ID de Contribution: **1** Type: **Non spécifié**

# Neural Network Accelerator Co-Design with FINN

*lundi 4 juillet 2022 15:00 (30 minutes)*

High-throughput and low-latency edge applications need co-designed solutions to meet the performance requirements. Quantized Neural Networks (QNNs) combined with custom FPGA dataflow implementations offer a good balance of performance and flexibility, but building such implementations by hand is difficult and time-consuming. In this presentation, we will introduce FINN, a framework for building fast and flexible FPGA accelerators using a flexible heterogeneous streaming architecture. It is an open-source experimental framework by Xilinx Research Labs to help the broader community to explore deep neural network (DNN) inference on FPGAs. It specifically targets QNNs, with emphasis on generating dataflow-style architectures customized for each network. It is not intended to be a generic DNN accelerator like xDNN, but rather a tool for exploring the design space of DNN inference accelerators on FPGAs. The key components are Brevitas for training quantized neural networks, the FINN compiler, and the finn-hlslib Vivado HLS library of FPGA components for QNNs.

**Orateur:** Dr PREUSSER, Thomas (AMD)