School of Computing
FACULTY OF ENGINEERING AND
PHYSICAL SCIENCES

UNIVERSITY OF LEEDS

Final Report

Optimizing Cloud Resource Allocation through a Proactive Combination of Horizontal Pod Autoscaling and Cluster Autoscaling in Kubernetes

Farha Rashid Sayed

Submitted in accordance with the requirements for the degree of
BSc Computer Science and Mathematics

2023/24

COMP3931 Individual Project

Summary

In today's fast-paced resource-demanding world, businesses face challenges in efficiently managing their resources to meet fluctuating application demands. Recent marketing strategies often focus on sale events, holidays or big product launches which drive sudden intense spikes in traffic. E-commerce platforms need to be duly prepared for such events that hold the potential to either make or break customer experiences, brand image and thus revenue potential. Ensuring seamless performance and cost-effectiveness becomes paramount.

This project aims to address the common issue of the lack of scalability of systems by exploring how integrating different autoscaling strategies can optimize resource allocation while simultaneously ensuring preparedness for sudden, high bursts of traffic. Inspired by a project called SmartScale, we want to contribute insight **into** the effectiveness of a combination of three specific autoscaling techniques, the Horizontal Pod Autoscaler (HPA), the Cluster Autoscaler (CA) and node overprovisioning. We consider factors such as costs, resource optimization and the rapid convergence of applications to desired scaling levels, even in the face of significant changes in workload intensity.

We want to conduct comprehensive performance testing to assess the efficiency, responsiveness, and overall viability through performance metrics such as reconfiguration time and resource utilization. Additionally, we also wanted to conduct a comparative analysis between the three autoscaling implementations to determine the most optimal solution, appreciating the differences and making recommendations for further enhancements.

This project conducted comprehensive load testing experiments, monitored metrics, and visualized data to compare strategies such as HPA, CA and Node Overprovisioning. Through these experiments, this project found the benefits of overprovisioning pods and allocating spare nodes for systems facing small, quick bursts of load. We found that overprovisioning pods in a system set up with an inclination towards cluster autoscaling proves to be the fastest solution with moderately high resource demands when it comes to autoscaling. Additionally, we also found that a combination of HPA and CA is the most ideal solution when a reasonably consistent amount of load is provided over a constant period of time.

Acknowledgements

I would like to thank my supervisor, Karim Djemame, for his ever-generous patience, continuous support, and guidance during each step of this project. I am extremely grateful for a supervisor that went above and beyond to ensure that I performed to the best of my ability.

I would also like to extend my thanks to my assessor, Sebastian Ordyniak, for his constructive feedback that helped me build additional strategies for improving this project.

Most importantly, I am profoundly grateful to my parents, sisters, and friends for their steadfast support throughout the duration of this project. Their unwavering encouragement has been invaluable and is deeply appreciated. In particular, I would like to extend special thanks to my father for inspiring me with his dedication and passion for this field.

Table of Contents