Contribution ID: **103** Type: **Standard Talk**

FINN+: Towards Hassle-Free Co-Design of FPGA DNN Inference Accelerators

Tuesday, 2 September 2025 13:00 (20 minutes)

Custom FPGA dataflow accelerators for DNN inference can enable unprecedented performance and efficiency for many applications. Dataflow accelerator compilers, such as the FINN framework, have improved in recent years and allow practitioners to explore this technology without requiring in-depth FPGA knowledge.

However, the overall design process remains quite tedious, time-consuming, and often requires significant manual intervention. This is primarily caused by limited flexibility and automation in the compiler, as well as the enormous size and complexity of the design space. In contrast to the typical exploration process, where a quantized DNN is manually trained and then passed through the compiler, requiring many iterations to reach an acceptable solution, we envision an automated co-design of DNN and FPGA accelerator based on Automated Machine Learning (AutoML) techniques.

In an effort to realize this vision while also facilitating the exploration of FPGA dataflow accelerators for energy efficient inference in the datacenter, we introduce FINN+, our custom fork of the FINN framework. Our work so far includes empirical resource and power consumption modeling, support for Transformer topologies, efficient deployment on datacenter FPGAs, Multi-FPGA acceleration, and general usability improvements.
In this talk, we will share recent highlights as well as remaining challenges of the project.

Author: JENTZSCH, Felix

Co-authors: WINTERMANN, Bjarne (Paderborn University); BERGANSKI, Christoph (Paderborn University); JUNGEMANN, Linus (Paderborn University)

Presenter: JENTZSCH, Felix

Session Classification: Contributed talks