DALHOUSIE
UNIVERSITY

CSCI 5410
Serverless Data Processing

Part 1
Research Paper Summary

Ritva Katrodiya
B00930131

Mitigating Cold Start Problem in Serverless Computing: A Reinforcement Learning Approach [1]

The author of this research paper addresses the initial Startup delay problem associated with serverless computing, specifically in the context of IoT applications [1]. They highlight the benefits of serverless computing, such as decreased maintenance expenses and system development complexity, as well as its potential for IoT devices that are limited in resources. However, the cold start delay continues to be a significant performance challenge. Existing approaches to reducing delay have limitations, such as resource waste and fixed mechanisms that do not adapt to the dynamic cloud environment [1]. To address that problem, the author proposes a novel two-layer responsive technique. The first layer applies a comprehensive learning reinforcement method to learn function invocation patterns over time and determine the best time to keep containers warm. The second layer applies a long short-term memory (LSTM) model to predict future function invocation times and calculate the number of prewarmed containers required. The proposed method mitigates the usage of memory while improving the execution operations on prewarmed containers [1].

The cold start delay in serverless computing is one of the concerns discussed in the paper. When a function is called, a new container is established, the container environment is configured, memory and computing resources are assigned, and the function is loaded into it and executed [1]. The cold start is prompted by these initial steps [1]. While the scale-to-zero feature of serverless computing reduces costs by redirecting idle functions, it also contributes to the cold start challenge, according to the paper. Each new invocation requires repeating the initialization steps, and if numerous invocations occur at the same time, various instances of the function must be generated and executed individually. This causes longer cold start delays, particularly as the number of invocations increases. The study overcomes this issue by demonstrating a smart strategy for determining the most suitable strategy for keeping containers warm over time based on function invocations, decreasing cold start times, and improving the utilization of resources in serverless environments [1].

By minimizing cold start occurrences and maximizing resource usage, the two-layer solution that has been presented seeks to solve the cold start issue [1]. The first layer implements an adaptive reinforcement learning technique to identify the best idle-container window while taking the utilization of resources into account. The second layer utilizes an LSTM model to estimate future postponed requests and enable container prewarming, which leads to shorter cold start times. By considering the problem's frequency and delay characteristics, this method offers flexibility and efficiency [1].

The author examined their method on a pair of data sets: one with sequential and one with simultaneous invocations. The analysis focused on important parameters such as idle-container window, number of cold starts, and memory consumption [1]. In comparison to the Openwhisk platform's fixed window, the first layer's dynamic assessment of the idle-container window resulted to improved efficiency [1]. The successful completion of invocations on prewarmed containers enhanced by 22.65% because of the second layer's intention to reduce cold start latency situations through container prewarming. These results demonstrate how successfully the authors' methodology tackled cold start challenges related to serverless computing [1].

References:

[1] “EzProxy - Libraries,” Dalhousie University. [Online]. Available:https://ieeexplore-ieee-org.ezproxy.library.dal.ca/stamp/stamp.jsp?tp=&arnumber=9749611&tag=1. [Accessed:06 June 2023].