# Optimal Resource Allocation for Multimedia Application in Single and Multiple Cloud Computing Service Providers

Cuong T. Do*, Duy T. Do†, Nguyen H. Tran*, Dai H. Tran*, Kyi Thar*, Choong Seon Hong*
*Department of Computer Enginneering, Kuyng Hee University, 446-701, Korea
†Faculty of Information Technology, University of Transport and Communications, Vietnam
Email: {dtcuong, nguyenth, dai.tran, kyithar, cshong}@khu.ac.kr*, atlantic1295@gmail.com†

*Abstract*—**In this paper, we optimize resource allocation for multimedia cloud based on queuing model. Specifically, we optimize the resource allocation in both single multimedia service provider (MSP) scenario and multiple MSPs scenario. In each scenario, we formulate and solve the MSPs' revenue maximization problem under eviction probability constraint of users. Numerical results demonstrate that the proposed optimal allocation scheme can optimally utilize the cloud resources to achieve a maximum revenue.**

## I. Introduction

In cloud computing, Zhu et.al. focus on how cloud can provide QoS provisioning for multimedia applications and services [1]. Cloud service providers use their own resources as utilities to process multimedia requests. Then the computing multimedia results are returned to users who do not need to pay for costly computing devices. By using multimedia cloud service, the users can process multimedia applications on powerful cloud servers of the provider and pay for the utilized resources by their usage period.

There are two major concerns for multimedia cloud service providers. The first concern is the QoS which includes response time, probability of immediate service, and mean number of tasks in the system [2]. Thus, it is important for multimedia cloud service providers to choose an appropriate queue system to meet users' requirements on service response time. The second concern is the cost of the allocated could resources. For multimedia could service providers, the cloud service should be scheduled and computed in an optimal manner. Inappropriate resource allocations will result in resource waste and revenue degradation of the multimedia service providers (MSPs). Therefore, it is challenging for MSPs to optimally allocate resources to maximize their revenue and satisfy users QoS requirements at the same time.

In cloud computing, several approaches have been proposed to tackle with resource management [3], [4]. Recently, several works have addressed resource management based on queueing performance analysis [5], [6], [7], [8], [9]. In [10], the distribution of response time was obtained for a cloud center modeled as the classical open network, assuming that both interarrival and service times are exponential. In [11], the distribution of response time was obtained for a cloud center modeled as an M/G/m/m+r queuing system. However, when interarrival time and/or service time are not exponential, the analysis is more complex [11]. If both interarrival and service times were assumed to be exponentially distributed, and the system had a finite buffer of size then we have the M/M/m/m+r queueing model [12], [13].

Considering as a specific cloud, Multimedia cloud mainly addresses how cloud can process multimedia applications and provide QoS provisioning for multimedia services. In order to enable subscribers to receive television programs and video streams from anywhere, Lau et al. [14] develop an architectural framework to employ on-demand cloud resources on IPTV. In [15], an efficient video-based mobile location search application is implemented in cloud environment to perform scalable and adaptive online monitoring.

Various resource management techniques have been proposed for cloud resource management [3], [4], [9]. A self-organizing model to manage cloud resources is proposed in [3] without centralized management control. Authors in [4] focus on the maximization of the steady-state throughput by deploying resources for the independent equalized tasks in the cloud. In [9], 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 in both single MSP and multiple MSPs scenarios, and provide optimal resource allocation respectively to meet users' constraints and maximize the MSPs' revenue.

## II. System Model

This section presents our system models, including multimedia cloud computing architecture, queueing model and revenue model of the multimedia service provider.

### A. Multimedia Cloud Computing Architecture

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.

### B. Queuing Model

We model a multimedia cloud server farm as a M/M/m/m queuing system, presented in [16] and [17], which indicates the interarrival 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 $\frac{1}{\lambda}$. The service times of the requests are identical independently distributed (i.i.d.) random variables (r.v.s) following exponential distribution with parameter $\mu$ (service rate). The transitions represent the state transitions by the task request arrivals and departures. Let's $\Pi_i$ denotes the stationary distribution of the system having $i$ working servers. Thus, we have the global balance equation as

$$\lambda\Pi_{i-1} = i\mu\Pi_i, i = 1,2,...,m. \tag{1}$$

Therefore, we obtain

$$\Pi_i = \Pi_0 \left(\frac{\lambda}{\mu}\right)^i \frac{1}{i!}, i = 1,2,...,m, \tag{2}$$

augmented by the normalization equation

$$\sum\nolimits_{i=0}^{m} \Pi_i = 1. \tag{3}$$

Thus, we have

$$\Pi_0 = \left[\sum\nolimits_{i=0}^{m} \left(\frac{\lambda}{\mu}\right)^i \frac{1}{i!}\right]^{-1}. \tag{4}$$

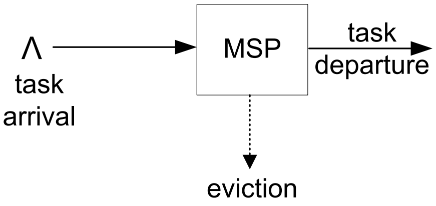

Fig. 1. Task arrival, evict and departure scheme.

