An Optimal Resource Allocation Scheme in Cloud Computing

Cuong T. Do, Pham Chuan, Choong Seon Hong
Department of Computer Engineering, Kyung Hee University
Email: {dtcuong,pchuan,cshong}@khu.ac.kr

Abstract

Multimedia cloud, as a strict QoS requirement cloud paradigm, addresses how cloud can effectively process multimedia services for multimedia applications. In this paper, we optimize resource allocation for multimedia cloud based on queuing model. Numerical results demonstrate that the proposed optimal allocation scheme can optimally utilize the cloud resources to achieve a maximum revenue.

1. Introduction

Recently, cloud computing becomes more and more popular in large-scale computing and has ability to share data and computations over a scalable network of nodes [1]. It is known that the underlying cloud computing environment is inherently a large numbers of independent computing and communication resources and data stores. In an open cloud computing framework, utilizing cloud resources and scheduling tasks that guarantees Quality of Service (QoS) constrains present a challenging technical problem. In cloud computing, the response time is a key QoS performance criterion for multimedia cloud applications. Due to limited resources of a mobile device, multimedia processing such as image/video retrieval, typically requires intensive computation, which is difficult to be performed on mobile devices.

Various resource management techniques have been proposed for cloud resource management [2], [3], [4]. A self-organizing model to manage cloud resources is proposed in [2] without centralized management control. Authors in [3] focus on the maximization of the steady-state throughput by deploying resources for the independent equalized tasks in the cloud. In [4], the authors optimize resource allocation for multimedia cloud based on the service response time in both single-class service case and multiple-class service case.

Compared to these previous work, our work demonstrates the following novelties: 1) we study the relationship between QoS, multimedia cloud provider's revenue and cloud resource allocation in different scenarios base on queuing model; 2) we analyze the cloud resource allocation and provide optimal resource allocation respectively to meet users' constraints and maximize the MSP's revenue.

2. Queuing Model

Most of multimedia clouds are built in the form a multimedia cloud server farm which consists a bunch of computing servers. Computing servers act as the real processors, which receive tasks through the multimedia cloud service load balancer and then process users' requests using their own resources and associated media data [1]. We assume the latency of internal communications between the multimedia cloud service load balancer and the multimedia cloud server farm is negligible. Thus, all tasks requested by user can be done simultaneously in parallel in the multimedia cloud server farm [1]. After processing, all the media service results will be transmitted back to users.

We model a multimedia cloud server farm as a M/M/m/m queuing system, presented in [5], which indicates the inter-arrival time and task service times of requests are exponentially distributed. The principal quantity of interest here is the probability that a request arrival will find all m servers busy and will therefore be evicted. The system under consideration contains m servers which render service in order of task request arrivals (FCFS). The capacity of system is m which means there is zero buffer size for incoming request. This is a reasonable model because the multimedia applications require immediate service response (i.e., no waiting in the input buffer) as a strict QoS requirement of real-time multimedia applications such as IPTV, voice over IP and online webinars applications. As the population size of a typical cloud center is relatively high while the probability that a given user will request service is

relatively small, the arrival process can be modeled as a Markovian process. It means that task interarrival time is exponentially distributed with a rate of $1/\lambda$. The service times of the requests are identical independently distributed (i.i.d.) random variables (r.v.s) following exponential distribution with parameter μ (service rate). Let's Π_i denotes the stationary distribution of the system having i working servers. According to [5], we have

$$\Pi_0 = \left[\sum_{i=0}^{m} \left(\frac{\lambda}{\mu}\right)^i \frac{1}{i!}\right]^{-1}. \quad (1)$$

The probability that a request arrival will find all m servers busy and will be evicted is

$$\Pi_m = \frac{(\lambda/\mu)^m / m!}{\sum_{i=0}^{m} (\lambda/\mu)^i / i!}. \quad (2)$$

3. Revenue Model

In a real-life scenario, cloud computational resources are shared among different cloud users who will pay for the services according to their usage of resource. Generally, the resource details are hidden from users through virtualization. Observed from user perspective, services are identical in terms of functionality and interface. In this paper, we employ a linear function to model the relationship between the payment of users to the service provider and allocated resources. Thus, the user pays the servicing fee of p per task unit.

