A Scalable RISC-V Hardware Platform for Intelligent Sensor Processing

Paul Palomero Bernardo,
Patrick Schmid, Oliver Bringmann
University of Tübingen
firstname.lastname@uni-tuebingen.de

Mohammed Iftekhar,
Babak Sadiye, Wolfgang Mueller
Paderborn University / Heinz Nixdorf Institute
firstname.lastname@hni.uni-paderborn.de

Andreas Koch
Technical University of Darmstadt
koch@esa.tu-darmstadt.de

Eyck Jentzsch
MINRES Technologies GmbH
eyck@minres.com

Axel Sauer, Ingo Feldner
Robert Bosch GmbH
firstname.lastname@de.bosch.com

Wolfgang Ecker
Infineon Technologies AG
wolfgang.ecker@infineon.com

***Abstract*—This paper presents a demonstrator chip for an industrial audio event detection application developed as part of the Scale4Edge project. The project aims at enabling a comprehensive RISC-V based ecosystem to efficiently assemble well-tailored edge devices. The chip is manufactured in Globalfoundries' 22FDX technology and contains a RISC-V CPU with custom Instruction-Set-Architecture Extensions (ISAX) for fast AI and DSP processing, a low power neural network accelerator, and a scalable PLL to fulfill real-time processing requirements. By automated integration of these specialized hardware components, we achieve a speedup of $\times$2.15 while reducing the power by 27% compared to the unoptimized solution.**

***Index Terms*—RISC-V, AI, IoT, edge applications, ecosystem**

I. Introduction

Internet of Things (IoT) applications became the reality over the last several years. In many application areas, like automotive, industry automation, and space, IoT implementations are most oftenly means of the solution. IoT devices rely frequently on edge processing on limited resources, which needs to perform efficiently with respect to energy and from the cost point of view. Moreover, with the outcome of network based connectivity and distributed IoT applications, the aspects of safety, security, and reliability became equally mandatory over the last years. For efficient application-specific processing IoT devices require the application of highly optimized processing architectures. Based on optimized utilizations they could provide up to 10 times better performance or power dissipation for demanding IoT applications.

In that context, RISC-V has been introduced some years ago as a novel, scalable, and extendible open Instruction Set Architecture (ISA). Meanwhile, we can find it as an adequate basis for advanced special purpose processors in a many research and commercial projects covering various applications. However, to enable efficient and effective use of RISC-V architectures for edge applications, the challenge is to provide a complete tool/IP ecosystem, which enables the implementation, verification, and test of highly scalable application-specific edge components. When considering complex electronic systems, the challenge does not only lie in the availability and management of different IP components but also in the availability of tightly integrated and scalable hardware and software tool chains with full automation support.

The Scale4Edge project was initiated to address those needs and introduced an integrated RISC-V ecosystem with hardware and software support for safe, secure, and reliable applications [1]. The project has currently finished the first project phase and continues with a second phase focusing on further automation and completion of the components. Fig.1 gives an overview of the different components and their interaction. The Scale4Edge ecosystem is based on a platform concept to supply efficient and cost-effective application-specific edge devices and to provide value-added services addressing different market segments. This is achieved through the automatic and very fine-grained adaptation of highly generic components to the application. As such, the Scale4Edge ecosystem covers highly scalable components and tools and extends them for application-specific edge components at three levels: (1) CPU instruction level defined by the RISC-V Instruction Set Architecture (ISA), (2) software level defined by the latest C programming language standard C11 with compilers and libraries open to complementary standards like MISRA-C, and (3) operating system and firmware level through system services, configuration interfaces, and drivers. The ecosystem platform is customizable to the individual application through the RISC-V ISA, which may define optional custom instructions, e.g., to support non-interruptible instructions for individual applications. Consequently, Scale4Edge is based on a broadly scalable hardware addressing different pipeline architectures, multi-core architectures, co-processors, and hardware accelerators, like for AI and DSP applications.

In this paper, we introduce one of the Scale4Edge demonstrators, which applied the Scale4Edge ecosystem to generate a specific processor for an industrial audio-event detection application provided by Bosch. Hereby, we demonstrate the efficiency of the fast and well-tailored components generation for our application. The application started from an CoreDSL

This work was supported in part by the German Federal Ministry of Education and Research (BMBF) within the project Scale4Edge under contract no. 16ME0122-140.

