ACCELERATING SERVERLESS FUNCTIONS WITH GPU/TPU POWER

Rajitha Pasindu Warusavitarana

A dissertation submitted in partial fulfilment of the requirements for Bachelor of Engineering (Honours) degree in Software Engineering.

Department of Computing

Informatics Institute of Technology, Sri Lanka

in collaboration with

University of Westminster, UK

2021

Abstract

Serverless computing is rapidly growing technology in the present day. The technology adaption statistics reveals that many developers tend to follow the serverless architecture model to deploy software applications in cloud environments. PaaS (platform as a service), CaaS (container as a service), and FaaS (function as a service) are some serverless models which are widely used in the present day. The FaaS serverless model, which is also commonly known as serverless functions is the most abstracted version of cloud computing.

Based on an analysis of recent research, several problems of the FaaS serverless model were identified. Lack of GPU/TPU support in the FaaS model is a widely mentioned issue in both academic and grey literature. This is a major disadvantage for scientific computing, image processing, video analysis, and data analytics related tasks. Through an in-depth investigation of existing systems, several limitations related to GPU/TPU acceleration in the FaaS model were identified. Inability to access TPU/GPU resources directly, network usage issues with GPU resources, inability to share a GPU resource with multiple serverless functions, and performance degradation are some of them. Furthermore, it was identified that TPU acceleration in serverless FaaS model is a minimally researched area.

This research introduces a novel approach to accelerate serverless functions with GPU/TPU power. It addresses a computer architecture challenge which falls under hardware heterogeneity of the serverless FaaS model. The developed FaaS platform allows to share a single GPU resource with multiple serverless functions, and to mitigate the issues such as extensive use of the network and performance degradation. Both quantitative and qualitative evaluations were conducted to assess the developed FaaS platform. The test results verified that followed approach provided better results comparing with the micro-benchmarks of similar systems. Moreover, the research contribution was validated during the qualitative evaluation conducted with domain experts.

Key Words: Serverless, GPU Acceleration, TPU Acceleration, FaaS, Serverless Functions, Function as a Service.