It is not reasonable to provide the same QoS to the users who would like to pay more for better services. Unlike best effort service users, real-time multimedia cloud users request immediate service. However due to limited resources, the MSP has ability to provide m immediate services by m servers. Thus, the MSP has to evict all requesting-service if all m servers are occupied as illustrated by Figure 1. It means that the MSP does not satisfy the service level agreement with the users. Therefore, each evicted request of users is compensated by a reimbursement of ϵ, where $\epsilon = \beta p$. We assume $0 \le \beta \le 1$ to represent the tolerance of users. The less b is the more tolerance of users is. Then, the revenue of the MSP is given as follows

$$\Re = p\sum_{i=1}^{m} i\Pi_i - \varepsilon(\lambda/\mu)\Pi_m - c\sum_{i=1}^{m} i\Pi_i - mC. \quad (3)$$

The first part of the revenue Â is the average profit which the MSP obtains by charging users p per completed task request unit. The second part is the average cost due to eviction of a task request. The third part is the average computing cost where c is the computing cost of a working server. The fourth part, mC, is the cost for infrastructure deployment of m servers.

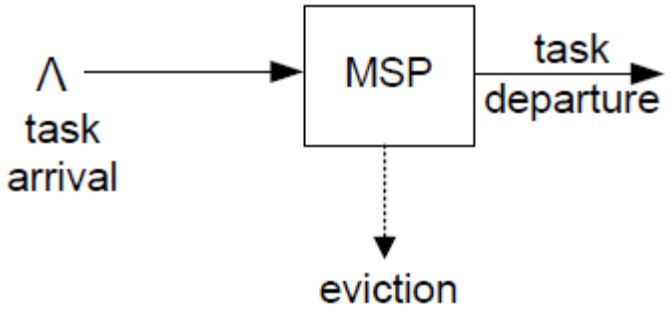

Fig.1 Task arrival, evict and departure scheme.

4. Optimal Resource Allocation for MSP

Since different applications often have different requirements on service response time, it is challenging for cloud providers to meet all users' requirements with the minimal resource cost. Therefore, we formulate the optimal resource allocation for multimedia application problem, which can be stated as: to maximize the total revenue of the MSP by determining the optimal number of servers under the eviction probability constraint of users. Mathematically, the optimal resource allocation for multimedia application problem can be formulated as Problem I as follows.

$$\max_{m} \quad \Re(m) = p\sum_{i=1}^{m} i\Pi_i - \varepsilon(\lambda/\mu)\Pi_m - c\sum_{i=1}^{m} i\Pi_i - mC$$

$$\text{s.t.} \quad \Pi_m \le I,$$

where I is a given upper bound of the probability of eviction. It means that I is the maximum average evicted task requests that the multimedia application can tolerate. Problem I is a integer maximization problem, however, it can be effectively solved by numerical method because it has only one variable m.

5. Numerical analysis

Numerical Setting: We perform numerical analysis to evaluate the proposed scheme. The resource of the computing server are charged by p =1$/request, the cost of employing a working server is c =0.2$/server, and the infrastructure cost C=0.1$/server, respectively. The mean arrival rate of users multimedia requests I is

given in range of 50 to 500 requests/second. The mean service rate μ = 1 request/second. The tolerance of user b is 0.5 and the upper bound of the probability of eviction l is 0.05.