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

$$\Pi_m = \frac{(\lambda/\mu)^m/m!}{\sum_{i=0}^{m} (\lambda/\mu)^i \big/ i!}. \tag{5}$$

### C. 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 $\varepsilon$, where $\varepsilon = \beta p$. We assume $0 \le \beta \le 1$ to represent the tolerance of users. The less $\beta$ is the more tolerance of users is. It is also possible to differentiate the reimbursement rate $\varepsilon_i$ between MSPs. Then, the revenue of the MSP is given as follows

$$\Re = p\sum\nolimits_{i=1}^{m} i\Pi_i - \varepsilon(\lambda/\mu)\Pi_m - c\sum\nolimits_{i=1}^{m} i\Pi_i - m\mathrm{C}. \tag{6}$$

The first part of the revenue $\Re$ 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, $m$C, is the cost for infrastructure deployment of $m$ servers.

## III. Optimal Resource Allocation for Multimedia Application

In this section, we use the proposed queueing models to study the resource allocation problems in single MSP scenario and multiple MSP scenario, respectively. In each scenario, we maximize the MSP's revenue under the users' QoS constraint.

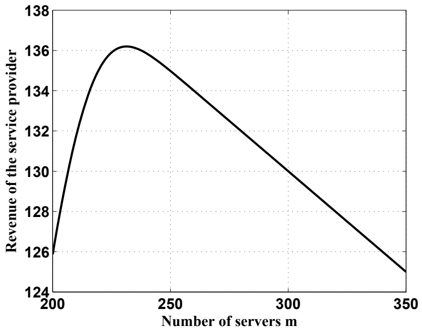

Fig. 2. Revenue of the MSP with λ = 200 requests/second.

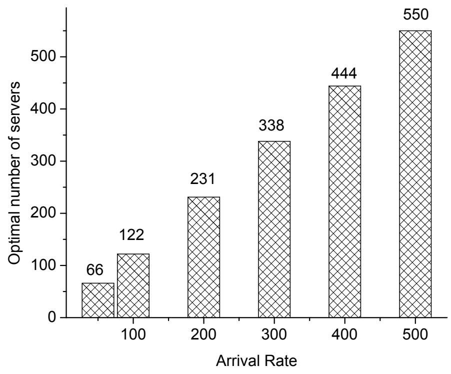

Fig. 3. Optimal Number of Servers vs. Arrival Rate of Request

## A. Single MSP Scenario

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 server under the eviction probability constraint of users. Mathematically, the optimal resource allocation for multimedia application problem can be formulated as follows.

$$\max_{m} \quad \Re(m) = p\sum_{i=1}^{m} i\Pi_i - \varepsilon(\lambda/\mu)\Pi_m - c\sum_{i=1}^{m} i\Pi_i - m\mathrm{C}, \tag{7}$$

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

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 (7) is a integer maximization problem, however, it can be effectively solved by numerical method because it has one variable $m$.

**Numerical Setting:** We perform numerical analysis to evaluate the proposed scheme with the parameters: $p$ =1\$/request, $c$ =0.2\$/server, $C$=0.1\$/server, λ is given in range of 50 to 500 requests/second, $\mu$ = 1 request/second, β is 0.5, I = 0.05.

Figure 2 shows the shape of the revenue function $\Re(m)$. We zoom in the curve of the revenue function $\Re(m)$ in the range of 200 to 300 servers in order to estimate the optimal number of server by observing. Thus, the revenue value $\Re(m)$ 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 λ of request to the optimal number of server $m^*$, we vary the value of λ 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.

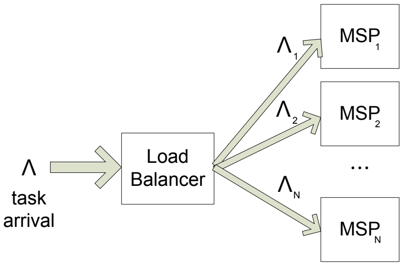

Fig. 4. Multiple media cloud service provider scenarios.

