Reconfigurable Hardware Accelerator for Machine Learning

Description

With the end of Dennard Scaling and the imminent end of Moore's Law, the search for new ways to improve performance in computing systems is increasing. Nowadays, the main approach is to use hardware accelerators to offload the application. They achieve higher performance at low energy efficiency. One such example is Systolic Arrays (SA), used for accelerating general matrix-matrix multiplication (GEMM), and kernel key in domains such as Machine Learning (ML). SAs are a simple and flexible architecture that can leverage the regular pattern of GEMM. However, while SAs define how computations are performed, there are still other aspects to analyze. These aspects include how to feed them the required data and how to control them. Moreover, with the goal of accelerating ML, it may be interesting to attach more functional support to the SA. In all, there are multiple considerations when it comes to designing hardware accelerators

Goals

With this project we would like to learn:

- Which tools are there available for the development of hardware accelerators?
- Which memory systems can better benefit SAs?
- How can we interface a SA-based accelerator so that our software can make use of it?
- What else can be done to accelerate ML applications?
- What other challenges can appear while developing hardware accelerators?

Pre-requisites

For the success in this project the student needs to have basic knowledge of computer architecture. Basic knowledge of Machine-Learning is an added benefit.

Methodology

- Analysis and selection of tools for development of hardware accelerators.
- Clear identification of the problem and formulation of the research question (the research question should come from the questions stated above in the description)

- Implementation of a SA running on a FPGA.
- Analysis of the implemented SA.
- Development of a software stack that can make use of the SA.
- Analysis of the developed platform and comparison with baseline.

Target group

D and E

Selective Relevant References

- Kung, Hsiang-Tsung. "Why systolic architectures?." Computer 15.01 (1982): 37-46.
- Ananda Samajdar, Yuhao Zhu, Paul Whatmough, Matthew Mattina, Tushar Krishna, "SCALE-Sim: Systolic CNN Accelerator Simulator", October 2018, arXiv:1811.02883 (https://github.com/scalesim-project/scale-sim-v2)
- Genc, Hasan, et al. "Gemmini: Enabling systematic deep-learning architecture evaluation via full-stack integration." 2021 58th ACM/IEEE Design Automation Conference (DAC). IEEE, 2021. (https://github.com/ucb-bar/gemmini)

Contact

Mateo Vázquez Maceiras
Computer Science and Engineering
maceiras@chalmers.se