Figure 2 shows the shape of the revenue function $\mathcal{R}(m)$. We zoom in the curve of the revenue function $\mathcal{R}(m)$ in the range of 200 to 300 servers in order to estimate the optimal number of server by observing. Thus, the revenue value $\mathcal{R}(m)$ of the MSP is increasing from m =1 to the optimal m^* = 231, then, is decreasing slowly if we continue increase number of server m. In order to observe the effect of arrival rate l of request to the optimal number of server m^*, we vary the value of l as {50,100,200,300,400,500}. As can be seen in Figure 3, when the arrival rate of request increases, we need to employ more servers to response. The optimal number of server likely increase linearly by the increasing of the arrival rate of request.

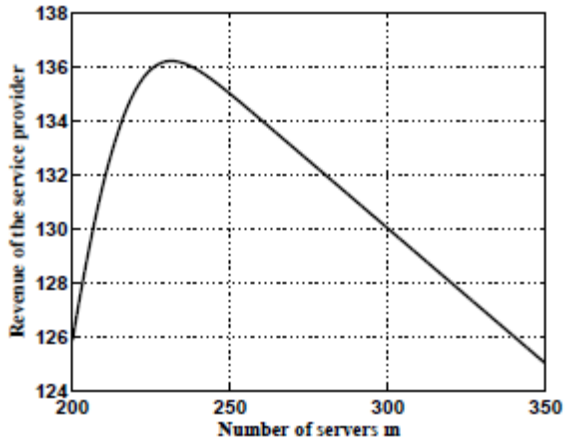

Fig.2 Revenue of the MSP with $\lambda = 200$ request/second.

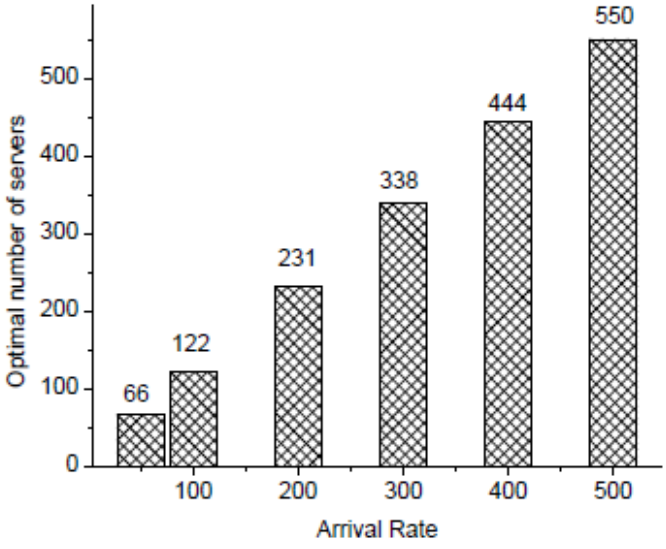

Fig.3 Optimal number of server vs. Arrival rate of request λ.

Acknowledge

This research was funded by the MSIP (Ministry of Science, ICT & Future Planning), Korea in the ICT R&D Program 2014. *Dr. CS Hong is the corresponding author

References

[1] W. Zhu, C. Luo, J. Wang, and S. Li, "Multimedia cloud computing," Signal Processing Magazine, IEEE, vol. 28, no. 3, pp. 59–69, 2011.

[2] W. Lin and D. Qi, "Research on resource self-organizing model for cloud computing," in Internet Technology and Applications, 2010 International Conference on. IEEE, 2010, pp. 1–5.

[3] H. Shi and Z. Zhan, "An optimal infrastructure design method of cloud computing services from the bdim perspective," in Computational Intelligence and Industrial Applications, 2009. PACIIA 2009. Asia-Pacific Conference on, vol. 1. IEEE, 2009, pp. 393–396.

[4] X. Nan, Y. He, and L. Guan, "Optimal resource allocation for multimedia cloud in priority service scheme," in Circuits and Systems (ISCAS), 2012 IEEE International Symposium on. IEEE, 2012, pp. 1111–1114.

[5] D. P. Bertsekas, R. G. Gallager, and P. Humblet, Data Networks, 2nd ed. Prentice-hall Englewood Cliffs, 1992.