dimensionen: die präzise Vorhersage von Fahrzeiten der Fahrzeuge in Abhängigkeit des derzeitigen Verkehrsaufkommens sowie das sichere, automatisierte Anfahren der Bushaltestelle. Der Beitrag der OTH Amberg-Weiden beschäftigt sich mit der Optimierung von sensornaher Datenverarbeitung mittels Edge Computing.

Im Folgenden werden die Ergebnisse des ersten Projektjahres (Mai 2019 bis Mai 2020) zusammengefasst. Aufbauend auf der Erläuterung der Vorgehensweise bei der Anforderungsanalyse und für das Systemdesigns, wird auf die von der OTH Amberg-Weiden entwickelte Testumgebung für Intelligente Transportsysteme (ITS) genauer eingegangen, welche technologische Grundlagen für MaaS bereitstellt.

2 Anforderungsanalyse und System Design

Im ersten Projektjahr beschäftigte sich der Großteil der Arbeiten des Konsortiums gemäß des V-Modells mit der Anforderungsanalyse der Anwendungsfälle und dem Systemdesign. Um insbesondere industrieübergreifende Merkmale zu identifizieren, wurde eine standardisierte Analysemethode mittels der ISO 25010 (“*Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – System and software quality models*”) für die Evaluation herangezogen. Die Charakteristika, welche durch die ISO spezifiziert werden, wurden um zusätzliche, für KI besonders relevante Merkmale ergänzt. Hierunter fallen eine Abschätzung des Grades der Autonomität des Systems und der Datenqualität, welche für Machine Learning (ML) Verfahren von äußerster Wichtigkeit sind.

Während funktionale Anforderungen für alle Partner im Vordergrund stehen, sind Metriken für Modulariät, Zeitverhalten und Ressourcenausschöpfung von hoher Priorität für Mobility as a Service (MaaS) Betrachtungen, welche den Hauptanwendungsfall für das Forschungsteam der OTH Amberg-Weiden darstellen. Die Ergebnisse aus dieser Analyse sind mittels der Volere Anwendungsspezifikation [4] verfeinert und quantifiziert worden, so dass diese im letzten Arbeitspaket, der Systemvalidierung, abgeglichen werden können.

