J. Vis. Commun. Image R. 25 (2014) 928–942

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Queueing model based resource optimization for multimedia cloud

Xiaoming Nan *, Yifeng He, Ling Guan

Ryerson University, Toronto, Ontario M5B 2K3, Canada

ARTICLE INFO

Article history:
Received 17 October 2013
Accepted 16 February 2014
Available online 28 February 2014

Keywords:
Multimedia cloud
Resource optimization
Queueing model
Response time
Resource cost
Quality of service (QoS)
Convex optimization
Priority service

ABSTRACT

Multimedia cloud is a specific cloud computing paradigm, focusing on how cloud can effectively support multimedia services. For multimedia service providers (MSP), there are two fundamental concerns: the quality of service (QoS) and the resource cost. In this paper, we investigate these two fundamental concerns with queueing theory and optimization methods. We introduce a queueing model to characterize the service process in multimedia cloud. Based on the proposed queueing model, we study resource allocation problems in three different scenarios: single-service scenario, multi-service scenario, and priority-service scenario. In each scenario, we formulate and solve the response time minimization problem and the resource cost minimization problem, respectively. We conduct extensive simulations with practical parameters of Windows Azure. Simulation results demonstrate that the proposed resource allocation schemes can optimally allocate cloud resources for each service to achieve the minimal response time under a certain budget or guarantee the QoS provisioning at the minimal resource cost.

1. Introduction

Recent years have witnessed the fast development of cloud computing. As the emerging computing paradigm, cloud computing manages a shared pool of servers to provide on-demand computation, communication, and storage resources as services in a scalable manner. According to the report from International Data Corporation (IDC) [1], the worldwide public cloud computing services will approach $100 billion by 2016 and enjoy an annual growth rate of 26.4%, which is five times the traditional IT industry. Based on the service provisioning at different levels, three cloud service models have been proposed [2,3], namely Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS), among which the SaaS cloud is most familiar by individual users. In the SaaS cloud, users send requests to cloud data center, which can offer software services and then deliver the processing results back to users. By using cloud-based services, users are free from application installation and software maintenance.

Among various cloud-based software services, multimedia services have strong demands for cloud computing. As well known, multimedia services, like media retrieval, video on demand (VOD), and free viewpoint video (FVV), typically require intensive computation and/or intensive bandwidth resources, which are burdens to client devices, especially to the resource-constrained mobile devices. The emergence of cloud computing provides a way to resolve this problem. By migrating multimedia processing to cloud, the hardware requirements on the user side have been greatly reduced. Users are able to access interested cloud services from anywhere at anytime. The elastic and on-demand natures of resource provisioning in cloud effectively satisfy the intensive resource demands of multimedia processing.

Multimedia services bring new issues to current general-purpose clouds. Nowadays, general-purpose clouds employ an utility based resource management to allocate computation resources (e.g. CPU, memory, bandwidth, etc.) for applications. The only guaranteed parameter in the Service Level Agreement (SLA) is the resource availability, like Amazon EC2 [4]. However, in addition to the computation resources, another important factor for multimedia services is the stringent quality of service (QoS) requirement. Therefore, simply using the general-purpose cloud to provide multimedia services without considering QoS requirements may suffer from the unacceptable user experience. To better utilize cloud for multimedia services, Zhu et al. [5] introduced the concept of multimedia cloud computing. In multimedia cloud, the key is how to supply various multimedia services and guarantee the QoS requirements for all users.

For the *multimedia service provider* (MSP), there are two major concerns: the *QoS* and the *resource cost*. Generally, different multimedia services have different QoS requirements. For instance, video streaming services usually take the visual quality, the packet loss ratio, and the delay as QoS measurements, image/video retrieval services concern about the accuracy and the time

* Corresponding author.
E-mail addresses: xnan@ee.ryerson.ca (X. Nan), yhe@ee.ryerson.ca (Y. He), lguan@ee.ryerson.ca (L. Guan).

http://dx.doi.org/10.1016/j.jvcir.2014.02.008
1047-3203/

consumed for each query, and the remote rendering services, like cloud gaming [6] or virtualized screen [7], are evaluated by the interaction delay and the rendered image quality. Delay is an important QoS metric for many multimedia services such as image/video retrieval, video streaming, and cloud gaming. A low delay will lead to a high QoS, while a high delay will degrade the user experience. In this paper, we focus on delay-sensitive cloud-based multimedia services, in which the response time in data center is the dominant component in end-to-end service delay. Thus, the response time is taken as the QoS factor to measure the performance of cloud-based multimedia services. The response time in cloud is defined as the duration from the time when the request arrives at the data center to the time when the service result completely departs from the data center. A lower response time means a faster service for the user's request. Therefore, it is important for the MSP to meet different response time requirements for all users. Besides the QoS requirement, another concern is the resource cost. In order to provide services, the MSP needs to rent the required resources from the *cloud service provider* (CSP) according to the demands. If the allocated resources are far more than the real demands, resources cannot be fully utilized, leading to the unnecessary resource cost. Conversely, if the allocated resources are less than the demands, the QoS requirements cannot be guaranteed. Therefore, it is significant for the MSP to learn the dynamic demands from users and accordingly determine the optimal resource allocation, which can simultaneously satisfy QoS requirements and minimize the resource cost.

