ABSTRAK

Penggunaan internet dalam kehidupan sehari-hari semakin meningkat. Layanan aplikasi yang digunakan juga semakin beragam dan jumlah pengguna tidak dapat diprediksi. Hal ini menyebabkan penyedia aplikasi kesulitan dalam melakukan manajemen aplikasi khususnya dalam aspek skalabilitas. Aspek skalabilitas aplikasi dapat diimplementasikan dengan manajemen *container* aplikasi yang handal seperti Kubernetes. Kubernetes menjadi pilihan yang umum untuk mengatur dan mengelola *container* aplikasi. Skalabilitas layanan *web* dalam Google Cloud Platform dapat menggunakan Google Kubernetes Engine. Pada penelitian ini dilakukan pengujian aspek skalabilitas terhadap *cluster* Kubernetes yang dibangun di atas Google Kubernetes Engine. Tujuan dari penelitian ini ialah melihat aspek skalabilitas dari *cluster* Kubernetes yang menggunakan *horizontal pod autoscaler* dan *cluster* tanpa *horizontal pod autoscaler*. Konfigurasi *horizontal pod autoscaler* yang digunakan pada *cluster* Kubernetes adalah utilisasi CPU 50 persen. Pengujian terhadap *cluster* dilakukan sebanyak tiga kali setiap variasi *user*. Variasi *user* dimulai dari 100 *user* hingga 1.000 *user* dengan kelipatan 100 *user*. Dari hasil pengujian terhadap *cluster* Kubernetes diperoleh hasil bahwa *cluster* dengan HPA memiliki aspek skalabilitas lebih baik daripada *cluster* non-HPA. Aspek skalabilitas yang dipertimbangkan yaitu jumlah *pod*, jumlah transaksi, *transaction rate*, *response time*, *longest transaction* dan *shortest transaction*. *Cluster* HPA memiliki jumlah transaksi dengan *transaction rate* lima kali lebih banyak, *response time* dua kali lebih cepat, waktu *longest transaction* dan *shortest transaction* sedikit lebih cepat daripada *cluster* non-HPA. Untuk pengembangan penelitian ini dapat dilakukan metode *autoscaling* berbeda, alat *load testing* bukan *siege* dan menggunakan parameter yang berbeda.

Kata Kunci : *container*, *horizontal pod autoscaler*, *cluster*, CPU, HPA.