979-8-3503-4859-0/DATE24/© 2024 EDAA

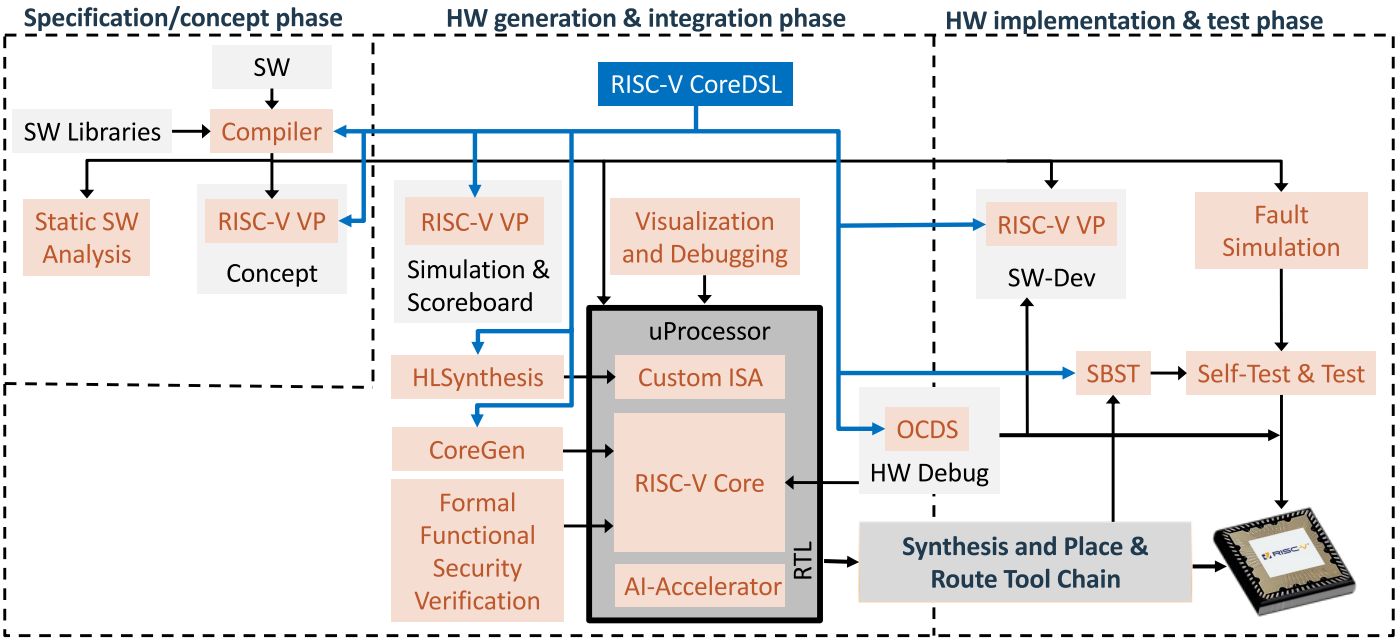


Fig. 1. Overview of the Scale4Edge ecosystem [1].

specification as an extension of the RISC-V standard ISA (ISAX) for fast AI and DSP processing. From CoreDSL we created a software compiler and also used it as an input for virtual prototyping, hardware synthesis, and physical implementation by commercial tools from Cadence and Siemens EDA. The final chip was fabricated by Globalfoundries' 22FDX technology, which was successfully evaluated by an audio-event detection software, which takes advantage of the extended ISA of the RISC-V processor. The software currently does not utilize the AI accelerator. While the chip is optimized for the audio-event detection, the RISC-V processor and the included ISAX as well as the AI accelerator can also be used for other applications.

The remainder of the paper is structured as follows. The next chapter introduces the different components of the chip architecture after which we give basic numbers of the chip layout and present the evaluation of the software application use case. The paper finally closes with a summary and an outlook.

II. Platform Design

To support a wide variety of AI and DSP applications, we created a scalable embedded platform that is designed by composing highly flexible and automatically generated components from the Scale4Edge ecosystem. The hardware generation allows for an easy and fast adaption of the platform to different use cases. It features a TGC RISC-V core [2] with four custom ISA extensions generated from CoreDSL, an instance of the AI hardware accelerator UltraTrail, and a programmable PLL. The underlying System-on-Chip (SoC) is a modified PULPissimo [3].

