eki Research Project Overview

ENERGY EFFICIENT AI
BY DNN APPROXIMATION
FOR FPGAs

eki-project.tech

Felix Jentzsch (felix.jentzsch@upb.de)
Computer Engineering Group
Paderborn University, Germany

Project Data

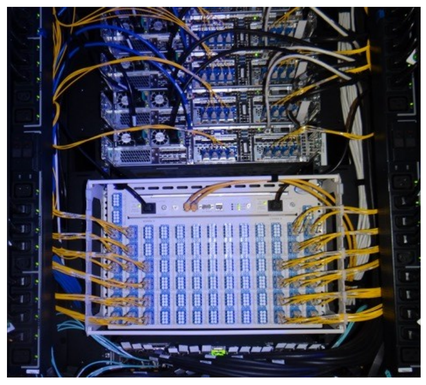
- Funded by German Federal Ministry for the Environment, Climate, Nature Conservation, Nuclear Safety and Consumer Protection
- Funding line “AI Lighthouses – Resource-efficient AI”
- Runtime 01/2023 to 12/2025
- **Goal:** Increase the energy efficiency of AI systems for deep neural network (DNN) inference through approximation techniques and mapping to high-end FPGA systems in the data center
- **Partners:**

Paderborn Center for Parallel Computing

High-performance FPGA Cluster

- 1124 servers x 2 AMD Milan 64 core CPUs
- 48 Xilinx Alveo U280, each with two optical ports
- Calient S320 all-optical switch: 320 ports, fully non-blocking, 100 GBps, low ns latency

Noctua 2 cluster at PC²

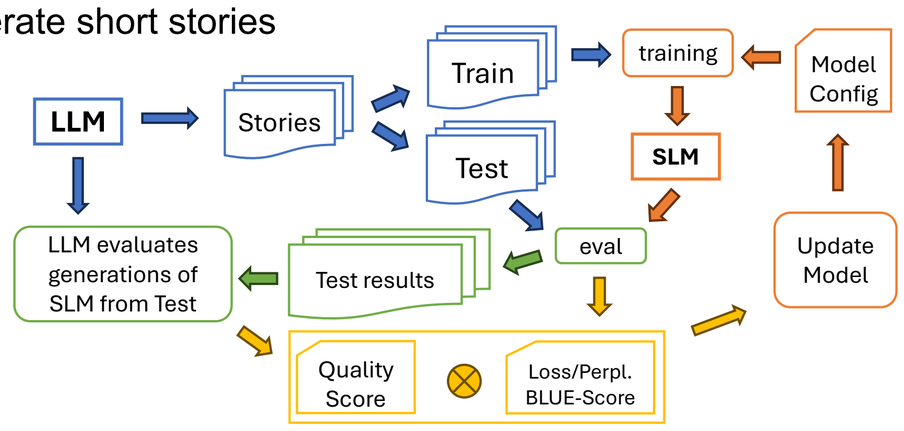


All-optical switch

Example Use Cases

- **RadioML**: Tiny DNNs for extreme throughput processing of radio signals (modulation classification, fingerprinting, etc.)
- **Precision farming:** CNN-based drone image processing for detection & classification of sugar beet plant health

- **Natural language processing:** Training tiny Transformers to generate short stories



Main Research Topics

- Focus on DNN Approximation for FPGAs via quantization & pruning
- Specifically explore streaming dataflow architectures, using the FINN framework as a vehicle

- **FINN integration into the Noctua 2 cluster**
  → *Highlighted in poster from Linus Jungemann (PC²)*
- **Multi-FPGA acceleration**
  → *Highlighted in poster from Bjarne Wintermann (PC²)*
- **FINN support for the transformer model architecture**
  → *Highlighted in poster from Christoph Berganski (CEG)*
- **Energy characterization & estimation**
  → *Highlighted in poster from Felix Jentzsch (CEG)*

- **Hardware-aware AutoML for energy optimization**
  - Extend search space to include accelerator-specific settings for quantization, resources, parallelism, etc.
  - Integrate search algorithms with the FINN compiler to create an end-to-end tool stack for DNN deployment on datacenter FPGAs
- **Empirical evaluation**
  - Employ public models (e.g., ResNet-50) and custom case studies developed within eki

Past & Present Student Projects using FINN

- Bachelor theses
  - **Effizienzanalyse leichtgewichtiger Neuronaler Netze für FPGA-basierte Modulationsklassifikation**, Florian Simon-Mertens
  - **Development of a Power Analysis Framework for Embedded FPGA Accelerators**, Lucas Reuter
  - **Demonstrator for Dataflow-based DNN Acceleration for Vision Applications on Platform FPGAs**, Marvin Oviasogie
  - **Efficient Automatic Speech Recognition on FPGAs for Datacenters**, Tobias Erhart
- Master theses
  - **Design and Implementation of a RadioML Demonstrator based on an RFSoC Platform**, Salem AlAidroos
  - **Exploring Custom FPGA Accelerators for DNN-based RF Fingerprinting**, Luca-Sebastian Henke
- Student project group (8 students for 1 year)
  - **WiFi-based Human Sensing using FPGA-accelerated Lightweight Neural Networks**

Supported by:

Federal Ministry
for the Environment, Nature Conservation,
Nuclear Safety and Consumer Protection

based on a decision of
the German Bundestag