ABSTRACT

Kubernetes is an open-source tool used to manage containerized workloads and services, for both configuration and automation. Containers are similar to VMs (Virtual Machine) where they are stored in their own filesystem, CPU, memory, process space and more. A problem occurs when the webserver breaks while being accessed for data and services where said services are hosted by servers. A university website being viewed by an influx of users applying or submitting a form at the same time frame is an example of this; if the server cannot accommodate a significant number of users, the service may degrade leading to a bottleneck and users will suffer some latency when accessing the website.

This thesis proposes a performance analysis by simulating traffic in a High Scalability Kubernetes-based cluster. This scenario is used to check whether a Kubernetes-based cluster would be an efficient tool for networking due to its more lightweight management. The data is then analyzed to know whether it follows the standard of a good QoS and to deduct its cost efficiency in real world applications.

The results for Cluster Autoscaling, Vertical Pod Autoscaling can be analyzed from node provisioning and by memory and CPU use events. QoS in the cluster is calculated by looking into the system metrics and calculating the difference of data sent and data received based on the number of virtual users within a given time frame. This data is analyzed in making sure that the number of HTTP requests results in HTTP Status Code 200 and that 95% of the requests are below 2 seconds.

Keywords: *Kubernetes, High Availability Clusters, Cost-Efficiency, Autoscaling, Stability*