A. CoreDSL

For the audio software application, we investigated the required extra functionality and specified a set of three instructions specialized for ML applications and a zero-overhead loop feature to add to the base TGC ISA. These new instructions have been formulated in CoreDSL [4], a language to describe both base and extended ISAs, which serves as single-source-of-truth for downstream implementations. The CoreDSL language has a C-like syntax familiar to developers allowing a high-level specification for each instruction's functionality. For a multiply-accumulate ISA extension (ISAX), the CoreDSL description in Fig.2 can be used.

B. ISAX Integration and High Level Synthesis

For our demonstrator chip, ISA extensions for neural network inference as well as the addition of zero-overhead loops have proven especially beneficial greatly improving performance and energy efficiency with minimal area cost.

All the instructions were manually implemented as RTL and then automatically integrated into the base core using the SCAIE-V tooling. SCAIE-V is a highly portable open source ISAX hardware interface for extending RISC-V based processing cores with custom instructions [5]. In addition to enabling a new "R"-type ISAX, it also supports advanced features, such as custom control flow and memory accesses as well as decoupled execution. In parallel, the CoreDSL descriptions, which served as the specification for the RTL description, could already be used to create instruction-set simulators (ISSs), which can execute the new instructions and thus allow software development even before the actual hardware design was complete.

In ongoing work, we are focusing on the development of a tool flow that is able to accept the CoreDSL descriptions and can then automatically generate the underlying RTL hardware using specialized high-level synthesis (HLS) techniques. This ISAX HLS tool leverages the state-of-the-art compiler

```
MAC {
  encoding: 7'b0000000 :: rs2[4:0] :: rs1[4:0] :: 3'b010 :: rd[4:0] :: 7'b0101011;
  assembly: "{name(rd)}, {name(rs1)}, {name(rs2)}";
  behavior: {
    signed<65> result = (signed)X[rs1] * (signed)X[rs2] + (signed)X[rd];
    if(rd != 0) X[rd] = result[31:0];
  }
}
```

Fig. 2. CoreDSL example that defines a custom MAC instruction for a RISC-V based instruction set architecture.

frameworks MLIR [6] and CIRCT[1] for the hardware synthesis and SystemVerilog export and targets SCAIE-V for the ISAX interfaces. Thus, analogously to the manually created ISAX RTL described above, the HLS created ISAX hardware can then also be integrated into the base core using the SCAIE-V tooling.

In initial experiments, the quality-of-results of the ISAX HLS flow matched those of the carefully manually designed hardware, both in terms of performance as well as in area efficiency.

C. TGC RISC-V Cores

The good core (TGC) is a highly flexible, scalable, and configurable RISC-V based core family [2]. The five basic TGC configurations combine 3, 4, or 5 pipeline stages with different standard RISC-V instructions and additional features. Each of these cores can be optimized for area or performance and further customized. All standard cores can be easily tailored to specific application requirements using CoreDSL including the addition of custom ISA extensions through the SCAIE-V interface.

The TGC series comes with a broad software development kit allowing rapid prototyping and fast bring-up. Based on the CoreDSL descriptions of both the core and custom instructions, an ISS is generated. The ISS contains a static pipeline model and trace capabilities to allow hotspot analysis, e.g., using kcachegrind[2]. Together with a tailored LLVM Clang version and a Board Support Package (BSP), this enables reliable performance analysis early in the design phase.

Using the Scale4Edge ecosystem, the TGC cores are extensively verified and analyzed to ensure functional safety and security. Cross level verification against the CoreDSL specification using simulation and formal methods allows us to quickly achieve high coverage and automatically detect all functional bugs leading to a very high functional quality. The verification is complemented by standard approaches from the OpenHW group and RISC-V international. Functional safety and security are ensured by GapFree and by the UPEC (Unique Program Execution Checking) approach from Siemens EDA and Kaiserslautern University, respectively.

D. AI Accelerator UltraTrail

UltraTrail is a configurable hardware accelerator for real-time inference of temporal convolutional networks on edge devices [7]. Designed for near-sensor signal processing on energy constrained platforms it features an optimized dataflow for one-dimensional convolution with a total power consumption in the low microwatt range. The parameterizable architecture combined with a hardware-aware neural architecture search (NAS) allows an automatic generation of domain-specific accelerator instances [8]. Accurate models for power, performance, and area enable a fast design space exploration.

