UNIVERSITI TEKNOLOGI MARA

VERIFICATION OF ENERGY-DRIVEN MICROSERVICE AUTOSCALING POLICIES IN CLOUD ENVIRONMENT USING PROBABILISTIC MODEL CHECKING

SITI NURAISHAH BINTI AGOS JAWADDI

Thesis submitted in fulfilment
of the requirements for the degree of
Master of Science
(Computer Science)

College of Computing, Informatics and Mathematics

August 2023

ABSTRACT

Microservice autoscaling is one of the critical mechanisms in a cloud system that needs to be analyzed and verified to ensure that it is working as expected. This is a challenging task for the cloud engineer at design time especially if they need to understand the impact on the energy consumption since the actual request demands are unknown which causes non-deterministic scaling action. Probabilistic model checking (PMC), a branch of formal verification has been widely recognized as a suitable technique to analyze stochastic systems exhaustively. Since microservice autoscaling can be characterized as a stochastic system, this technique is naturally suitable to be applied for verification and analysis purposes. However, the application of this technique is not a straightforward implementation since the autoscaling behavior needs to be formally specified, and the objective to be measured needs to be formally quantified. Therefore, this research addresses this challenge by proposing a formalism of microservice autoscaling decision process to enable verification and analysis of the decision in relation to energy efficiency level. Four formal models based on the Markov decision process (MDP) have been developed by embedding distinct scaling constraints. The models are then encoded, verified, and analyzed using a PMC model checker, known as PRISM. The analyses conducted focus on the efficiency measures by comparing the four models that consider different sets of scaling constraints to drive the autoscaling decision-making process. The measures determine how far the models can minimize the host energy consumption and how frequently the decisions made by the models cause energy violations. The inputs of each model are based on the variation of incoming workloads at a normal hour and peak hour. The analysis results prove that it is insufficient for the autoscaling decision-making process to only consider energy constraints in the baseline process to minimize the host energy consumption and frequency of energy violations. Meanwhile, taking into account the latency tactic along with the energy constraints encourage the frequent selection of scaling action. This eventually speeds up the scaling process and improves the energy efficiency level of the autoscaling decision-making process regardless of the number of demands.

ACKNOWLEDGEMENT

Praise be to Allah for bestowing good health, opportunities, and many other blessings that enabled me to successfully complete my research and master's journey. I would like to express my immense gratitude to my dedicated and helpful supervisor; Assoc. Prof. Dr. Azlan bin Ismail and Dr. Mohd Suffian bin Sulaiman for their guidance and encouragement throughout my learning process. My special thanks go to Assoc. Prof. Dr. Valeria Cardellini for her contribution in writing one of the research manuscripts with me and Assoc. Prof. Dr. Azlan Ismail to fulfill one of the requirements to complete this research. Finally, my special appreciations go to my family and friends for their unconditional love, support, and prayers that have eased my life journey this while.

TABLE OF CONTENTS

CHAPTER ONE
INTRODUCTION

This chapter emphasizes the overview of research entitled Verification of Energy-Driven Microservice Autoscaling Policies in Cloud Environment using Probabilistic Model Checking which discussed the background of the study, problem statements, research objectives, research questions, the scope of research, and significance of the research.

1.1 Background of Study

Microservice is an architecture that consists of loosely-coupled and fine-grained structures as well as considered the most promising architecture to develop modern and large-scale cloud software systems (Jamshidi et al., 2018). Apart from microservice being an extension of service-oriented architecture (SOA), it is more advance than monolithic architecture in terms of agility, reliability, scalability, and domain-specific development (Mazzara et al., 2020). Moreover, it is often implemented as a container-based application (Salah et al., 2017). In the report presented by Gartner Inc in April 2021, the worldwide end-user spending on public cloud services is forecasted to grow 23.1% in 2021 with a total estimation of USD 332.32 billion compared to USD 270 billion in 2020 whilst containerization is one of the emerging technologies that become the common user preference when spending on public cloud services (Costello & Rimol, 2021). Meanwhile, Docker and Kubernetes are the common open-source container management systems whilst according to the State of Cloud Native Development Report, Kubernetes has been adopted by 5.6 million developers to develop and manage container-based applications including microservices (Witkowski & Korakitis, 2021).

Despite the advantages, microservice applications had increased the number of workloads exchanged in clouds due to the independent execution process performed by the services. Thus, to maintain or improve the scalability of a microservice-based application, a resource management method such as autoscaling is performed by using the container management system. Autoscaling has been introduced by the cloud service providers (CSPs) to enable the client (i.e. application owners) to dynamically