Subject

CSCI 5410 (Serverless Data Processing)

Professor

Dr. Saurabh Dey

Assignment 1

Part A

Purvisha Patel

(B00912611)

Synopsis on the "Mitigating Cold Start Problem in Serverless Computing: A Reinforcement Learning Approach"

A paper on **"Mitigating Cold Start Problem in Serverless Computing: A Reinforcement Learning Approach**" by **Parichehr Vahidinia, Bahar Farahani** and **Fereidoon Shams Aliee** explores methods for reducing the negative effects of delays that cause cold and suggests a two-layer adaptive strategy to address the problem. In order to find the function invocation patterns over time and determine the ideal time to keep the containers warm, the first layers uses a comprehensive reinforcement learning approach. The other layer predicts function invocation timings in the future and calculates the required number of pre-warmed containers using a Long Short-Term Memory (LSTM) [1].

Researchers have proposed an efficient method that chooses the optimum strategy for keeping the containers warm in accordance with the function calls over time to reduce cold start latency with consideration of resource consumption. To reduce the cold start delay, two broad solutions can be considered. First, by lowering the cold start, and second, by lowering the occurrence.

Firstly, the technique for decreasing cold start delay uses pre-warmed stem cell containers on the Openwhisk platform, which classifies the containers based on memory and programming language. As a result, the preparation time for the container and cold start delay is reduced. By using Akkus's [2] application-level separation approach, all of an application's functions are housed on the same container to decrease the delay of internal events as well as the delay of functions calling one another.

Secondly, the Cold Start Occurrence Reduction layer's main goal is to steadily reduce the number of cold starts. It uses the Temporal Difference (TD) Advantage Actor-Critical method because the state and action spaces are continuous. As a result, the intelligent agent analyses the function's invocation pattern as well as the gaps between invocations in the past to determine the value of the idle-container window in the future.

Cold start delay is one of today's biggest challenges, thus minimizing it in real-time systems is essential because it directly impacts both performance and customer happiness. The most popular solutions are ineffective, so a method is needed that figures out how long it takes to keep the containers warm by figuring out how often the functions are called. The study's researchers have proposed a two-layer approach to achieving the objectives. Discovering the function invocation pattern and determining the idle-container window are part of the first layer. Based on the LSTM's predicted time of the upcoming invocation, the second layer calculates the required number of pre-warmed containers.

References

[1] P. Vahidinia, B. Farahani, and F. S. Aliee, “Mitigating cold start problem in serverless computing: A reinforcement learning approach,” IEEE Internet Things J., pp. 1–1, 2022.

[2] I. E. Akkus, R. Chen, I. Rimac, M. Stein, K. Satzke, A. Beck, P. Aditya, and V. Hilt, “{SAND}: Towards high-performance serverless comput ing,” in 2018 {Usenix} Annual Technical Conference ({USENIX}{ATC}
18), 2018, pp. 923–935