UltraTrail provides a TVM UMA backend to streamline the deployment of neural network models onto the hardware. The Universal Modular Accelerator Interface (UMA) offers an easy-to-use API to integrate new hardware accelerators into Apache TVM [9]. It was developed as part of the Scale4Edge project and has been integrated into TVM mainline since then.

E. Analog PLL

The analog PLL (Phase Locked Loop) is a frequency synthesizer and consists of a linear phase frequency detector (PFD), a current starved ring voltage-controlled oscillator (VCO), and a 7-bit programmable divider for frequency hopping, where the PFD uses positive edge-triggered True Single Phase Clock (TSPC) D flip-flops. To ensure the robust performance and stability of the PLL for Process Voltage Temperature (PVT) variations the PFD was designed from 10–200 MHz with a 96–89 % clock phase margin and a power consumption of 10 mW. The current starved ring VCO tuning range was designed from 180–1200 MHz and the programmable divider until 1.66 GHz. The PLL was optimized for a worst-case frequency range of 200–1000 MHz with 20 MHz frequency hopping with a rise time and fall time of 37–18 ps and a maximum locktime of 5 µs with a duty cycle variation of 49–51 %.

The analog PLL was integrated to ensure proper compatibility with the digital components. The power supply was decoupled with the digital section to prevent coupling noise between them with a solid and common grounding. To prevent interference and to maintain the integrity of the analog signal, the PLL was isolated from the fast switching digital components.

III. Chip Layout

To demonstrate the scalability of the presented hardware platform, an optimized implementation was taped out in 22FDX technology for an industrial audio-event detection software application provided by Bosch. It features a RV32IMC compliant 4-stage TGC-C core for IoT applications with four custom

[1]https://circt.llvm.org/
[2]https://github.com/KDE/kcachegrind

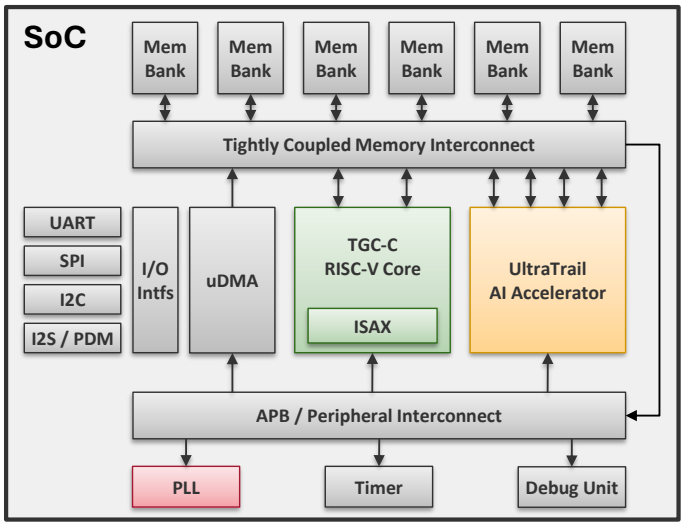


Fig. 3. Block diagram of the PULPissimo-based SoC platform [3].

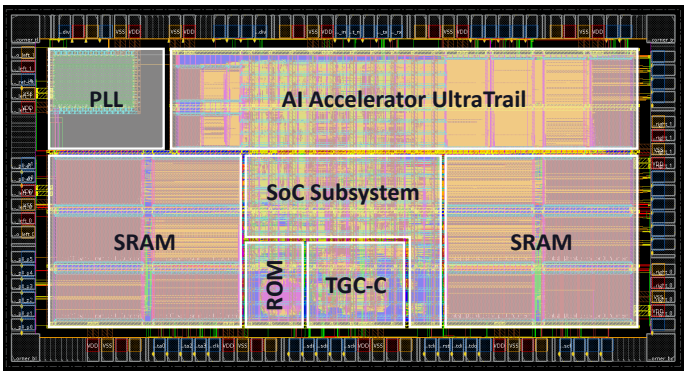


Fig. 4. Layout of the 22FDX demonstrator chip.

ISA extensions, an instance of UltraTrail with 144 multiply-accumulate units, and the configurable PLL to meet the real-time requirements. Fig. 3 shows the block diagram of the PULPissimo-based platform.