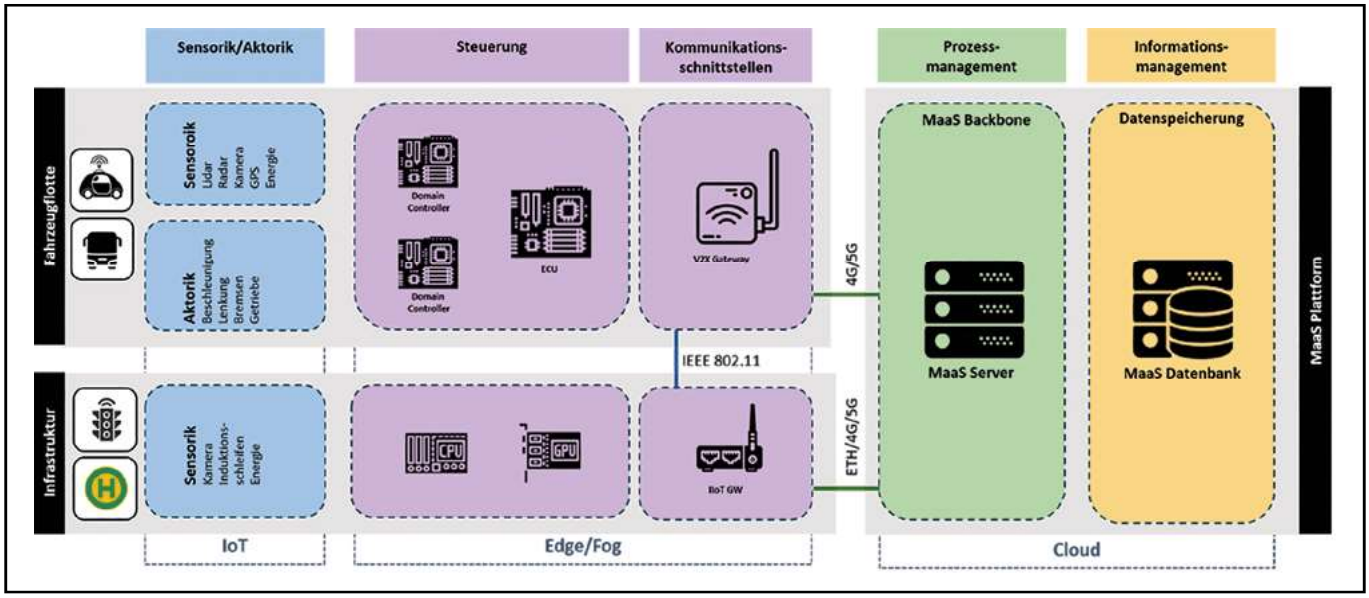

Abbildung 1: High-level Architektur für MaaS Anwendungsfall [8]

Für das Systemdesign verständigte sich das AI4DI-Konsortium auf eine harmonisierte Vorgehensweise für die Spezifikation einer hybriden Referenzarchitektur. Diese basiert auf drei bestehenden Modellen: *der Reference Architectural Model Industrie 4.0 (RAMI 4.0)* [5], der *IoT/IIoT 3D Reference Architecture* [6] und der *Industrial Internet Reference Architecture (IIRA)* [7]. Definiert werden die folgenden drei Dimensionen: die funktionalen Domänen und deren Partitionierung bezüglich der Systemkomponenten, die Systemeigenschaften, welche sich aus den Anforderungscharakteristiken ableiten, und die übergreifenden Systemfunktionalitäten wie u. a. Datensicherheit. Die Referenzarchitektur wurde dann gemäß des jeweiligen Anwendungsfalls spezifiziert (vgl. Abbildung 1).

3 Testumgebung für Intelligente Transportsysteme (ITS)

Um reproduzierbare MaaS-Szenarien testen zu können, wurde eine Testumgebung implementiert. Diese Umgebung modelliert Aspekte wie die Verteilung von Rechenprozessen auf Plattformen in der Edge und berücksichtigt dabei u. a. die Mobilität der Rechenknoten. Hiermit lassen sich Anwendungsfälle in skalierbarer Art und Weise in Bezug auf die Anzahl an Verkehrsteilnehmern (Fahrzeuge, Infrastrukturelemente und Fußgänger) simulieren – dies ist in der Realität aufgrund der beschränkten Ressourcen und der hohen Komplexität nicht möglich. Durch die Reproduzierbarkeit wird weiterhin gewährleistet, dass die entwickelten Algorithmen objektiv miteinander verglichen werden können. Somit wird ein iterativer Ansatz zur Problemoptimierung ermöglicht.

In dieser Testumgebung für ITS wird der *Hardware-in-the-Loop* Ansatz verfolgt. Hierbei werden Teile des Systems durch reale Komponenten dargestellt, wohingegen andere Teile z. B. durch bereits vorhandene Messdaten modelliert werden. Konkret bedeutet dies für den Aufbau, dass alle Rechenplattformen physikalisch in eine Simulationsumgebung eingebunden sind. Sensordaten hingegen, welche auf den Plattformen verarbeitet werden, werden durch historische Messdaten (z. B. Zähler von Induktionsschleifen) und der Mobilitätssimulation dem Testaufbau bereitgestellt.

Tabelle 1 listet einen Ausschnitt der verwendeten heterogenen Rechenplattformen und Hardwarebeschleuniger sowie deren wichtigsten Kenndaten auf. Hierunter lassen sich neue Architekturen wie Goolgle's *Tensor Processing Unit* (TPU) und die *Vision Processing Unit* (VPU) von Intel finden. Ergänzt wird das Portfolio durch geläufige *Graphics Processing Units* (GPUs) von NVIDIA. Alle Entwicklungsplattformen verfügen zudem über energieeffiziente *Advanced Risc Machine* (ARM) basierende *Central Processing Units* (CPUs). Ziel der Untersuchung wird es sein, insbesondere die Performance der Ausführung (engl. inference) von *Machine Learning* (ML) Modellen zu benchmarken. Der Trainingsschritt dieser Modelle ist auf den leistungsfähigeren Plattformen wie dem NVIDIA Jetson AGX Xavier möglich und wird in neuen Ansätzen, wie z. B. beim *Federated Learning*, angewandt [9]. Dieser Anwendungsfall ist allerdings fernab der derzeitigen Betrachtung, da die Modelle zentral auf einer leistungsstarken Serverplattform trainiert werden.

