Performance Analysis of Cloud Computing Centers for Bulk Services

Veena Goswami, School of Computer Application, KIIT University, Bhubaneswar, India
Sudhansu Shekhar Patra, School of Computer Application, KIIT University, Bhubaneswar, India
G. B. Mund, School of Computer Engineering, KIIT University, Bhubaneswar, India

ABSTRACT

Cloud is a service oriented platform where all kinds of virtual resources are treated as services to users. Several cloud service providers have offered different capabilities for a variety of market segments over the past few years. The most important aspects of cloud computing are resource scheduling, performance measures, and user requests. Sluggish access to data, applications, and web pages spoils employees and customers alike, as well as cause application crashes and data losses. In this paper, the authors propose an analytical queuing model for performance evaluation of cloud server farms for processing bulk data. Some important performance measures such as mean number of tasks in the queue, blocking probability, and probability of immediate service, and waiting-time distribution in the system have also been discussed. Finally, a variety of numerical results showing the effect of model parameters on key performance measures are presented.

Keywords: Blocking Probability, Bulk Data, Cloud Computing, Performance Evaluation, Queuing Model

1. INTRODUCTION

Cloud computing is an emerging commercial infrastructure paradigm which provides shared information and communication technology services. The use of virtualization and resource time sharing offers a number of desirable properties. Clouds serve, with a single set of physical resources, a large user base with different needs. Thus, clouds have the potential to provide to their owners the benefits of an economy of scale and scaling the hardware resources affirming a system with demand as well as providing high availability through hosting in geographically dispersed data centers. It entails a service-oriented architecture, reduces information technology overhead for the end-users, provides great flexibility, cuts down total cost of ownership, and may provide on demand services (Vouk, 2008).

The cloud computing is the development of distributed computing, parallel computing and grid computing that has brought platform virtualization (Buyya, Yeo, & Venugopal, 2008)

DOI: 10.4018/ijcac.2012100104
technology into the field of high performance computing. Virtualization offers both flexibility and security through custom user images and user isolation. Due to commercial interests, a cloud provider normally has a proprietorship where a user makes an image that only runs on the provider’s site. Many cross-site technologies applied in grid computing can easily be adopted by cloud computing. The actual configuration and management of cloud computing differs from traditional computing paradigms as it is scalable, can be encapsulated as an abstract entity, and the services are dynamically configurable. Cloud Computing infrastructure allows users to achieve more efficient use of their IT hardware and software investments with super-user privileges on-demand. This is accomplished by analyzing the physical barrier inherent in isolated systems, automating the management of the group of the systems as a single entity.

Any task request sent to the cloud management system (CMS) is serviced within a suitable service node; after the completion of the service, the task leaves the center. A service node may contain resources such as web servers, database servers, directory servers and others. A service level agreement (SLA) outlines all aspects of cloud service usage and the obligations of both service providers and clients, including various descriptors collectively referred to as Quality of Service (QoS). Cloud computing may be considered as a resource available as a service for virtual data centers. For example, in Amazons S3 Storage Service cloud computing and virtual data centers are not the same (Rittinghouse, & Ransome, 2010).

QoS is the criterion which measures the satisfaction of users using the Cloud Computing services. Due to dynamic nature of cloud environments, diversity cloud computing delivers three kinds of services: Infrastructure as a service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS) (Foster, Fidler, Roy, Sander, & Winkler, 2004, Rimal & Eunmi, 2009). These services are available to users in a Pay per-use-on-demand model, which can access shared IT resources like server, data storage, application, network and so on through internet. Iosup et al. (2011) analyzed the performance of cloud computing services for scientific computing workloads. Users of cloud computing services are faced with the problem of transferring large amounts of data across a geographically dispersed network under time and budget constraints. This problem arises because cloud computing services work on the scale of terabyte and are composed of geographically dispersed and distant sites. Cloud partners like to combine large amounts of data, from multiple distributed locations to a single datacenter. There are many reasons to transfer data between sites, for example, users may have to perform initial transfers, cloud service provider migrations, site backups, or user data migrations. Furthermore to perform their data transfers within a certain time deadline and to minimize the cost of the transfer. In particular, the challenge of reducing the latency of the transfer, without incurring excessive costs is the bottleneck for cloud users (Chung, 2000). Cloud providers charge their cloud users for bandwidth utilization when the size of data transfer is high, or when the transfers are frequent.

A variety of cloud applications require bulk data transfer, such as scientific computing, internet protocol television (IPTV) and real time multimedia streaming, etc. For such applications, data resources provide large amounts of data and users to access the data which is geographically dispersed in most cases. To improve the performance of the system, they must be processed in bulk in accordance to the service level agreement (SLA). In case of IPTV service, it simply relies on the availability of sufficient network and data resource which must be operated within the limitation of time constraints, otherwise the corresponding application requirements may not be guaranteed. Therefore, the reservation and utility optimization of the combined resources become critical issues to guarantee the bulk data transfer in cloud computing. It provides optimized and efficient computing through enhanced collaboration, agility, scalability and availability.
Related Content

Service Concepts
www.igi-global.com/chapter/service-concepts/18306?camid=4v1a

IT Audit for Information Logistics
Małgorzata Pankowska (2010). Service Science and Logistics Informatics: Innovative Perspectives (pp. 128-143).
www.igi-global.com/chapter/audit-information-logistics/42639?camid=4v1a

A Study on Relationships Among Software Engineering Capabilities in Japan Using Panel Analysis
www.igi-global.com/article/study-relationships-among-software-engineering/59880?camid=4v1a
Business Models for Insurance of Business Web Services
Liu Wenyin, An Liu, Qing Li and Liusheng Huang (2011). Service Intelligence and Service Science: Evolutionary Technologies and Challenges (pp. 261-272).
www.igi-global.com/chapter/business-models-insurance-business-web/47366?camid=4v1a