## B. Multiple MSPs Scenario

In this subsection, we study the resource allocation problem in multiple MSPs scenario which is described in Figure 4. Suppose there are N MSPs sharing a market having the multimedia task arrivals of requests follow a Poisson process with mean arrival rate Λ. We assume that there is a third party provider, named Load Balancer, has a role in collecting all multimedia task requests and then distribute it to N MSPs by probabilities. According to decomposition property, the multimedia task arrivals of each MSP requests follow a Poisson process with mean arrival rate $\lambda_1$, $\lambda_2$,..., $\lambda_N$ respectively such as $\sum_{n=1}^{N}\lambda_n \leq \Lambda$. We assume that there is some revenue sharing contract between the MSPs and the Load Balancer, therefore, we formulate the total revenue maximization problem in multiple MSPs scenarios as follows

$$\min_{\{\lambda_1,\lambda_2,...,\lambda_N\},\{m_1,m_2,...,m_N\}} \quad \sum_{n=1}^{N}\lambda_n\Re_n(\lambda_n) - \varepsilon\left(\Lambda - \sum_{n=1}^{N}\lambda_n\right), \tag{8}$$

$$\text{s.t.} \quad \Pi^n_{m_n}(\lambda_n) \leq \mathrm{I}, \forall n = 1,...,N,$$
$$\sum_{n=1}^{N}\lambda_n \leq \Lambda,$$

where $m_n$ is the number servers of the MSP $n$. Parameter ε is the eviction payment of a evicted task request from the Load Balancer in case of there are too many requests to Load Balancer such that all MSPs together cannot server. However, Problem (8) is difficult due to the complexity of revenue $\Re(m)$ and the eviction probability $\Pi^n_{m_n}(\lambda_n)$ as functions of the number of server as described in (5) and (6), respectively. Thus, in this work, we assume that the number of server $m_n$ of the MSP $n$ is given and fixed. This assumption is reasonable

TABLE I
OPTIMAL ARRIVAL RATE IN A SCENARIO $m_1$ = 50, $m_2$ = 100, FOUR CASES: (I) $\Lambda$ = 50; (II) $\Lambda$ = 100; (III) $\Lambda$ = 150; (IV) $\Lambda$ = 200.

| Cases | $\lambda_1^*$ | $\lambda_2^*$ | The maximum revenue | $\Pi_{m_1}^1(\lambda_1)$ | $\Pi_{m_2}^2(\lambda_2)$ |
|---|---|---|---|---|---|
| $\Lambda$ = 50 | 16.67 | 33.33 | 138 | 0.00 | 0.00 |
| $\Lambda$ = 100 | 44.53 | 55.47 | 1603 | 0.05 | 0.00 |
| $\Lambda$ = 150 | 44.53 | 95.24 | 2749 | 0.05 | 0.05 |
| $\Lambda$ = 200 | 44.53 | 95.24 | 2749 | 0.05 | 0.05 |

TABLE II
OPTIMAL ARRIVAL RATE IN B SCENARIO $m_1$ = 100, $m_2$ = 50, FOUR CASES: (V) $\Lambda$ = 50; (VI) $\Lambda$ = 100; (VII) $\Lambda$ = 150; (VIII) $\Lambda$ = 200.

| Cases | $\lambda_1^*$ | $\lambda_2^*$ | The maximum revenue | $\Pi_{m_1}^1(\lambda_1)$ | $\Pi_{m_2}^2(\lambda_2)$ |
|---|---|---|---|---|---|
| $\Lambda$ = 50 | 33.33 | 16.67 | 555 | 0.00 | 0.00 |
| $\Lambda$ = 100 | 83.33 | 16.67 | 4644 | 0.01 | 0.00 |
| $\Lambda$ = 150 | 95.24 | 44.53 | 6027 | 0.05 | 0.05 |
| $\Lambda$ = 200 | 95.24 | 44.53 | 6027 | 0.05 | 0.05 |

because the Load Balancer cannot have ability to choose the number of server of the MSPs. Given fixed $m_n \forall n = 1, ..., N$, we have a equivalent problem to Problem (8) as follows

$$\begin{aligned} \min_{\{\lambda_1,\lambda_2,...,\lambda_N\}} \quad & \sum_{n=1}^{N} \lambda_n \Re_n(\lambda_n) - \varepsilon\left(\Lambda - \sum_{n=1}^{N} \lambda_n\right), \qquad (9) \\ \text{s.t.} \quad & \Pi_{m_n}^n(\lambda_n) \leq \text{I}, \forall n = 1, ..., N \\ & \textstyle\sum_{n=1}^{N} \lambda_n \leq \Lambda, \end{aligned}$$

**Numerical Setting:** The parameters are set as: $p_1 = 1$, $p_2 = 0.5$ \$/request, $c_1 = c_2 = 0.2$ \$/server, $C_1 = C_2 = 0.1$ \$/server, $\Lambda$ is given in range of 50 to 200 requests/second, $\mu_1 = \mu_2$ = 1 request/second, $\beta_1 = \beta_2 = 0.5$, $\varepsilon = 0.5$ and I = 0.05.

