Student assignment

Type: Graduation assignment
Contact: Barry de Bruin (e.d.bruin@tue.nl)

Topic: Energy-efficient multi-granular arithmetic on the Blocks CGRA for Binary Neural Networks inference.

Background:
Binary Neural Networks (BNN) have become a potent candidate for energy-efficient computer vision on battery-constrained edge devices. These networks generally utilize wide variety of reduced-precision number formats to maintain acceptable quality-of-service. On top of that, the memory access pattern and computational order (i.e. schedule) becomes quite complex. Specialized hardware support for multi-granular arithmetic and complex schedules is required to run these networks efficiently.

Task description:
In this MSc thesis project you are tasked to investigate, implement and evaluate potential hardware optimizations to support energy-efficient multi-granular arithmetic on the state-of-the-art Blocks Coarse-Grained Reconfigurable Array (CGRA). This includes instruction set extensions and modifications to the CGRA interconnect to efficiently support these operations. These optimizations will be evaluated on a large-scale BNN application.

The expected outcomes of this MSc thesis are:

1. An extensive literature study on existing approaches for energy-efficient multi-granular arithmetic support on programmable or reconfigurable architectures.
2. An evaluation of the most promising approaches using an analytical model or synthesized RTL-level implementation in a modern ASIC technology.
3. An application-level evaluation of the most promising approach implemented on the Blocks CGRA while (partially) running a large BNN.
4. A detailed report and/or potential paper.

COVID disclaimer:
This project can be fully carried out remotely. The second half of the project requires a stable connection to a TU/e server (over the VPN) to access the ASIC tooling for simulation and/or logic synthesis.
Recommended courses and experience

- Intelligent architectures (IA).
- Embedded Computer Architecture (ECA).
- Some experience with the Linux command line and RTL coding and some scripting language like Python is a plus.

Interesting references

- “Multi-granular arithmetic in a coarse-grain reconfigurable architecture” – S. Louwers et al. (DSD 2016)
- “Binary Neural Networks: A Survey” – H. Qin et al. (ArXiv 2020)
- “Review and Benchmarking of Precision-Scalable Multiply-Accumulate Unit Architectures for Embedded Neural-Network Processing” – V. Camus et al. (JETCAS 2019)