CustoNN2: Customizing Neural Networks on FPGAs 2

High-Performance IT Systems group

Dr. Tobias Kenter
Prof. Dr. Christian Plessl

Kickoff Meeting
11 October 2018

UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

Paderborn
Center for
Parallel
Computing

Recap – CustoNN2 Goals

- Hot topic neural networks
- 32 brand new and huge Stratix 10 FPGAs at PC²
- Programming via high-level synthesis tool flow OpenCL
- Many others research in this area!
- We need…
    - OpenCL code specifying the CNN execution on FPGA
    - Functional execution of established CNNs on FPGA hardware
    - Models to understand the performance of solutions
    - (Re) training of adapted CNNs on GPU/CPU
- **We want to achieve cool new results with this setup**
    - Scaling over multiple FPGAs via host or point-to-point
    - Codesigned/ Specialized topologies or applications
    - Fixed precision/binary CNNs or sparse weights
- Goals will be refined for project plan

Evaluation Criteria

- **Overall group result**
  - ➢ Final project report
  - ➢ Source code, documentation
  - ➢ Performance figures, scaling
  - ➢ Trained/adapted CNNs

- Your individual contributions
  - – Code you contribute
  - – Documents, presentations, decisions that you bring forward
  - – Participation in tutorial, solutions to exercises
- ➢ **Individual interviews with each participant**
  - Briefly every ~3 months
  - At project end

Organization

- It's your project
  - Among the learning goals:
    - Self organization
    - Collaboration
    - Project organization

- It's our joint project
  - Platforms, tools and design methods are central to our research
  - Topic suitable for publications and follow-up projects

Thus, we start giving directions...

… but we expect you to take over step by step

Time is everything

- PG in CS: 2 x 10 ECTS
- PG in CE: 2 x 9 ECTS

~ 2 x 1.7 "(3+2)" lectures + self-study
~ 2 x 1.5 "(3+2)" lectures + self-study

- 1 ECTS = 30 hours time effort, e.g. 2 hours/week during term
- comparison: new CS master lecture (e.g. 3+2) = 6 ECTS

➢ **>= 2 full work days for PG**

- **Common time slots Tuesday 9am-12 and Wednesday 13-16**
  - Tutorials, group meetings, discussions, joint hands-on sessions
  - Flexible allocation?
  - Additional individual+group work in lab and from home

Skills and know-how you will need

- OpenCL - concepts, host and kernel code
- FPGAs - architecture, resources, general design flow
- Programming FPGAs with OpenCL - concepts and tools
- Performance modeling
- CNNs - general architecture, compute patterns for different layers
- CNNs - benchmarks, existing fixed-point / binary NNs
- CNN training with frameworks

Tutorial phase – mix of prepared material and self-study

- CNN research
  - multi-FPGA scaling, custom data formats, sparsity, …

Research for project plan

Infrastructure

- Gitlab
  - central location for all code, scripts, makefiles, measurements, documentation
  - experienced gitmaster?
- Shared file system
  - training data (ImageNet, Cifar, …), OpenCL binaries
- Mailing List
- Slack?
- Lab:
  - Access tokens
  - Your laptops + some monitors + Icy box USB
    - Keep the infrastructure usable, plug monitors, keybords back in…
  - Remote access to synthesis resources and FPGA hardware
  - Custom computing infrastructure + Noctua cluster

Getting Started

Self-evaluation and preparation

- https://app.codility.com/programmers/lessons/2-arrays/
- Two tasks
  - different programming languages, please try several, including C
  - first task states efficiency is not relevant
    - please do consider it
  - second task emphasizes efficiency
    - it depends on a little trick, but solutions with O(N^2) are still ok
  - save a copy of your evaluation results and bring to the next meeting

First tutorial session

- Tuesday 16 October 9:15-11:45 am **O4.267**