The layout of the design is shown in Fig. 4. It has a total area of 2500 µm × 1350 µm and contains 107 I/O-Pins, including supply voltages (1.8 V I/O, 0.9 V core). The PLL dimension was 300 µm × 200 µm. The clock source can be varied through an on-board oscillator (20–200 MHz) or through the on-chip PLL (160–1180 MHz). The digital design was optimized for a worst-case frequency range of 20–700 MHz to apply different clock frequencies, which can be dynamically scaled either via software or through configuration pins. Experimental results have shown, that we can safely overclock the chip up to 1 GHz. During idle times, the TGC-C and UltraTrail are aggressively clock-gated. In total, the system features 384 KiB of SRAM for the SoC and 136 KiB of SRAM for UltraTrail.

IV. Results and Evaluation

The demonstrator was evaluated with an application software for audio event detection provided by Bosch. The evaluation board is shown in Fig. 5(a). An external PDM microphone is connected to the I2S interface of the SoC. For the measurements, power values are reported for the chip without the PCB. Performance is measured using the SoCs performance counters.

Fig. 5(b) shows the execution time of the application at 200 MHz with and without utilizing the specialized ISA extensions. Due to the application latency constraint of 100 ms, it is not possible to run the application without the internal PLL. A comparison of performance, power, and area (PPA) between the ISAX and the non ISAX variant is shown in Fig. 5(c). The 2.15× increase in performance allows us to meet the real-time constraints with a significantly slower frequency resulting in an overall power reduction of 30 %. The area overhead of the additional ISA extensions is minimal with less than 1 % for the entire SoC and 14.8 % for the TGC-C alone.

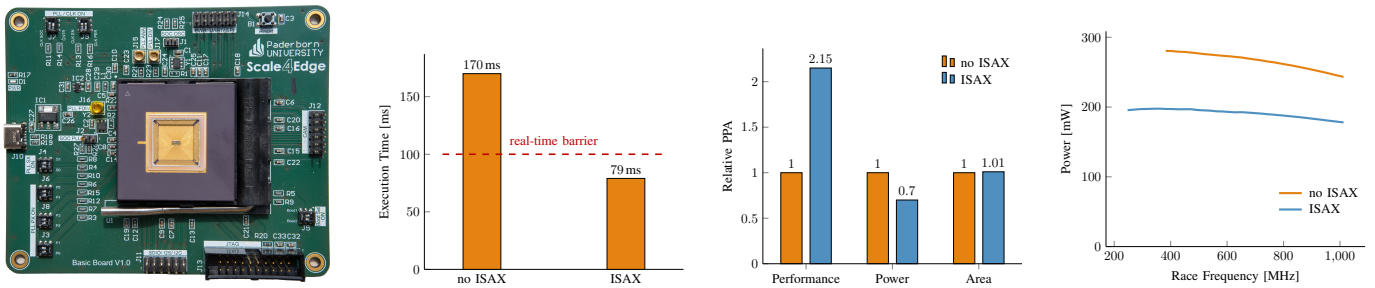
To further reduce power, we utilize dynamic frequency scaling to briefly run the application at high speed (race frequency) and switch to an idle state for the remaining time. Fig. 5(d) shows the total chip power for different race frequencies. For the application with ISAX, we achieve a 10 % power reduction from 198 mW at 247.5 MHz to 178 mW at 1012.5 MHz. For the non ISAX variant, dynamic frequency scaling is more effective with a 13 % power reduction resulting in an overall power reduction of 27 % between the two best-case scenarios.

V. Summary and Outlook

The paper presented results from the Scale4Edge project, which targets at the introduction of an ecosystem for RISC-V based edge applications composed of multiple components and evaluated by different demonstrators. It gave an overview of one of the project demonstrators and on the successful and powerful application of some of the ecosystem components. The here given audio-event use case executes software on a customized RISC-V processor architecture, which was manufactured as a chip in 22FDX technology. This demonstrated the applicability of the CoreDSL approach based on the efficient orchestration of several highly scalable ecosystem components for the hardware development in combination with the scalable software toolchain for software generation and debugging.

As such, the paper gave an overview of just one project demonstrator which applied parts of the Scale4Edge ecosystem. Other project demonstrators focus on power control and on high reliable (HiRel) electronic systems for space, high altitude avionics, nuclear etc. [10]. Other project activities focus on AI components and on alternative accelerators like the SpiNNedge accelerator [11]. A general overview of the complete Scale4Edge ecosystem and its demonstrators is given in [1] with continuous updates at [12].

