CLUSTER AUTOSCALER FOR UNMANAGED KUBERNETES CLUSTER DEPLOYMENT ON CLOUD

S. A. Chathura Madhushanka Siriwardhana

199366F

Degree of Master of Science

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

August 2020

CLUSTER AUTOSCALER FOR UNMANAGED KUBERNETES CLUSTER DEPLOYMENT ON CLOUD

S. A. Chathura Madhushanka Siriwardhana

199366F

Project report submitted in partial fulfillment of the requirements for the degree Master of Science in Computer Science Specialising in Cloud Computing

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

August 2020

DECLARATION

I declare that this is my own work and this project report does not incorporate without acknowledgment any material previously submitted for a 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 acknowledgment is made in the text. Also, I hereby grant to the University of Moratuwa the non-exclusive right to reproduce and distribute my thesis, 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 or books).

Signature: Date:

The above candidate has carried out research for the Masters Project report under my supervision.

Name of the supervisor: Prof. Gihan Dias

Signature of the supervisor: Date:

ABSTRACT

This project report comprises details of the research "Cluster Autoscaler for Unmanaged Kubernetes Cluster Deployment on Cloud". Underutilization of server resources is a huge issue in enterprise data centers. When it comes to Kubernetes, underutilization and overutilization issue exists as is. In Kubernetes main course of having this issue is the use of fixed number of Kubernetes worker nodes. The Kubernetes community provides a cluster autoscaler solution to reduce underutilization and overutilization on Kubernetes clusters. This solution is only supported by a few major cloud providers like Google, AWS, DigitalOcean, and few others. Also, this solution is tightly bound to the auto scale group concept in those clouds. Hence this solution provided by the Kubernetes community cannot be used elsewhere. Therefore, there is a necessity for a general auto scaling approach that can be used on a wide range of cloud platforms and hardware virtualization platform. This research is to design and develop a Kubernetes Cluster Autoscaler which can be used on any cloud platform. This is achieved by removing the tightly bound auto scale group in the solution proposed by this research. Proposed solution use API and SDK provided by cloud provider and using libvirt which is a general purpose API library to manage KVM, Xen, VMWare ESXi and QEMU.

ACKNOWLEDGMENT

Firstly, I would like to express my sincere gratitude to Professor Gihan Dias at the University of Moratuwa for his valuable support and contribution in my research and the dedicated guidance provided to me throughout the research in terms of motivation, and immense knowledge.

My sincere thanks goes to the Cloud Native Computing Foundation, Google Inc. and Docker for developing Kubernetes and container standards, and introducing it to the opensource community.

Last but not the least, I would like to thank my family and fellow colleagues for the precious support given to me throughout this period in order to successfully complete the research work.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

LXC	Linux containers
OCI	Open Container Initiative
CNCF	Cloud Native Computing Foundation
HPA	Horizontal Pod Autoscaler
CI/CD	Continuous Integration / Continuous Delivery
GCE	Google Compute Engine
GKE	Google Container Engine
ASG	AutoScaleGroups
AWS	Amazon Web Services
EKS	Elastic Kubernetes Service
API	Application Program Interface
AKS	Azure Kubernetes Service