However, it is challenging to satisfy these two concerns. *First*, there exist different types of multimedia services, which have heterogenous resource demands. It is a challenge for the MSP to quantify the resource demands and optimally configure resources for each service. *Second*, different multimedia services have different requirements on response time. For example, compared to the media content delivery, the cloud-based video gaming needs a lower response time such that users can enjoy a smooth and real-time interactive experience. The MSP should optimize resources to meet different response time requirements for various types of multimedia services. *Third*, there is a trade-off between the response time and the resource cost. The under-provisioned resources would slow down the service and deteriorate QoS, while the over-provisioned resources would result in the unnecessary waste. For the MSP, it is difficult to provide satisfactory services at a low resource cost. *Last but not least*, multimedia services may demand different priorities in the practical cloud. For example, some urgent applications, like real-time health monitoring, demand a higher priority to process abnormal events. However, current resource management schemes are mainly for the first-come first-served (FCFS) service discipline, which cannot effectively adapt to services with heterogeneous priorities. Therefore, a challenge for the MSP is how to optimize resources for the differentiated services and satisfy different QoS requirements.

To address above mentioned challenges, we study the optimal resource allocation for multimedia cloud in this paper. Our contributions can be summarized as follows:

1. We investigate the cloud resource allocation problem using queueing theory. Specifically, we introduce a queueing model to characterize the service process in multimedia cloud. The proposed queueing model consists of three concatenated queueing systems, namely the *schedule queue*, the *computation queue*, and the *transmission queue*. We theoretically analyze the equilibrium demands for the schedule, computation, and bandwidth resources of multimedia services and derive the relationship between the allocated resource capacity and the response time.
2. Based on the proposed queueing model, we study the resource optimization for multimedia cloud in three different scenarios: the single-service scenario, the multi-service scenario, and the priority-service scenario. In each scenario, we formulate and solve the response time minimization problem and the resource cost minimization problem, respectively, to determine the optimal resource allocation for each multimedia service.
3. We conduct extensive simulations to evaluate the performance of our proposed resource optimization schemes in the realistic settings. The simulation results indicate that the proposed resource optimization schemes can optimally allocate cloud resources to achieve the minimal response time under a certain budget or provide the satisfactory services at the minimal resource cost.

The remainder of this paper is organized as follows. Section 2 reviews the related work. Section 3 presents the proposed system models, including the data center architecture, the queueing model, and the resource cost model. Based on the proposed models, we optimize resources for multimedia cloud in Section 4 under the single-service scenario, the multi-service scenario, and the priority-service scenario, respectively. Section 5 presents extensive performance evaluations. Finally, we conclude the paper in Section 6.

2. Related work

In this section, we review the related literature in the areas of cloud-based multimedia services, cloud resource allocation schemes, and resource optimization for cloud-based multimedia services.

2.1. Cloud-based multimedia services

Cloud-based multimedia services have been widely studied in recent years [5,8–11]. Zhu et al. [5] discussed the challenges imposed by cloud multimedia processing. The current general-purpose cloud uses a utility like mechanism to manage resources and cannot guarantee QoS requirements for multimedia services. Zhu et al. [5] stated that simply using the general-purpose cloud to deal with multimedia services may lead to unacceptable media QoS. Therefore, they introduced a concept of multimedia cloud computing from the perspectives of the multimedia-aware cloud and the cloud-aware multimedia. In addition, they proposed a media edge cloud computing architecture, which physically placed computing servers at the edge of a cloud in order to reduce transmission delays between users and data centers. But the detailed resource management schemes are not discussed in [5]. With powerful computing resources, cloud computing is an ideal platform to support multimedia applications on mobile devices. Miao et al. [8] proposed a cloud-based free viewpoint video (FVV) rendering framework for mobile devices over cellular networks. In the framework, they performed a joint rendering allocation between cloud and client based on the required QoE. The cloud rendering is conducted during the stationary-viewing-time to achieve a high visual quality, while the local rendering is applied during the switch-viewing-time in order to conceal the interaction delay. Wang et al. [9] studied cloud mobile gaming, an approach that enables multi-player games on mobile devices. In cloud mobile gaming, the computation intensive tasks, like three-dimensional model rendering, are executed on cloud servers in response to the control commands from players, and the rendered frames are compressed and streamed to mobile devices as video streaming. Cloud computing has also been widely used in social media applications. Wu et al. [10] presented a generic mobile social TV framework, known as *CloudMoV*, which makes use of both an IaaS cloud and a PaaS