ABSTRACT

In this digital era, video conferencing is common in society. This technology is very helpful for humans in carrying out their daily activities. Because so many users access, it is not uncommon for this service to experience downtime. One of the causes of downtime is that the resources on the server have been used up.

The solution to overcome downtime is to use an infrastructure built with a container orchestration tool called Kubernetes. The Kubernetes Cluster will run on top of Linode which functions as a Cloud Service. Kubernetes runs services on nodes. There is the smallest component in the node, namely the pod.

This study aims to test the service using Kubernetes which is in charge of managing the clusters on the server. *Micro* Kubernetes *cluster* is the latest development from Kubernetes which is smaller and faster in managing clusters. With the *micro* Kubernetes *cluster*, a Video Conference service will be built.

The service implemented is WebRTC. This service will be tested and compared with several parameters. Parameters that are considered are CPU Usage, response code, response time, and throughput. Based on these parameters, the results obtained are that service run on a monolithic architecture has a better response time with a value of 2668 ms, a response code of 106, and CPU Usage in the range of 35 - 68%. Service running on a *micro* Kubernetes *cluster* architecture with the HPA (Horizontal Pod Autoscaler) feature has more stable CPU Usage due to load balancing that occurs by replicating pods by HPA.

Keywords: Kubernetes, High Availability, Cluster, WebRTC.