Chapter 11
Quality of Services and Optimal Management of Cloud Centers with Different Arrival Modes

V. Goswami
KIIT University, India

S. S. Patra
KIIT University, India

G. B. Mund
KIIT University, India

ABSTRACT

In Cloud Computing, the virtualization of IT infrastructure enables consolidation and pooling of IT resources so they are shared over diverse applications to offset the limitation of shrinking resources and growing business needs. Cloud Computing is a way to increase the capacity or add capabilities dynamically without investing in new infrastructure, training new personnel, or licensing new software. It extends Information Technology's existing capabilities. In the last few years, cloud computing has grown from being a promising business concept to one of the fast growing segments of the IT industry. For the commercial success of this new computing paradigm, the ability to deliver guaranteed Quality of Services is crucial. Based on the Service Level Agreement, the requests are processed in the cloud centers in different modes. This chapter deals with Quality of Services and optimal management of cloud centers with different arrival modes. For this purpose, the authors consider a finite-buffer multi-server queuing system where client requests have different arrival modes. It is assumed that each arrival mode is serviced by one or more virtual machines, and different modes have equal probabilities of receiving services. Various performance measures are obtained and optimal cost policy is presented with numerical results. A genetic algorithm is employed to search optimal values of various parameters for the system.

DOI: 10.4018/978-1-4666-4715-2.ch011

1. INTRODUCTION

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.

Cloud Computing Technology has been developed from virtualization, utility computing, Infrastructure as a Service (IaaS), Platform as a Service (PaaS), Software as a Service (SaaS), etc. (Chen & Zheng, 2009). It provides IT business model where the users can acquire IT services through the internet. Cloud computing platform utilizes the high-speed internet to deliver the computing, storage, software and services (Foster, 2005) which are distributed all over the world, to the terminal users. It integrates the mass computing resources to compose one resource pool and serve the users dynamically with virtualized resources including computing, storage and services through network (Schiller, 2011). A user can rent all the services such as software, hardware, data and information from the cloud. The cloud computing platform can be subdivided into three layers. SaaS delivers the software through web browsers as a service of cloud computing platform. PaaS provides one platform for the users and developers with application development, test and deployment (Buyya et al., 2008). The platform includes database, middleware and development tools, for example, the Google Map platform and APP platform. IaaS provides the hardware infrastructure as servers, storage and hardware through internet. It is created based on virtualization technology as server and storage virtualization, e.g. EC2 of Amazon (Amazon, 2011) is one famous IaaS platform of cloud computing technology (Azeez, 2009).

The cloud computing platforms are of three types. Public Cloud serves the users distributed all over the world across the border of enterprises and areas. The public cloud platform is large-scale and composed of a few data centers in different areas to provide IaaS, PaaS or SaaS service (Hand, 2007). Private Cloud only serves for one company or organization. The widely used private cloud includes VCloud, VSphere of VMware and XEN Cloud of Citrix (Lee et al. 2006). Mixed Cloud owns the properties of public cloud and private cloud. It connects the resources of private clouds including its data, application and service through public cloud. It can guarantee the security of private cloud and support the permitted resources that can be exposed to the internet. OpenNebula is one famous mixed cloud platform (Delic, 2005).

There are some critical Quality of Service (QoS) parameters to be considered in cloud computing environment, such as time, cost (service charge for the user and servicing charge for provider), reliability and trust/security. In particular, QoS requirements are not static and need to be updated dynamically over the time due to continuous changes in the operating environments (Delic et al., 2007). That is, greater importance should be given to user's time as they pay for using services from the clouds based on time. In addition, dynamic negotiation of service level agreement (SLAs) between the users and the service provider is not completely supported in the cloud computing environment. Venugopal et al. (2008) have developed negotiation mechanisms based on alternate offers protocol for establishing SLAs. A 21st century vision of computing has been presented in Buyya et al. (2008). They also have identified various computing paradigms promising to deliver the vision of computing utilities. Cloud computing definitions and the architecture for creating market-oriented Clouds by leveraging technologies such as VMs are also discussed. Buyya et al. (2005) have proposed scheduling policies to address the time minimization and cost minimization problem in the context