Metriken bezüglich der Performance der Rechenplattformen werden durch verschiedene Software und Hardwarekomponenten dauerhaft in Echtzeit gemessen. Hierunter fallen Messgrößen, welche die Auslastung der Plattformen (u .a. CPU und RAM) und der Kommunikation untereinander (Datendurchsatz und Latenz) beschreiben. Weil ein Aspekt der Untersuchungen darin liegt, die Daten energieeffizient in der Edge zu verarbeiten, wurde das Messsystem um eine Energiemessung erweitert. Bedingt durch die hohe Anzahl an Hardwareplattformen, welche gleichzeitig gemessen werden sollen, kann nicht auf kommerziell verfügbare Messsysteme zugegriffen werden. Stattdessen wurde eine prototypische Plattform entwickelt (vgl. Abbildung 2), welche gleichzeitige, nicht-invasive, Messungen von bis zu 18 Geräten mittels INA3221 Sensoren ermöglicht. Neben Ethernet, Wi-Fi und den Universal Asynchronous Receiver/Transmitter (UART), unterstützt

Tabelle 1 Übersicht der verwendeten Edge-Rechenplattformen [10] [11] [12]

	RaspberryPi 4B + Intel Movidius NCS2	Google Coral Dev Board	NVIDIA Jetson Nano	NVIDIA Jetson AGX Xavier
				
CPU	Quad-Core ARM Cortex A72	Quad-Core ARM Cortex A53	Quad-Core ARM Cortex A57 MPCore	8-core NVIDIA Carmel ARMv8
Speicher	4 GB LPDDR4	1 GB LPDDR4	4 GB LPDDR4	32 GB LPDDR4x
KI-Chip	Intel Movidius Myriad X Vision Processing Unit	Google Edge TPU Coprozessor	128 Core Maxwell GPU	512-core NVIDIA Volta GPU mit 64 Tensor Cores
Schnittstelle	USB 3.0	PCIe	PCIe	PCIe
Chip OPs	4 TOPs	4 TOPs	472 GFLOPs	32 TOPs
Leistung	2W + 7.6W	2W + 3.5W	5W/10W	10W/15W/30W
ML-Toolkit	OpenVino	TensorFlow Lite	TensorFlow, TensorRT	TensorFlow, TensorRT

das Gerät die kabellose Übertragung von Messwerten über weite Strecken mit der Long Range (LoRa) Funktechnologie. Diese Funktionalität ist insbesondere für den späteren Einsatz im realen Umfeld von Vorteil. Im weiteren Verlauf des Projektes soll außerdem eine „Dongle"-Variante entwickelt werden. Hiermit kann dann eine beliebige, einzelne Plattform in der Infrastruktur platziert und dauerhaft aus der Ferne überwacht werden.

Abbildung 2: Messsystem für Energieverbrauch

Die Testumgebung wird durch eine Mobilitätssimulation vervollständigt. Hierfür wurde sich mit den Projektpartnern auf die Simulationssoftware *Simulation of Urban Mobility* (SUMO) [13] verständigt. Diese ermöglicht es, die Positionen und Routen von Fahrzeugen und Infrastrukturelementen zu simulieren. In der Simulation wird dabei das Stadtzentrum von Tampere (Finnland) abgebildet, welches auch anderen Partner für Untersuchungen bezüglich des MaaS-Anwendungsfalls von Nutzen ist. Die Umgebung wurde möglichst realitätsnah in SUMO modelliert und erlaubt es, eine hohe Anzahl an Fahrzeugen, Infrastrukturelementen, Radfahrern und Fußgängern reproduzierbar zu simulieren. Die Simulation kann durch verschiedenste, real verfügbare Sensordaten (u. a. Induktionsschleifen) parametriert werden. Für die Zukunft ist eine Anbindung dieser Umgebung an eine Unity3D Simulation in Zusammenarbeit mit Partner Vaisto geplant. Diese ermöglicht es realitätsnahe Kameraaufnahmen, z. B. für das Training von Objektdetektionsalgorithmen, zu erzeugen. Durch weitere Schnittstellen, z. B. an einen Netzwerksimulator wie OMNet++ mittels Veins [14], können weitere Aspekte wie die Konnektivität (z. B. über IEEE 802.11x oder LTE) der Verkehrsteilnehmer im Softwarestack der Simulation integriert werden. Nichtsdestotrotz bleibt zwischen Realität und Simulation weiterhin eine gewisse Modellierungslücke (engl. "*reality gap*") vorhanden. Die Exploration dieser Lücke stellt derzeit ein eigenen Forschungsaspekt dar und kommt insbesondere beim Transfer von ausschließlich in einer Simulation trainierten Reinforcement Learning Modellen in die reale Umgebung zu tragen [15].

