Success Story:
Energy-efficient Keyword-Spotting

Power-efficient edge AI applications have an enormous market potential. Key factors for a competitive hardware solution in this domain are energy efficiency and low silicon area in relation to computing performance. SpiNNcloud Systems GmbH is closely collaborating with TU Dresden in this domain for transferring cutting-edge research and development on AI hardware accelerators into commercially viable applications.

Within the Scale4Edge project, TU Dresden has developed the SpiNNedge AI accelerator IP for audio keyword spotting applications and beyond. Key focus of SpiNNedge is the reduction of on-chip memory and processing effort by hardware-optimized preprocessing and exploitation of sparsity in neural network based classification. With on-chip memory being one of the core drivers of silicon area and static power, SpiNNedge helps to significantly improve in both factors.

But an AI accelerator alone is worth almost nothing in isolation. Integration in the Scale4Edge ecosystem made SpiNNedge useable and created a commercial potential. The Scale4Edge RISC-V ecosystem core by MINRES runs the application with an offloading of key processing tasks for keyword spotting to the SpiNNedge accelerator. The customizable RISC-V core solution could be perfectly adapted to the needs of the keyword spotting application. Moreover, Scale4Edge made a joint effort to provide software support for hardware extensions of microcontrollers in the widely adopted machine learning compiler Apache TVM. This is highly beneficial for the useability of an accelerator like SpiNNedge, as it bridges the gap between hardware IP and high-level software frameworks for machine learning that users employ to develop their edge AI applications.

TU Dresden has successfully implemented a test chip for audio keyword spotting, integrating the SpiNNedge accelerator with a customized MINRES RISC-V core into an overall processing chain from microphone input to detected keywords. The chip realizes a hierarchical processing approach, first detecting speech in the microphone input, and then starting keyword classification. Speech detection is performed by the ZEN accelerator module, which won a 1st prize in the BMBF German innovation competition "energy efficient AI systems". The chip has been implemented in GlobalFoundries 22 FDX technology, employing adaptive body biasing (ABB) IP for leakage power reduction, developed and provided by Racyics, a spin-off of TU Dresden.

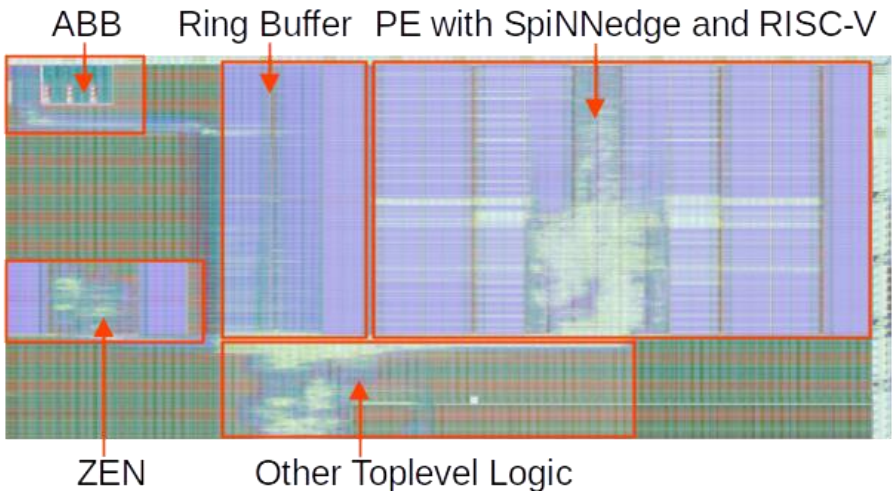

Figure 1: Core Layout of the keyword spotting test chip by TU Dresden. Core part is the processing element (PE) with TU Dresden's SpiNNedge accelerator and the MINRES RISC-V core.

This work has been developed in the ZuSE project Scale4Edge. Scale4Edge is funded by the German ministry of education and research (BMBF) (reference numbers: 16ME0122K-16ME0140+16ME0465). The authors are responsible for the content of this publication.

The processing element with the MINRES RISC-V core and the SpiNNedge accelerator performs real-time keyword spotting in 34 uW, achieving a classification accuracy on the widely-used Google speech command dataset of over 95%, which outperforms most of the existing low-power hardware solutions for this use case. A demo board has been implemented to showcase the solution.

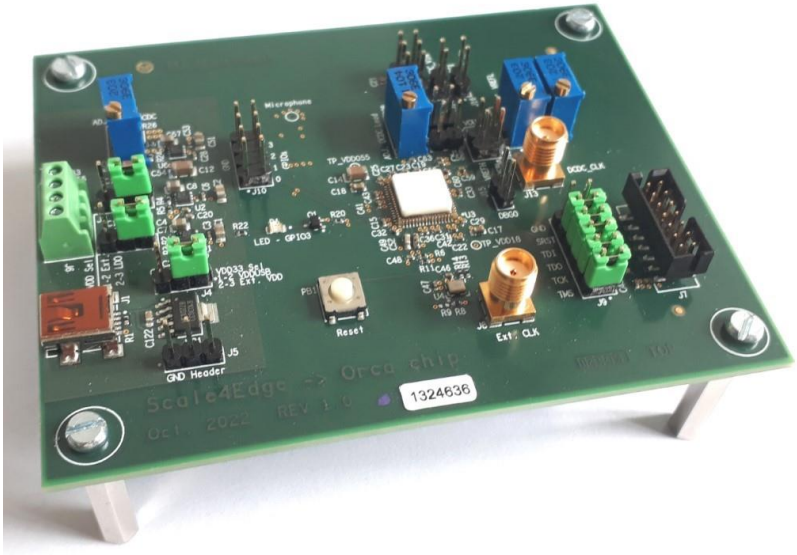

Figure 2: Demo board for keyword spotting with the test chip by TU Dresden (white, center)

The mission of SpiNNcloud Systems is to design energy-efficient AI systems by leveraging practical inspiration from the brain. Its core product is a holistic computing solution ranging from chip design, software development to deployment of full data center servers. Despite our large-scale supercomputers being built leveraging Arm IP, our constant commitment to pursue innovation in an Edge2cloud continuum has led us to evaluate different solutions at different scales. SpiNNcloud is highly interested in commercializing Edge AI solutions, in which the Edge IP built in Scale4Edge stands as a clear and attractive building block. The flexibility of these custom-extended RISC-V cores allow a path to achieve ultra efficient operations at the edge. Furthermore, the Scale4Edge Ecosystem provides a quick start for us into an efficient Edge exploration including know-how from compilers, functional safety, ISA-extensions and chip design methodology among others. Additionally, it introduces us to a wide variety of best practices that can boost development speed and reduce technical risks, all crucial to ensure commercial success. Strategically the consortium enables a fully European ecosystem, especially a German one, with technological independence that increasingly becomes more important to protect and foster companies in the field of computing.

Cont@ct:
TU Dresden | Dr.-Ing. Johannes Partzsch | johannes.partzsch@tu-dresden.de

MINRES | Eyck Jentzsch | eyck@minres.com

SpiNNcloud Systems GmbH | Matthias Lohrmann | matthias.lohrmann@spinncloud.com

https://www.edacentrum.de/scale4edge/

Further Scale4Edge partners and sub-contractors

This work has been developed in the ZuSE project Scale4Edge. Scale4Edge is funded by the German ministry of education and research (BMBF) (reference numbers: 16ME0122K-16ME0140+16ME0465). The authors are responsible for the content of this publication.