TOWARDS OPTIMAL RESOURCE ALLOCATION FOR DIFFERENTIATED MULTIMEDIA SERVICES IN CLOUD COMPUTING ENVIRONMENT

Xiaoming Nan, Yifeng He, and Ling Guan

Ryerson University, Toronto, Canada

ABSTRACT

Cloud-based multimedia services have been widely used in recent years. As the growing scale, users often have quite diverse quality of service (QoS) expectations. A key challenge for differentiated services is how to optimally allocate cloud resources to satisfy different users. In this paper, we study resource allocation problems for differentiated multimedia services. We first propose a queueing model to characterize differentiated services in cloud. Based on the model, we optimize cloud resources in the first-come first-served (FCFS) scenario and priority scenario. In each scenario, we formulate and solve the optimal resource allocation problem to minimize resource cost under response time constraints. We conduct extensive simulations with practical parameters of Amazon EC2. Simulation results demonstrate that the proposed resource allocation schemes can optimally configure resources to provide satisfactory services at the minimal resource cost.

1. INTRODUCTION

In recent years, we have witnessed the rapid development of cloud computing. As the emerging computing paradigm, cloud computing manages a shared pool of servers to provide on-demand computation resources. Due to the elastic and on-demand natures of resource provisioning, cloud computing can effectively satisfy the intensive computation demands for multimedia processing. Therefore, an increasing number of multimedia services have been migrated to cloud side. As an example, Netflix, a major media streaming provider in North America, moved its streaming services to Amazon Web Services (AWS) public cloud to deal with the ever-growing customers [1].

In cloud-based multimedia services, the *multimedia service provider* (MSP) rents a bunch of virtual servers and supplies applications as services. As customers, users can enjoy the interested services by sending requests to cloud. Due to the delay sensitive characteristic, the *response time* is commonly used as a critical quality of service (QoS) metric, which measures the duration from the time when the user's request arrives at cloud to the time when the request is completely served. The lower response time means the faster response to user's request, leading to the higher user experience.

As the growing scale, users often have quite diverse QoS expectations. Some enterprise users rely on the cloud-based services for business, and thus they are willing to pay a higher access fee for the faster services. Meanwhile, most individual users prefer to pay as low as possible for the fundamental services. Thus, differentiated services are required to satisfy different users. In the differentiated multimedia services, each request is assigned a tag, indicating the QoS class which the request belongs to. Upon arriving at a virtual server, the request will be processed according to its assigned class. Typically, a request with a higher QoS class requires a lower response time. Therefore, a fundamental concern for the MSP is how to optimally allocate the cloud resources such that all requests can be satisfactorily served at the minimal resource cost.

However, it is challenging to find the optimal resource allocation. *First*, there exist various multimedia services, which have heterogeneous resource demands. It is difficult to quantify the resource demands and optimally configure resources for each service. *Second*, different classes of requests have different requirements on response time, which imposes challenges for QoS provisioning. *Third*, there is a trade-off between the response time and the resource cost. If more resources are allocated to one service, the service can be sped up at a high cost. The under-provisioned resources would slow down the service, while the over-provisioned resources would lead to the unnecessary cost. It is challenging for the MSP to determine the optimal resource allocation, which can satisfy response time requirements for all classes of requests at a minimal resource cost.

To address the aforementioned challenges, we study resource allocation problems for differentiated multimedia services. Our contributions can be summarized as follows. We first propose a queueing model to characterize service process in cloud. Based on the queueing model, we optimize cloud resources in two different scenarios: first-come first-served (FCFS) scenario and priority scenario. In each scenario, we formulate and solve the optimal resource allocation problem to minimize the resource cost under the response time constraints. Evaluation results demonstrate that the proposed resource allocation schemes can optimally configure resources to guarantee QoS provisioning for all requests at the minimal resource cost.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

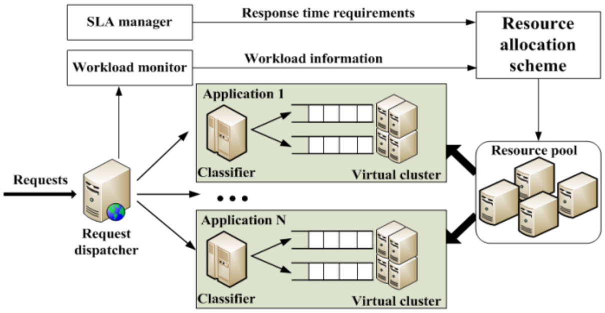


Fig. 1 An illustration of data center architecture for differentiated multimedia services