Table I shows that when the total arrival of request $\Lambda$ is small value as 50 and 100, the sum $\lambda_1^* + \lambda_2^*$ is equal to $\Lambda$. However, when there are too many request such as 150 and 200, due to the probability of eviction constraint the optimal arrival rate sum $\lambda_1^* + \lambda_2^*$ in (iii) and (iv) are both less than $\Lambda$. The reason that $\lambda_1^*$ less than $\lambda_2^*$ in all four cases is $m_1 = 50$ less than $m_2 = 100$. In vice versa, $\lambda_1^*$ is greater than $\lambda_2^*$ in all four cases when $m_1 = 100$ is greater than $m_2 = 50$ as shown in Talbe II. The total maximum revenue in B scenario is larger than that in A scenario because the price value $p_1 = 1.0$ is greater than price value $p_2 = 0.5$ in both scenarios and the number of server $m_1$ and the optimal arrival request of MSP 1 in B scenario are both greater than that in A scenario.

## IV. CONCLUSION

In this paper, we have studied the resource optimization problem in multimedia cloud to provide services to obtain the maximum revenue of MSP. We use the queuing model to capture the relationship between arrival task request, eviction of users and the number of servers. For the future works, the interactions between MSPs become an interesting challenge for us. We will investigate the competitive and cooperative behaviors between MSPs by using game theory.

## V. ACKNOWLEDGEMENT

This work was supported by the ICT R&D program of MSIP/IITP, Republic of Korea. (2014-044-011-003, Open control based on distributed mobile core network). 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] L. Wang, G. Von Laszewski, A. Younge, X. He, M. Kunze, J. Tao, and C. Fu, "Cloud computing: a perspective study," *New Generation Computing*, vol. 28, no. 2, pp. 137–146, 2010.

[3] 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.

[4] 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.

[5] C. Do, N. Tran, M. V. Nguyen, C. seon Hong, and S. Lee, "Social Optimization Strategy in Unobserved Queueing Systems in Cognitive Radio Networks," *Communications Letters, IEEE*, vol. 16, no. 12, pp. 1944–1947, 2012.

[6] N. H. Tran, C. S. Hong, S. Lee, and Z. Han, "Optimal Pricing Effect on Equilibrium Behaviors of Delay-Sensitive Users in Cognitive Radio Networks," *Selected Areas in Communications, IEEE Journal on*, vol. 31, no. 11, pp. 2266–2579, 2013.

[7] C. T. Do, N. H. Tran, C. S. Hong, and S. Lee, "Finding an Individual Optimal Threshold of Queue Length in Hybrid Overlay/Underlay Spectrum Access in Cognitive Radio Networks," *IEICE Transactions on Communications*, vol. 95, pp. 1978–1981, 2012.

[8] C. Do, N. Tran, Z. Han, L. Le, S. Lee, and C. S. Hong, "Optimal Pricing for Duopoly in Cognitive Radio Networks: Cooperate or Not Cooperate?" *Wireless Communications, IEEE Transactions on*, 2014, in press.

[9] 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.

[10] K. Xiong and H. Perros, "Service performance and analysis in cloud computing," in *Services-I, 2009 World Conference on*. IEEE, 2009, pp. 693–700.

[11] H. Khazaei, J. Misic, and V. B. Misic, "Performance analysis of cloud computing centers using m/g/m/m+ r queuing systems," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 23, no. 5, pp. 936–943, 2012.

[12] B. Yang, F. Tan, Y.-S. Dai, and S. Guo, "Performance evaluation of cloud service considering fault recovery," in *Cloud Computing*. Springer, 2009, pp. 571–576.

[13] J. M. Smith, "M/g/c/k blocking probability models and system performance," *Performance Evaluation*, vol. 52, no. 4, pp. 237–267, 2003.

[14] P. Y. Lau, S. Park, J. Yoon, and J. Lee, "Pay-as-you-use on-demand cloud service: An iptv case," in *Electronics and Information Engineering (ICEIE), 2010 International Conference On*, vol. 1. IEEE, 2010, pp. V1–272.

[15] Z. Ye, X. Chen, and Z. Li, "Video based mobile location search with large set of sift points in cloud," in *Proceedings of the 2010 ACM multimedia workshop on Mobile cloud media computing*. ACM, 2010, pp. 25–30.

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

[17] B. C. ROBERT, "Introduction to queueing theory," 1981.