Chapter 8

Conclusions

The primary aim of our work was to analyze the performance of virtual machines for efficient resource scheduling. In this chapter, we summarize the work carried out in the thesis and briefly outline the possible directions for future research work.

We model the cloud systems by considering virtual machines as service centers and the web applications as queues. We employed finite buffer multi-server queueing system with queue dependent heterogeneous virtual machines. The service load in the cloud system can be dynamically scaled up and down depending upon end users service requests. Steady state queue size distribution is obtained using a recursive method assuming Markovian processes of arrival and service times. It has been shown that our queueing based model is effective in the web applications on cloud and helps in deciding the point of migration.

To measure performance metrics of a private cloud computing system, we applied finite source finite buffer multiple server queueing model. Two service policies have been proposed along with an analytical resource prediction model for private cloud system. Various performance measures such as utilization of the system, utilization of server, the expected number of client requests in the system, the expected number of client requests in the queue, mean number of idle servers, mean waiting time, response time and block-

ing probability are obtained. The numerical computations under a range of parameters show that the proposed models can optimize the organizational resources in cloud computing system.

In a cloud computing system, we consider two categories of users / clients: premier user and normal user. 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 determine the optimum number of VMs and the optimum system capacity to minimize the total expected cost. The numerical searching approach for the cost function is implemented using a genetic algorithm.

In comparison to the single tier applications the problem of resource allocation in multi-tier applications is harder as the tiers are not homogenous and a performance bottleneck in one tier may reduce the overall performance of the system thus violating the SLA. There is a need for automatic adaption of self-managing techniques to dynamically assign resources among applications of different clients on the basis of short-term demand estimates. We proposed an optimal autonomic virtual machine provisioning architecture for cloud data center to minimize the congestion in the network by varying the service rate of the virtual machines. An analytical model is developed for a cloud system with heterogeneous servers to dynamically scale up or scale down the number of servers depending upon end users service requests. The objective is to improve the efficiency and flexibility in cloud environment for resource provisioning.
The following directions may be pursued for future research work.

- With the increase in the popularity of cloud computing systems, virtual machine migrations across data centers and diverse resource pools will be greatly beneficial to data center administrators. The modeling and decision-making processes used by the process can be improved to confirm not only the changes in number of VMs but also changes in

each VM capacity. Furthermore, the queueing model can be improved to model the composite services and access to Cloud storage. The work can be extended to decide the point of migration for a migrating virtual machines.

- From the users perspective it is desirable to use as few resources as possible to minimize their pay-as-per-usage bill. Note that resources are provided in the form of virtual machines with a certain resource configurations. Therefore, it may be interesting to find out the types of virtual machines that should be used to meet application-level requirements. The solution to this may help the cloud users in taking decisions.

- Our work can be extended to analyze the resource scheduling policy for preemptive schedulers, allowing a high-priority task to interrupt the execution of a lower priority one.

- The proposed model can be extended for deployment of complex multi-tier applications in a cloud computing infrastructure.