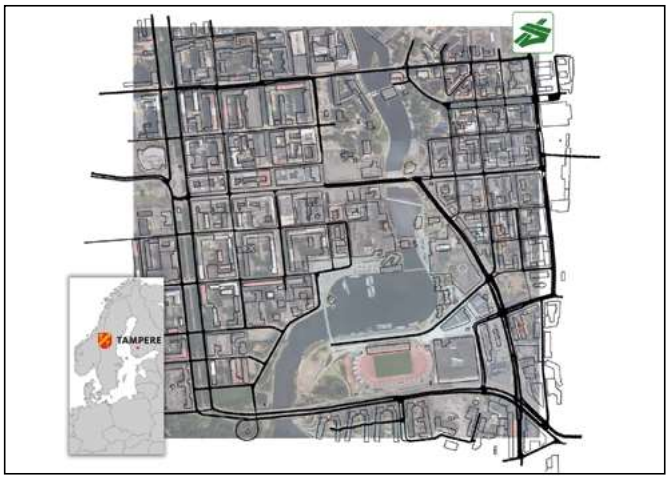

Abbildung 3: Modellierung der Testfelds in Tampere (Finnland) mittels SUMO

4 Ausblick

Im kommenden Projektjahr wird die Testumgebung für ITS dazu verwendet, die Performance der Inference verschiedenster ML-Verfahren (u. a. Objektdetektion und Segmentierungsalgorithmen) zu benchmarken. Diese Zwischenergebnisse werden darauffolgend verwendet, um konsekutive Datenverarbeitungspipelines zu definieren und auf eine statische Edge Netzwerktopologie zu verteilen. Des Weiteren werden dynamische Szenarien definiert, in welchen sich die Verfügbarkeit von Rechenplattformen oder die Art der Verarbeitungsschritte verändert. Ein intelligenter Agent soll basierend auf diesen Änderungen evaluieren, ob die Zuordnung zwischen Rechenprozessen und den Verarbeitungsplattformen optimiert werden kann und gegebenenfalls die notwendigen Operationen, z. B. die Verlagerung von Rechenschritten auf eine andere Plattform, durchführen.

Referenzen:

[1] Presse- und Informationsamt der Bundesregierung, „Podcast – Merkel: Wir müssen uns sputen", 12 3 2016. [Online]. Available: https://www.bundeskanzlerin.de/bkin-de/mediathek/bundeskanzlerin-merkel-aktuell/merkel-wir-muessen-uns-sputen-1009212. [Zugriff am 13 7 2020].

[2] „AI4DI", [Online]. Available: http://ai4di.eu/. [Zugriff am 13 7 2020].

[3] M. Schneider, S. Saadani und A. Höß, "Artificial Intelligence for Digitalizing Industry", in Forschungsbericht 2020, Amberg, Ostbayerischen Technische Hochschule Amberg-Weiden, 2019, pp. 6 – 11.

[4] J. Robertson und S. Robertson , "Volere Requirements Specification Template", Atlantic Systems Guild, [Online]. Available: https://www.volere.org/templates/volere-requirements-specification-template/. [Zugriff am 26 1 2020].

[5] P. Adolphs, H. Bedenbender, D. Dirzus, M. Ehlich, U. Epple, M. Hankel, R. Heidel, M. Hoffmeister, H. Huhle, B. Kärcher, H. Koziolek, R. Pichler, S. Pollmeier, F. Schewe, A. Walter, B. Waser und M. Wollschlaeger, "Reference architecture model industrie 4.0 (rami4. 0)", ZVEI and VDI, Status report, 2015.

