WORKLOAD, RESOURCE, AND PRICE AWARE PROACTIVE AUTO-SCALAR FOR DYNAMICALLY-PRICED VIRTUAL MACHINES

Daham Positha Pathiraja

179340D

M.Sc. in Computer Science

Department of Computer Science and Engineering

University of Moratuwa

Sri Lanka

March 2019

WORKLOAD, RESOURCE, AND PRICE AWARE PROACTIVE AUTO-SCALAR FOR DYNAMICALLY-PRICED VIRTUAL MACHINES

Daham Positha Pathiraja

179340D

This report is submitted in partial fulfillment of the requirements for the Degree of
Master of Science in Computer Science specializing in Cloud Computing

M.Sc. in Computer Science

Department of Computer Science and Engineering

University of Moratuwa

Sri Lanka

March 2019

Declaration

I, D.P. Pathiraja, hereby declare that this is my own work and this report does not incorporate without acknowledgement any material previously submitted for the degree or diploma in any other university or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant University of Moratuwa the non-exclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles of books).

Signature: ……………………. Date: ……………………….

Name: D. P. Pathiraja

I certify that the above candidate has carried out research for the Masters thesis under my supervision.

Signature: ………………. … Date: ………………………

Name of the supervisor: Dr. H. M. N. Dilum Bandara

Abstract

Proactive Cloud auto-scalers forecast future conditions and initiate scaling response in advance leading to better service quality and cost savings. Their effectiveness depends on the forecast accuracy and penalty due to miss prediction. However, such solutions assume fixed prices for virtualized Cloud resources to be provisioned. Hence, they are unable to benefit from dynamically-priced resources such as Amazon Spot Instances which are introduced by Cloud providers to deal with fluctuating workloads cost effectively. Moreover, users have the risk of losing resources when the dynamically-adjusted market price of resources exceeds the user-defined maximum bid price. Therefore, proactive auto-scalers should also forecast market price of dynamically-priced resources to minimize the cost further while retraining service quality. However, predicting the market price (to set the maximum bid price) is quite complicated given highly varying workload and resource demands. We present a proactive auto-scalar for dynamically-priced virtual machines by combing the workload and resource prediction capabilities of an existing auto-scalar named InteliScaler, and a novel technique for forecasting Spot price. We retrieve Spot price history from Amazon and use it to forecast the future prices using Recurrent Neural Networks. Next, we selected the maximum price for a given decision window as the bid value to make Spot request. To demonstrate the utility of the proposed solution, we tested the performance of the enhanced auto-scaler using a synthetic workload generated using the Rain toolkit and the RUBiS auction site prototype. Proposed auto-scaler with dynamically-priced virtual machines reduced the total cost by ~75% compared the same auto-scalar with fixed priced instances. Moreover, no noticeable change in service quality was observed.

Acknowledgement

I owe my deepest gratitude to my supervisor, Dr. Dilum Bandara of Department of Computer Science, Faculty of Engineering, University of Moratuwa, for his invaluable support in providing relevant knowledge, advice and supervision throughout the project. His continuous guidance, inspiring advices and constructive feedback provided this project a great value and an encouraging background throughout the project. This would not have been possible without his expertise and remarkable support.

I take this opportunity to convey our gratitude towards my batch mates, family members and friends who helped me to do my best towards achieving success during the entire project time period.

Finally, I would like to thank my colleagues at Sysco Labs (Pvt.) Ltd for covering my work and helping me to balance the workload. Without them, this project would not have been possible.

Last but not least, I am grateful for all the people who supported me throughout this research in various means.

Table of Contents

List of Figures

List of Tables

Abbreviations

APE	Absolute Percentage Error
ANN	Artificial Neural Network
ARIMA	Autoregressive Integrated Moving Average
ARMA	Auto Regressive
AWS	Amazon Web Services
CMDP	Constrained Markov Decision Process
EC2	Elastic Compute Cloud
IaaS	Infrastructure as a Service
IOPS	IO Operations per Second
LA	Load Average
MAPE	Mean Absolute Percentage Error
PaaS	Platform as a Service
POC	Proof of Concepts
PTPM	Price Transition Probability Matrix
QoS	Quality of Service
RMSE	Root Mean Square Error
RUBiS	Rice University Bidding System
VM	Virtual Machine