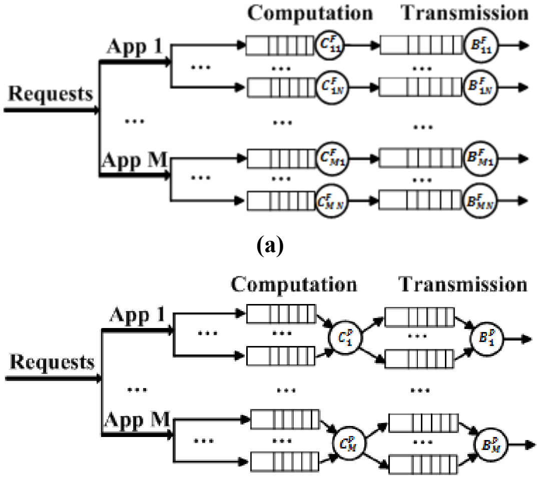
2. RELATED WORK

According to the forecast from International Data Corporation (IDC) [2], the worldwide public cloud computing services will edge towards $100 billion by 2016 and enjoy an annual growth rate of 26.4%, which is five times the traditional IT industry. With the advance in cloud, there is an upsurge of research interests in resource allocation for cloud-based multimedia services. Zhu *et al.* [3] proposed the concept of multimedia cloud. Wu *et al.* [4] presented a cloud-based video-on-demand (VoD) system, in which they used a Jackson Network to characterize the dynamic viewing behaviors. Wang *et al.* [5] presented a framework for cloud assisted live media streaming to adaptively adjust resources according to dynamic demands. Nan *et al.* [6] studied the optimal cloud resource allocation in priority service scheme. Hui *et al.* [7] presented a load-balancing technique for cloud-based multimedia systems to allocate resources in the shortest time. Compared to existing works, our study is different in the following senses: 1) we optimize cloud resources for differentiated multimedia services; 2) we use queueing model to characterize the differentiated services in cloud; 3) we minimize the resource cost in the FCFS scenario and priority scenario, respectively.

3. SYSTEM MODELS

3.1. Data Center Architecture

Fig. 1 illustrates the data center architecture for differentiated multimedia services. When requests arrive at data center, the dispatcher will schedule requests to the corresponding application, in which a classifier assigns a tag for each request to indicate the QoS class. Requests will be queued and processed by the virtual cluster. Privileged requests demand lower response time than normal requests. In the data center, the workload monitor performs a live monitoring on time-varying load, and the service level agreement (SLA) manger determines the upper bound of response time. Given the workload information and response time requirements, the proposed resource allocation scheme is applied to configure resources for each application such that all requests can be processed at the minimal resource cost.



(a)

(b)

Fig. 2 Proposed queueing model: (a) FCFS scenario, and (b) priority scenario.

In this paper, we study two different service scenarios: the *FCFS scenario*, in which each class of requests are processed in a FCFS order by the separated cluster, and the *priority scenario*, in which requests are served by the shared cluster but privileged requests have pre-emptive priority to receive service. The allocated cloud resources include computation resources and bandwidth resources, which are major resources provided by most cloud vendors like Amazon EC2 [8] and Windows Azure [9].

3.2. Queueing Model

We propose queueing models to represent service processes in FCFS scenario and priority scenario, which are illustrated in Fig. 2(a) and Fig. 2(b), respectively. In the proposed queueing models, the allocated cloud resources are represented by the servers in each queue. The *waiting time* in the queue corresponds to the waiting time of a request for available resources, while the *service time* represents the processing time of the request. As a result, the response time is the sum of waiting time and service time.

Let M be the number of applications in cloud. For each application, suppose that there are N classes of differentiated services provided by the MSP. A smaller class number corresponds to a higher privilege. Since two consecutive requests may be sent from two different users, the number of requests occur in non-overlapping intervals are independent random variables. Therefore, the request arrivals can be modeled as a Poisson process [10]. The average arrival rate of class-j ($j = 1, \ldots, N$) requests in type-i ($i = 1, \ldots, M$) application is denoted as λ_{ij}.

In FCFS scenario, each class of requests are computed or transmitted by a specific cluster. For type-i application, the computation time of a class-j request is assumed to be exponentially distributed with an average of $1/C_{ij}^F$, where the service rate C_{ij}^F represents the allocated computation resource at the virtual cluster. Similar assumptions can be found in [4]. The average transmission time is denoted as $1/B_{ij}^F$, in which B_{ij}^F is the allocated bandwidth resource. According to queuing theory, the service process in FCFS scenario can therefore be modeled as two concatenated M/M/1/∞/FCFS queueing systems [10].

