FastWarm

A Hybrid Approach to Mitigate the Cold-starts in FaaS Platforms

Hirun Kodituwakku

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

School of Computing

Informatics Institute of Technology, Sri Lanka
in collaboration with
University of Westminster, UK

2023

ABSTRACT

FaaS is an emerging cloud computing paradigm that enables consumers to deploy and execute tasks as individual functions. The service provider manages the management of functions such as resource allocation and scaling, making the deployment of functions simpler and quicker than integrating a dedicated server. In addition, cloud service providers offer resource scaling policies, such as the scale-to-zero policy, in which the client is only billed for actual usage and not for periods of non-use. Consequently, this paradigm has become widespread in cloud application development.

The cold start problem is a common issue in FaaS platforms, which occurs when the datacentre lacks the necessary number of function instances to manage the invocation load. Despite the fact that numerous techniques have been proposed to mitigate this issue, cloud service providers typically employ a small number of simple techniques to reduce the cold start. Among the numerous techniques proposed to address this issue, use of model-based approaches to mitigate this has the disadvantage of requiring training data that cannot be obtained for some time after function is deployed. The author proposes a hybrid scaling mechanism in which a novel algorithm based on historical data is used until sufficient data is collected, after which the aforementioned model-based approach is implemented.

Due to the limitations of prominent FaaS platforms, the hybrid scaling mechanism was implemented using the open-source FaaS platform Knative. Author tested and benchmarked it against the platform default KPA (Knative Pod Autoscaler) and was able to reduce cold starts during the data collection period (algorithmic scaling) by more than 60% and during the model applied period (model-based scaling) by more than 85%.

Key Words – Serverless Cold-start, Cold Start in Function as a Service, Cloud Computing, Serverless Computing, Time Series Analysis