Master’s Thesis Evaluating the Impact of Integrating GPU Acceleration in Serverless Computing

Context
Serverless computing offers a range of advantages over traditional cloud computing models. For instance, relieving developers from operational tasks such as resource management and planning while enabling new properties such as high elasticity, pay-per-use pricing, and fully managed application runtimes. In particular, the emerging model of Function-as-a-Service (FaaS), where developers write small functional blocks of code that run in response to events, is rising in popularity.

State of the Art & Problem
Serverless computing, and FaaS in particular, were initially intended for stateless microservice applications. However, industry and research have expanded the use of these platforms for big-data processing and machine learning. Though, some use-cases, such as machine learning and inference, are restricted by the current FaaS setups. Specificity, state of the art FaaS platforms offer no hardware acceleration such as GPUs for FaaS runtimes.

Thesis Topic & Goal
In this thesis, we want to investigate if FaaS platforms can use hardware acceleration by integrating GPU access in an OpenSource FaaS Platform. We are particularly interested in how GPU acceleration can be integrated into a volatile FaaS environment and how a FaaS provider is impacted by adding this feature.

Skills: Linux, Virtualization/Container, GPU Programming, Serverless Computing, Benchmarking

Contact: Sebastian Werner und Anselm Busse
{sw, ab}@ise.tu-berlin.de

Our Mission:
Our lectures cover fundamental methods and techniques in the areas of service computing, cloud computing, and enterprise computing. We like to engage students in hands-on building of distributed information systems and to take an interdisciplinary approach to evaluating such systems. Through a close mentoring of students, especially in our seminars, we aim to introduce students to our ongoing research and to excite them to do future studies and research with us.