
Guofeng Yan, School of Computer and Communication, Hunan Institute of Engineering, Hunan, China & School of Computer, National University of Defense Science and Technology, Changsha, China

Yuxing Peng, School of Computer, National University of Defense Science and Technology, Changsha, China

Shuhong Chen, School of Information Science and Engineering, Central South University, Changsha, China

Pengfei You, School of Computer, National University of Defense Science and Technology, Changsha, China

ABSTRACT

Quality of service (QoS) optimization for end-to-end (e2e) services always depends on performance analysis in cloud-based service delivery industry. However, performance analysis of e2e services becomes difficult as the scale and complexity of virtualized computing environments increase. In this paper, the authors present a novel hierarchical stochastic approach to evaluate the QoS of e2e virtualized cloud services using Quasi-Birth Death structures, where jobs arrive according to a stochastic process and request virtual machines (VMs), which are specified in terms of resources, i.e., VM-configuration. To reduce the complexity of performance evaluation, the overall virtualized cloud services are partitioned into three sub-hierarchies. The authors analyze each individual sub-hierarchy using stochastic queueing approach. Thus, the key performance metrics of e2e cloud service QoS, such as acceptance probability and e2e response delay incurred on user requests, are obtained.

Keywords: Cloud Services, Performance Evaluation, Quality of Service, Quasi-Birth Death Structures, Stochastic Model, Virtualized Computing Environments

INTRODUCTION

In most cloud services, virtual machine instances normally share physical processors and I/O interfaces with other instances. It is expected that virtualization can impact the computation and communication performance of cloud services. Currently, most of the research work related to cloud computing has dealt with implementation issues, while performance-related issues have

DOI: 10.4018/IJWSR.2015010103
received much less attention (Maguluri et al., 2012). Walker (2008) compares the application level performance of Amazon EC2 against another high performance computing cluster, NCSA, and shows that Amazon EC2 has much worse performance than traditional scientific clusters. To overcome the major delays during data exchange for large data sets in the cloud, Broberg et al. (2009) and Fedak et al. (2009) propose MetaCDN and BitDew, respectively. MetaCDN reduces the complexity of dealing with multiple storage providers on the cloud and provides a unified name space, and BitDew addresses the issue of large-scale data management in the cloud and the grid environments. Based on empirical measurements, Wang and Eugene Ng (2010) make the first study focusing on the end-to-end (e2e) networking performance of Amazon EC2 instances and on understanding the impact of virtualization on the data center network performance. They analyze the abnormal large delay variations and unstable TCP/UDP throughput caused by end host virtualization. Their findings are helpful to explain the observations in (Walker, 2008). Mei et al. (2013) first show that current implementation of virtual machine monitor does not provide sufficient performance isolation to guarantee the effectiveness of resource sharing across multiple virtual machine instances running on a single physical host machine, especially when applications running on neighboring VMs are competing for computing and communication resources. Then we present the detailed analysis on different factors that can impact the throughput performance and resource sharing effectiveness.

To study the response time in terms of various metrics, such as the overhead of acquiring and realizing the virtual computing resources, Yigitbasi et al. (2009) design and implement C-Meter, a portable, extensible, and easy-to-use framework for generating and submitting test workloads to computing clouds. Using C-Meter, researchers can be easy to evaluate different scheduling algorithms. In addition, theoretical analyses on cloud services quality mostly rely on performance evaluation of M/G/m queuing systems (Miyazawa, 1986; Yao, 1985).

In (Xiong & Perros, 2009), the cloud centers are modelled as M/G/m queuing systems, which have been used to compute the distribution of response time based on the assumption that inter-arrival time and service time are both exponential. Xiong and Perros (2009) use the distribution of response time to study the relationship among the minimal service resource, the maximal request number, and the service rank. For M/G/m systems, however, the response time distribution and the queue length cannot obtain the closed form solutions. As a result, suitable approximations are sought. Using continuous time Markov chains (CTMCs), Yang et al. (2013) analyze the transient behaviors and steady states of three different virtual machine-based architectures, i.e., load balance server architecture (LBSA), isolated component server architecture (ICSA), and Byzantine fault tolerant server architecture (BFTSA). Most of these provide reasonably accurate estimates of mean response time only when the number of servers is comparatively small, (say, less than twenty or so), as well as small coefficient of variation of service time (less than unity), but fail for a large number of servers and higher coefficient of variation (Khazaei et al., 2012). Humphrey et al. (2013) present the Cloud Distributed Research Network (namely Cloud-DRN) which leverage the cloud for availability, reliability, scalability, and improved security as compared to legacy distributed systems while still supporting site autonomy. In (Ghosh et al., 2010; Khazaei et al., 2012; Ren et al., 2013), the researchers evaluate the quality of service (QoS) of end clients by analyzing the performance of cloud server farms with general service time. The researchers propose a general analytic model for e2e performance of cloud services (Ghosh et al., 2010). However, the proposed model is limited to the single arrival of requests and the startup delay of cold physical machines (PMs) has not been captured. Also, the effect of virtualization is not reflected in their results (Lu et al. 2013; Mi et al. 2013).
Visual Inspection System and Psychometric Evaluation with Correlation for Multiple Perceptions
www.igi-global.com/chapter/visual-inspection-system-psychometric-evaluation/53283?camid=4v1a

Reputation Management for Composite Services in Service-Oriented Systems
www.igi-global.com/article/reputation-management-composite-services-service/55235?camid=4v1a