Optimal Management of Cloud Centers with Different Arrival Modes for Cloud Computing Environment

Veena Goswami, KIIT University, India
Sudhansu Shekhar Patra, KIIT University, India
G. B. Mund, KIIT University, India

ABSTRACT

Cloud computing is a new computing paradigm in which information and computing services can be accessed from a Web browser by clients. Understanding of the characteristics of computer service performance has become critical for service applications in cloud computing. For the commercial success of this new computing paradigm, the ability to deliver guaranteed Quality of Services (QoS) is crucial. Based on the Service level agreement, the requests are processed in the cloud centers in different modes. This paper analyzes a finite-buffer multi-server queuing system where client requests have two arrival modes. It is assumed that each arrival mode is serviced by one or more Virtual machines, and both the modes have equal probabilities of receiving service. Various performance measures are obtained and optimal cost policy is presented with numerical results. The genetic algorithm is employed to search the optimal values of various parameters for the system.

Keywords: Cloud Computing, Cost Model, Genetic Algorithm, Queuing, Virtual Machines

1. INTRODUCTION

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, n.d.) 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 of Grid computing. Mukherjee and Sahoo (2010) studied a mathematical model for Market-Oriented Cloud Computing and proposed a Bee and Ant colony system based scheduling policy. Resource scheduling in cloud computing based on stochastic integer programming technique has been discussed in Li and Guo (2010). Cao et al. (2009) introduces an optimized algorithm for task scheduling based on activity based costing in cloud computing and its implementation. The theoretical foundations of ethical and entrepreneurial adoption behavior in cloud computing services have been discussed in Ratten (2012).

In this paper, we examined various system measures when two categories of client requests are coming to the cloud center, one may be from the premier subscribers and other from the ordinary subscribers. We analyze the cloud model using a finite-buffer multi-server queuing system where client requests have two arrival modes. It is assumed that each arrival mode is serviced by one or more Virtual machines, and the two arrival modes have equal probabilities of receiving service. The inter-arrival and service-times are exponentially distributed. A cost model is developed to determine the optimum number of VMs and the optimal system capacity which minimizes the total expected cost. The numerical searching approach for the cost function is implemented using the genetic algorithm. A genetic algorithm is a search technique used in computing to find exact or approximate solutions to optimization and search problems. Various performance measures are provided and
Related Content

Towards Federation and Interoperability of Cloud Storage Systems
Sebastian Dippl, Michael C. Jaeger, Achim Luhn, Alexandra Shulman-Peleg and Gil Vernik (2013). Data Intensive Storage Services for Cloud Environments (pp. 60-71).
www.igi-global.com/chapter/towards-federation-interoperability-cloud-storage/77431?camid=4v1a

Supporting Service Level Agreement with Service Discovery
www.igi-global.com/chapter/supporting-service-level-agreement-service/52191?camid=4v1a

Business to Consumer E-Services: Australian Accounting Practices and their Web Sites
www.igi-global.com/article/business-consumer-services/2155?camid=4v1a
Business Integration as a Service

[www.igi-global.com/article/business-integration-service/64633?camid=4v1a](www.igi-global.com/article/business-integration-service/64633?camid=4v1a)