Abstract

In recent years due to sharp increase in demand for cloud computing services, we need effective platforms for management of services. As a large number of users are to access these services, we need to manage and handle a large volume of requests. In this project, using Kubernetes as a powerful platform in cloud computing services, we intend to create an ecosystem for automated management of microservices so that in the peak times of system workload, the serving nodes or Microservices to be automatically scaled up. By scaling up the serving nodes or microservices, we are able to load balance workload among serving nodes or microservices, preventing a huge workload on every serving node or microservice. As a result of auto-scaling, the availability of microservices rises and the response time of them reduces. Therefore, more requests can be responded by these microservices. This ecosystem includes the base Kubernetes platform, api-server, metric server, load monitoring objects such as Linkered and Prometheus, load generation objects such as siege and microservices that are built based on Web microservices.

Key Words: microservices, availability, response time, auto-scaling, Kubernetes, cloud computing services