Abstract

Recently cloud computing has received significant attention as a promising approach for delivering ICT (information and communication technologies) services that rent computing resources on-demand, bill on a pay-as-you-use basis, and can multiplex many users on the same physical infrastructure. These cloud computing environments provide an illusion of infinite computing resources to cloud users so that the users vary the resource consumption rate according to their demands.

At the same time, the cloud environment has many challenges. Two actors of the cloud computing environment are cloud providers and cloud users, with different goals; providers wish to meet the service level agreement (SLA) and to maximize revenue with utmost resource utilization, while the cloud users want to meet their performance requirements with minimum expenditure. But, it is difficult to allocate resources in a mutually optimal way due to the lack of information sharing between the providers and the users. The other difficulties in resource scheduling are ever-increasing heterogeneity and variability of the environment which poses more challenges for the cloud providers. In this thesis, we develop an analytical model and measure the performance of the Virtual Machines (VMs) in different cloud computing environments. The various performance measures may help the cloud providers to tune the various performance parameters of the system to achieve high resource utilization and also users meet their application performance.

In a cloud system, each cloud computing user (CCU) requests cloud computing service provider (CCSP) for use of resources. A cloud user may need to wait in the system to get service from a provider. If CCU finds the server busy, then the user has to wait till its turn. This may result in increase of queue length as well as waiting time, which may lead to request drop. To handle this problem, CCSP needs to find ways to reduce waiting time. Multiple VMs may be engaged to serve the user's request. We introduced a finite multi-server queueing model with queue dependent heterogeneous servers

where the web applications are modeled as queues and the virtual machines are modeled as service providers. Cloud computing service providers may use multiple servers and the number of busy servers may vary depending on the queue length to reduce the queue length and waiting time. This helps the system to dynamically create and remove VMs in order to scaling up and down.

A private cloud is that in which an internal datacenter is assembled within an organization and the infrastructure is devoted to a particular organization and not dealt with other organizations. We developed two service policies along with an analytical resource prediction model for private cloud system. Various performance measures of the cloud system for finite population environment indicate that the proposed provisioning techniques may help the cloud operators in tuning the resources accordingly to improve the quality of service (QoS) targets.

We developed a finite-buffer multi-server queueing system where client requests have two arrival modes corresponding to premier users and the normal users. For such a system, we obtain various system measures such as average number of requests in the system, average number of requests in the buffer, the blocking probability and average waiting time in the system. A cost model is developed to search the optimal values of various parameters for the system using genetic algorithm.

As most of the applications deployed in the cloud adopt n tier architecture there is a need to provide QoS performance guarantee for each class of services in each tier. We derived an efficient provisioning technique for serving the client requests efficiently at each tier. We also presented an analytical model for bulk services in cloud system.