[6] O. Vermesan, M. Eisenhauer, M. Serrano, P. Guillemin, H. Sundmaeker, E. Z. Tragos, J. Valino, B. Copigneaux, M. Presser, A. Aagaard, R. Bahr und E. C. Darmois, "The Next Generation Internet of Things – Hyperconnectivity and Embedded Intelligence at the Edge", in Next Generation Internet of Things – Distributed Intelligence at the Edge and Human Machine-to-Machine Cooperation, River Publishers Series in Communications, 2018.

[7] Industrial Internet Consortium, "The Industrial Internet of ThingsVolume G1: Reference Architecture", 2019.

[8] AI4DI Consortium, "D2.2 Report on HW/SW partitioning and subsystem level key architecture designs (initial report)", 2020.

[9] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konecny, S. Mazzocchi, B. McMahan, T. Van Overveldt, D. Petrou, D. Ramage und J. Roselander, "Towards Federated Learning at Scale: System Design", in Proceedings of the 2nd SysML Conference, Palo Alto, CA, USA, 2019.

[10] NVIDIA Corporation, "Technical Specifications", [Online]. Available: https://developer.nvidia.com/embedded/develop/hardware. [Zugriff am 13 7 2020].

[11] Intel Corporation, "Intel Neural Compute Stick 2", [Online]. Available: https://ark.intel.com/content/www/us/en/ark/products/140109/intel-neural-compute-stick-2.html. [Zugriff am 13 7 2020].

[12] Google LLC, "Dev Board datasheet", [Online]. Available: https://coral.ai/docs/dev-board/datasheet/. [Zugriff am 13 7 2020].

[13] P. A. Lopez, M. Behrisch, L. Bieker-Walz, J. Erdmann, Y.-P. Flötteröd, R. Hilbrich, L. Lücken, J. Rummel, P. Wagner und E. Wießner, "Microscopic Traffic Simulation using SUMO", in 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, Hawaii, USA, 2018.

[14] H. Noori, "Realistic urban traffic simulation as vehicular Ad-hoc network (VANET) via Veins framework", in 2th Conference of Open Innovations Association (FRUCT), Oulu, Finland, 2012.

[15] F. Muratore, M. Gienger und J. Peters, "Assessing Transferability from Simulation to Reality for Reinforcement Learning", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019.

Projektpartner:

Acknowledgement:

AI4DI has received funding within the Electronic Components and Systems for European Leadership Joint Undertaking (ECSEL JU) in collaboration with the European Union's H2020 Framework Programme and National Authorities, under grant agreement n° 826060.

Das Bundesministerium für Bildung und Forschung (BMBF) fördert das Projekt unter dem deutschen Titel „Elektronik-systeme für Künstliche Intelligenz in der digitalen Industrie – AI4DI" mit dem Teilvorhaben an der OTH Amberg-Weiden „KI-Algorithmen für die Optimierung von verteilten Datenverarbeitungsketten in heterogenen Rechnernetzen" unter der Fördernummer 16ESE0343.

ECSEL JU

GEFÖRDERT VOM Bundesministerium für Bildung und Forschung

Kontakt:

Mathias Schneider, M.Eng.

Ostbayerische Technische
Hochschule (OTH) Amberg-Weiden
Fakultät Elektrotechnik,
Medien und Informatik
Kaiser-Wilhelm-Ring 23
92224 Amberg

mat.schneider@oth-aw.de

Seifeddine Saadani, M.Eng.

Ostbayerische Technische
Hochschule (OTH) Amberg-Weiden
Fakultät Elektrotechnik,
Medien und Informatik
Kaiser-Wilhelm-Ring 23
92224 Amberg

se.saadani@oth-aw.de

Ruben Proschka, B.Eng.

Ostbayerische Technische
Hochschule (OTH) Amberg-Weiden
Fakultät Elektrotechnik,
Medien und Informatik
Kaiser-Wilhelm-Ring 23
92224 Amberg

ru.proschka@oth-aw.de

Prof. Dr.-Ing. Alfred Höß

Ostbayerische Technische
Hochschule (OTH) Amberg-Weiden
Fakultät Elektrotechnik,
Medien und Informatik
Vizepräsident Forschung und
Technologietransfer,
wissenschaftlicher Nachwuchs
Kaiser-Wilhelm-Ring 23
92224 Amberg

a.hoess@oth-aw.de