The future work of the still ongoing Scale4Edge project will focus on optimizing, expanding, and further adapting the different Scale4Edge ecosystem components, particularly regarding optimized AI solutions. In addition, the project will investigate the opening up of further application domains while also heading for open access and open source solutions for some of the components.



(a) (b) (c) (d)

Fig. 5. Evaluation board with the chip in a CPGA-144 package. The I/O-Pins of the chip are routed to DIP switches or GPIO pins (a). Comparison of execution time (b), PPA (c), and total chip power utilizing dynamic frequency scaling (d) for the audio event detection application with and without ISA extensions.

REFERENCES

[1] W. Ecker, P. Adelt, W. Mueller, R. Heckmann, M. Krstic, V. Herdt, R. Drechsler, G. Angst, R. Wimmer, A. Mauderer, R. Stahl, K. Emrich, D. Mueller-Gritschneder, B. Becker, P. Scholl, E. Jentzsch, J. Schlamelcher, K. Grüttner, P. P. Bernardo, O. Bringmann, M. Damian, J. Oppermann, A. Koch, J. Bormann, J. Partzsch, C. Mayr, and W. Kunz, "The scale4edge risc-v ecosystem," in *2022 Design, Automation and Test in Europe Conference and Exhibition (DATE)*, 2022, pp. 808–813.

[2] MINRES, "The good folk series." [Online]. Available: https://www.minres.com/products/the-good-folk-series/

[3] P. D. Schiavone, D. Rossi, A. Pullini, A. Di Mauro, F. Conti, and L. Benini, "Quentin: an ultra-low-power pulpissimo soc in 22nm fdx," in *2018 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*, 2018, pp. 1–3.

[4] MINRES, "Coredsl 2.0." [Online]. Available: https://www.minres.com/work/coredsl/

[5] M. Damian, J. Oppermann, C. Spang, and A. Koch, "Scaie-v: An open-source scalable interface for isa extensions for risc-v processors," in *2022 The 59th Design Automation Conference (DAC)*, 2022. [Online]. Available: https://doi.org/10.1145/3489517.3530432

[6] C. Lattner, M. Amini, U. Bondhugula, A. Cohen, A. Davis, J. Pienaar, R. Riddle, T. Shpeisman, N. Vasilache, and O. Zinenko, "MLIR: Scaling compiler infrastructure for domain specific computation," in *2021 IEEE/ACM International Symposium on Code Generation and Optimization (CGO)*, 2021, pp. 2–14.

[7] P. P. Bernardo, C. Gerum, A. Frischknecht, K. Lübeck, and O. Bringmann, "Ultratrail: A configurable ultralow-power tc-resnet ai accelerator for efficient keyword spotting," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 39, no. 11, pp. 4240–4251, 2020.

[8] G. Christoph, F. Adrian, H. Tobias, P. P. Bernardo, K. Lübeck, and B. Oliver, "Hardware accelerator and neural network co-optimization for ultra-low-power audio processing devices," in *2022 25th Euromicro Conference on Digital System Design (DSD)*. IEEE, 2022, pp. 365–369.

[9] T. Chen, T. Moreau, Z. Jiang, L. Zheng, E. Yan, H. Shen, M. Cowan, L. Wang, Y. Hu, L. Ceze *et al.*, "{TVM}: An automated {End-to-End} optimizing compiler for deep learning," in *13th USENIX Symposium on Operating Systems Design and Implementation (OSDI 18)*, 2018, pp. 578–594.

[10] M. Ulbricht, J. Chen, L. Li, M. Krstic, and W. Müller, "Der tetrisc soc - ein resilientes quad-core system auf pulpissimo-basis," in *5. ITG / GMM / GI -Workshop Testmethoden und Zuverlässigkeit von Schaltungen und Systemen*, 2023.

[11] H. Bauer, M. Stolba, S. Scholze, D. Walter, C. Mayr, A. Oefelein, S. Höppner, A. Scharfe, F. Schraut, and H. Eisenreich, "A risc-v mcu with adaptive reverse body bias and ultra-low-power retention mode in 22 nm fd-soi," in *20th International SoC Conference (ISOCC 2023)*, 2023.

[12] "Scale4edge homepage." [Online]. Available: https://www.edacentrum.de/scale4edge