ABSTRAK

Kubernetes adalah alat sumber terbuka yang digunakan untuk mengelola beban kerja dan layanan dalam *container*, baik untuk konfigurasi maupun otomatisasi. *Container* mirip dengan VM (Mesin Virtual) di mana mereka disimpan dalam sistem file, CPU, memori, ruang proses, dan lainnya sendiri. Masalah terjadi saat *server web* rusak saat diakses untuk data dan layanan di mana layanan tersebut dihosting oleh *server*. Sebuah situs web universitas yang dilihat oleh banyak pengguna yang mengajukan atau mengirimkan formulir pada waktu yang sama adalah contohnya; jika server tidak dapat mengakomodasi sejumlah besar pengguna, layanan dapat menurun yang menyebabkan kemacetan dan pengguna akan mengalami latensi saat mengakses situs web.

Tesis ini mengusulkan analisis kinerja dengan mensimulasikan lalu lintas dalam *cluster* berbasis Kubernetes Skalabilitas Tinggi. Skenario ini digunakan untuk memeriksa apakah klaster berbasis *Kubernetes* akan menjadi alat yang efisien untuk jaringan karena pengelolaannya yang lebih ringan. Data tersebut kemudian dianalisis untuk mengetahui apakah sesuai dengan standar *QoS* yang baik dan untuk mengurangi efisiensi biaya dalam aplikasi dunia nyata.

Hasil untuk *Cluster Autoscaling*, *Vertical Pod Autoscaling* dapat dianalisis dari penyediaan node dan berdasarkan peristiwa penggunaan memori dan CPU. *QoS* dalam cluster dihitung dengan melihat metrik sistem dan menghitung selisih data yang dikirim dan data yang diterima berdasarkan jumlah pengguna virtual dalam jangka waktu tertentu. Data ini dianalisis untuk memastikan bahwa jumlah permintaan HTTP menghasilkan Kode Status HTTP 200 dan bahwa 95% permintaan di bawah 2 detik.

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