In priority scenario, clusters are shared to process requests with different classes, and privileged requests have pre-emptive priority for service. For type-i application, the computation time follows the exponential distribution with an average of $1/C_i^P$, where C_i^P denotes the computation rate at the cluster. During service, class-$(j+1)$ requests can be processed only after all class-j requests have left the queue. In transmission phase, the average transmission time is denoted as $1/B_i^P$, where B_i^P is the transmission rate at the cluster. Therefore, the service in priority scenario can be modeled as two concatenated M/M/1/∞/PR queueing systems [10].

Based on the proposed queueing models, we will study the resource optimization problems to determine the optimal values of C_{ij}^F, B_{ij}^F ($\forall i = 1, \cdots, M, j = 1, \dots, N$) in FCFS scenario and C_i^P, B_i^P ($\forall i = 1, \cdots, M$) in priority scenario, respectively.

4. RESOURCE OPTIMIZATION FOR DIFFERENTIATED MULTIMEDIA SERVICES

4.1 Resource Optimization in FCFS Scenario

We first study resource optimization in FCFS scenario. Our objective is to provide satisfactory services at the minimal resource cost. The total resource cost can be formulated as $R_{tot}^F = \left(\sum_{i=1}^M \sum_{j=1}^N \theta_i C_{ij}^F + \sum_{i=1}^M \sum_{j=1}^N \rho_i B_{ij}^F\right)t$, in which θ_i and ρ_i are costs for processing and transmitting one request respectively, and t is the charge period, which is set as one hour in Amazon EC2 [8]. We take the response time as QoS metric. As presented in Sec. 3.2, the total response time is the sum of response time in computation phase and transmission phase. According to queueing theory [10], the response time for processing class-j requests in type-i application is given by $T_{ij}^{F_{com}} = 1/(C_{ij}^F - \lambda_{ij})$. To enable the queue stable, the processing rate should be higher than the request incoming rate, i.e. constraints $C_{ij}^F > \lambda_{ij}$ should be satisfied. The response time in the transmission queue is given by $T_{ij}^{F_{tra}} = 1/(B_{ij}^F - \lambda_{ij})$. To avoid local congestion, the allocated bandwidth resource must satisfy transmission demands, which is represented as $B_{ij}^F > \lambda_{ij}$. Therefore, the response time can be formulated as $T_{ij}^F = T_{ij}^{F_{com}} + T_{ij}^{F_{tra}}$, ($i = 1, \cdots, M, j = 1, \dots, N$). Based on above analysis, we can formulate the optimal resource allocation problem in FCFS scenario as follows.

$$\underset{\{C_{ij}^F, B_{ij}^F\}}{\text{Minimize}} \quad \left(\sum_{i=1}^M \sum_{j=1}^N \theta_i C_{ij}^F + \sum_{i=1}^M \sum_{j=1}^N \rho_i B_{ij}^F\right)t$$

subject to (1)

$$\begin{aligned} &\lambda_{ij} < C_{ij}^F, && i = 1,..,M, j = 1,..,N,\\ &\lambda_{ij} < B_{ij}^F, && i = 1,..,M, j = 1,..,N,\\ &\frac{1}{C_{ij}^F - \lambda_{ij}} + \frac{1}{B_{ij}^F - \lambda_{ij}} \le \tau_{ij}, && i = 1,..,M, j = 1,..,N, \end{aligned}$$

where τ_{ij} is the upper bound of response time for serving class-j requests in type-i application. The optimization problem (1) is a convex optimization, which can be solved by the Lagrange multiplier method [11]. The optimal analytical solution is given as follows.

$$\begin{aligned} C_{ij}^{F*} &= \frac{\sqrt{\theta_i} + \sqrt{\rho_i}}{\sqrt{\theta_i}\tau_{ij}} + \lambda_{ij}, \quad i = 1,..,M, j = 1,..,N,\\ B_{ij}^{F*} &= \frac{\sqrt{\theta_i} + \sqrt{\rho_i}}{\sqrt{\rho_i}\tau_{ij}} + \lambda_{ij}, \quad i = 1,..,M, j = 1,..,N. \end{aligned} \tag{2}$$

