School of Computing
FACULTY OF ENGINEERING AND
PHYSICAL SCIENCES

UNIVERSITY OF LEEDS

Final Report

Optimizing Cloud Resource Allocation through Combined Proactive 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

The candidate confirms that the following have been submitted*:*

Items	**Format**	**Recipient(s) and Date**
Final Report	*PDF file*	*Uploaded to Minerva (08/05/24)*
Link to online code repository	https://github.com/farha-sayed/FYP-K8s	*Linked in Final Report 08/05/24)*

The candidate confirms that the work submitted is their own and the appropriate credit has been given where reference has been made to the work of others.

I understand that failure to attribute material which is obtained from another source may be considered as plagiarism.

Farha Rashid Sayed

Summary

In today's fast-paced high-compute 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 experience, brand image and thus revenue potential. Ensuring seamless performance and cost-effectiveness becomes paramount. This project aims to address this common dilemma by exploring how integrating different autoscaling strategies can optimize resource allocation while simultaneously ensuring preparedness for 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 grateful for a supervisor that went above and beyond to ensure that I performed to the best of my ability.

Additionally, I would like to thank my parents, sisters, and friends for their unwavering support throughout the duration of this project. I would especially like to thank my father for influencing me with his dedication and passion for the field.

Table of Contents

List of References

1. Alsalem, L. A. A., 2021. *Auto Scaling in the Cloud: An Evaluation of Kubernetes,* s.l.: s.n.

2. Anon., n.d. *TeaStore on GitHub.* [Online]
   Available at: https://github.com/DescartesResearch/TeaStore

3. AWS, n.d. *What is a Private Cloud?.* [Online]
   Available at: https://aws.amazon.com/what-is/private-cloud/
   [Accessed 11 2023].

4. Chronosphere, 2023. *What is Prometheus and how does it help cloud native monitoring?.* [Online]
   Available at: https://chronosphere.io/learn/what-is-prometheus-and-how-does-it-help-cloud-native-monitoring/
   [Accessed 17 03 2024].

5. Codoid, 2019. *Performane Testing Terminologies.* [Online]
   Available at: https://codoid.com/performance-testing/performance-testing-terminologies/#:~:text=Saturation%20is%20a%20point%20during,state%20of%20optimized%20resource%20utilization.
   [Accessed 01 05 2024].

6. DescartesResearch, 2022. *TeaStore on GitHub.* [Online]
   Available at: https://github.com/DescartesResearch/TeaStore
   [Accessed 06 12 2024].

7. Docker, 2024. *containerd vs. Docker: Understanding Their Relationship and How They Work Together.* [Online]
   Available at: https://www.docker.com/blog/containerd-vs-docker/#:~:text=What's%20containerd%3F,%2C%20networking%20capabilities%2C%20and%20more.
   [Accessed 10 04 2024].

8. Google Cloud, n.d. *Containers.* [Online]
   Available at: https://cloud.google.com/learn/what-are-containers

9. IBM, n.d. *Virtual Machines.* [Online]
   Available at: https://www.ibm.com/topics/virtual-machines
   [Accessed 2024].

10. Kubecost, n.d. *Kubernetes HPA.* [Online]
    Available at: https://www.kubecost.com/kubernetes-autoscaling/kubernetes-hpa/
    [Accessed 01 03 2024].

11. Kubernetes, 2023. *Kubernetes Documentation / Overview.* [Online]
Available at: https://kubernetes.io/docs/concepts/overview/
[Accessed 17 03 2024].

12. Kubernetes, 2024. *ReplicaSet.* [Online]
Available at: https://kubernetes.io/docs/concepts/workloads/controllers/replicaset/
[Accessed 12 04 2024].

13. Li Ju, P. S. S. T., n.d. *Proactive Autoscaling for Edge Computing Systems with Kubernetes,* s.l.: s.n.

14. Medium, 2019. *Understanding Kubernetes Cluster Autoscaler.* [Online]
Available at: https://medium.com/kubecost/understanding-kubernetes-cluster-autoscaling-675099a1db92
[Accessed 20 04 2024].

15. Medium, 2023. *Performance Testing with JMeter.* [Online]
Available at: https://medium.com/@monusinghpersonal/performance-testing-with-jmeter-133de25fdd0a
[Accessed 12 2023].

16. Microsoft Azure, 2023. *Best practices for performance and scaling for small to medium workloads in Azure Kubernetes Service (AKS).* [Online]
Available at: https://learn.microsoft.com/en-us/azure/aks/best-practices-performance-scale
[Accessed 16 01 2024].

17. Microsoft Azure, n.d. *Cloud Computing.* [Online]
Available at: https://azure.microsoft.com/en-gb/resources/cloud-computing-dictionary/what-is-cloud-computing
[Accessed 04 2024].

18. Microsoft Azure, n.d. *What is a public cloud?.* [Online]
Available at: https://azure.microsoft.com/en-gb/resources/cloud-computing-dictionary/what-is-a-public-cloud#:~:text=The%20public%20cloud%20is%20defined,storage%2C%20or%20bandwidth%20they%20consume.
[Accessed 11 2023].

19. Mulugeta Ayalew Tamiru, Johan Tordsson, Erik Elmroth, Guillaume Pierre., 2020. *An Experimental Evaluation of the Kubernetes Cluster Autoscaler in the Cloud,* s.l.: s.n.

20. Nhat-Minh Dang-Quang, M. Y., n.d. *Deep Learning-Based Autoscaling Using Bidirectional Long Short-Term Memory for Kubernetes,* s.l.: s.n.

21. Red Hat, 2017. *What is a Kubernetes pod?.* [Online]
Available at: https://www.redhat.com/en/topics/containers/what-is-kubernetes-pod
[Accessed 10 04 2024].

22. Research Gate, 2018. *TeaStore Architecture.* [Online]
Available at: https://www.researchgate.net/figure/TeaStore-Architecture_fig1_328638010
[Accessed 20 03 2024].

23. Soham Kakade, G. A. O. P. A. D. N. G. S. p. P. B. S., n.d. *Proactive Horizontal Pod Autoscaling in Kubernetes using Bi-LSTM,* s.l.: s.n.

24. Sourav Dutta, Sankalp Gera, Akshat Verma, Balaji Viswanathan, 2012. *SmartScale: Automatic Application Scaling in Enterprise Clouds,* India: IBM Research.

25. VMWare, n.d. [Online]
Available at: https://www.vmware.com/topics/glossary/content/virtual-machine.html#:~:text=A%20Virtual%20Machine%20(VM)%20is,a%20physical%20%E2%80%9Chost%E2%80%9D%20machine.
[Accessed 26 April 2024].

26. VMware, n.d. *What is a Kubernetes cluster?.* [Online]
Available at: https://www.vmware.com/content/vmware/vmware-published-sites/us/topics/glossary/content/kubernetes-cluster.html.html
[Accessed 09 04 2024].

27. ZDNet, 2022. *What is Cloud Computing?.* [Online]
Available at: https://www.zdnet.com/article/what-is-cloud-computing-everything-you-need-to-know-about-the-cloud/
[Accessed 11 2023].