Modulbezeichnung: **Project on Accelerating Machine Learning and HPC on Cloud FPGAs (cFProj)** **10 ECTS**
(Project on Accelerating Machine Learning and HPC on Cloud FPGAs (cFProj))

Modulverantwortliche/r: Dietmar Fey
Lehrende: Burkhard Ringlein

Startsemester: SS 2022	Dauer: 1 Semester	Turnus: jährlich (SS)
Präsenzzeit: 100 Std.	Eigenstudium: 200 Std.	Sprache: Englisch

Lehrveranstaltungen:

How can performance demanding applications such as machine learning or scientific simulations be computed faster and more efficiently using modern accelerators in a cloud ecosystem? This project (Praktikum) is organized in cooperation with IBM Research Europe - Zurich and will use its research prototype cloudFPGA (https://www.zurich.ibm.com/cci/cloudFPGA/).

Project on Accelerating Machine Learning and HPC on Cloud FPGAs (cFProj) (SS 2022, Praktikum, 8 SWS, Anwesenheitspflicht, Burkhard Ringlein)

Empfohlene Voraussetzungen:

- Good programming skills in C++
- Basic knowledge of Python and Linux
- Basic knowledge of FPGAs/VHDL would be beneficial, but is not required
- Prior knowledge about the named example application is not required

Inhalt:

Data-center workloads are expanding exponentially and cloud services are becoming increasingly globalized from the core cloud to the cloud edge. At the same time, the hardware powering today's clouds offerings has evolved, too. Due to the limited performance and efficiency gains provided by new generations of CPUs, new compute architectures with innovations beyond technology scaling were needed to sustain the performance and efficiency gains demanded from every new generation. This has led to the emergence of accelerators ranging from GPUs and Field Programmable Gate Arrays (FPGAs) all the way to domain specific Application Specific Integrated Circuits (ASICs) for Artificial Intelligence (AI) acceleration. Consequently, in reaction to this demand, accelerators like FPGAs are offered as Infrastructure-as-a-Service (IaaS) offerings, too. Accelerator-based clouds have illustrated improvements in power and performance by accelerating compute-intensive workloads. A well-designed accelerator covers the broadest possible space of application, accelerating a domain rather than a single application. Adding domain-specific functions to a programmable accelerator, like FPGAs, provides the efficiency of the specialized function while retaining deployment flexibility and operationalization agility. This project will analyze the current trends in accelerated machine learning and accelerated HPC from an industry and research perspective as well as examine possible future developments.
Topics include:

- Examples of machine learning inference on FPGAs
- Distributed machine learning applications on FPGAs
- Examples of HPC kernels on FPGAs
- Resource abstraction of Cloud FPGAs
- Deployment and controlling of FPGAs in the Cloud
- Evolution of domain specific languages for FPGA usage

Afterwards, the students implement one example application in small groups on the IBM cloudFPGA platform (https://github.com/cloudFPGA). Example applications could be:

- Low-latency image processing/filtering
- 2D or 3D Jacobi simulation (HPC kernel)

- Monte Carlo simulations
- Deep Neuronal Networks on one or multiple FPGAs

Lernziele und Kompetenzen:

- Awareness of the current situation of technology scaling, its history and possible future development
- Understanding of latest acceleration technology and its application to machine learning and HPC
- Recognizing the advantages and disadvantages of resource provisioning via cloud infrastructure
- Using modern FPGA tools and generating FPGA bitstreams
- Programming/adaption of one application based on existing libraries
- Accessing cloud infrastructure
- Remote debugging of FPGAs, network traffic analysis
- Analyzing and comparing scientific research
- Ability to summarize the state of the art for one particular topic
- Ability to summarize the work done in this project
- Working independently on complex scientific topics
- Development and execution of a oral presentation of the results
- Development and execution of a written report

Literatur:

Will be announced at the beginning of the course.

Studien-/Prüfungsleistungen:

Project on Accelerating Machine Learning an HPC on Cloud FPGAs (cFProj) (Prüfungsnummer: 540766)

(englische Bezeichnung: Project on Accelerating Machine Learning an HPC on Cloud FPGAs (cFProj))

Praktikumsleistung
weitere Erläuterungen:
Graded oral presentation (50%), graded report (50%)
Prüfungssprache: Englisch

Erstablegung: SS 2022, 1. Wdh.: WS 2022/2023
1. Prüfer: Dietmar Fey

Organisatorisches:

This project (Praktikum) is divided into two parts: In the first part, the state of the art is examined theoretically and presented by the students. Afterwards, in the second part, selected examples will be implemented in small groups on the IBM cloudFPGA platform (https://github.com/cloudFPGA). After an introductory lecture at the beginning of the term, the students work independently on the theoretical background. The students will be assigned topics to prepare a presentation (20min), supported from the lecturers if needed. The talks will be given on one or two days at the end of the term. The second part will be done in one week full-time at the end of the term. In this week, the learned background will be put into practice by implementing selected examples of machine learning or HPC acceleration in small teams. Afterwards, the teams will submit a written report (15 pages).