4.2 Resource Optimization in Priority Scenario

In this subsection, we will optimize cloud resources in priority scenario. The total resource cost is given by $R_{tot}^P = (\sum_{i=1}^M \theta_i C_i^P + \sum_{i=1}^M \rho_i B_i^P)t$. As analyzed in Sec. 3.2, the service in priority scenario can be viewed as two concatenated M/M/1/∞/PR queueing systems. Thus, the computation response time for class-j requests in type-i application is $T_{ij}^{P_{com}} = \frac{1/C_i^P}{1-\sigma_{i(j-1)}^{com}} + \frac{\sum_{k=1}^j (\lambda_{ik}/(C_i^P)^2)}{\left(1-\sigma_{i(j-1)}^{com}\right)\left(1-\sigma_{ij}^{com}\right)}$, where $\sigma_{ij}^{com} = \sum_{k=1}^j \frac{\lambda_{ik}}{C_i^P}$. To avoid congestion, $\sum_{j=1}^N \frac{\lambda_{ij}}{C_i^P} < 1$ should be satisfied. The response time for transmission is given by $T_{ij}^{P_{tra}} = \frac{1/B_i^P}{1-\sigma_{i(j-1)}^{tra}} + \frac{\sum_{k=1}^j (\lambda_{ik}/(B_i^P)^2)}{\left(1-\sigma_{i(j-1)}^{tra}\right)\left(1-\sigma_{ij}^{tra}\right)}$, where $\sigma_{ij}^{tra} = \sum_{k=1}^j \frac{\lambda_{ik}}{B_i^P}$. Constraints $\sum_{j=1}^N \frac{\lambda_{ij}}{B_i^P} < 1$ are required to make sure results can be successfully sent back to users. Therefore, the response time in priority scenario is formulated as $T_{ij}^P = T_{ij}^{P_{com}} + T_{ij}^{P_{tra}}$. Based on the above analysis, the optimal resource allocation problem in priority scenario can be formulated as follows.

$$\underset{\{C_{ij}^F, B_{ij}^F\}}{\text{Minimize}} \quad \left(\sum_{i=1}^M \theta_i C_i^P + \sum_{i=1}^M \rho_i B_i^P\right)t$$

subject to (3)

$$\begin{aligned} &\sum_{j=1}^N \frac{\lambda_{ij}}{C_i^P} < 1, && i = 1,..,M,\\ &\sum_{j=1}^N \frac{\lambda_{ij}}{B_i^P} < 1, && i = 1,..,M,\\ &T_{ij}^{P_{com}} = \frac{1/C_i^P}{1-\sigma_{i(j-1)}^{com}} + \frac{\sum_{k=1}^j (\lambda_{ik}/(C_i^P)^2)}{\left(1-\sigma_{i(j-1)}^{com}\right)\left(1-\sigma_{ij}^{com}\right)}, && i = 1,..,M, j = 1,..,N,\\ &T_{ij}^{P_{tra}} = \frac{1/B_i^P}{1-\sigma_{i(j-1)}^{tra}} + \frac{\sum_{k=1}^j (\lambda_{ik}/(B_i^P)^2)}{\left(1-\sigma_{i(j-1)}^{tra}\right)\left(1-\sigma_{ij}^{tra}\right)}, && i = 1,..,M, j = 1,..,N,\\ &T_{ij}^P = T_{ij}^{P_{com}} + T_{ij}^{P_{tra}} \le \tau_{ij}, && i = 1,..,M, j = 1,..,N, \end{aligned}$$

where τ_{ij} is response time upper bound for class-j requests in type-i application. The optimization problem (3) is a convex optimization problem [11], which can be solved efficiently using the primal-dual interior-point methods [11].

5. PERFORMANCE EVALUATION

We perform simulations to evaluate the proposed resource allocation schemes. Amazon EC2 [8] is one of the biggest public cloud computing platform, which allows MSP to rent virtual servers for different services. To make our evaluation convincible, we use price rates and resource configurations of Amazon EC2 in our simulations. We consider two types of multimedia applications, the free-viewpoint video (FVV) and the transcoding. The processing time of these two applications are provided by Amazon EC2 Performance Benchmark [12]. In each application, there are two classes of requests, privileged requests and normal requests. Privileged requests demand lower response time. Currently, the state-of-the-art resource allocation scheme is based on utilization threshold [13], which has been applied in Amazon AWS Elastic Beanstalk [14]. In the utilization scheme, the resource provisioning will be triggered when the utilization is higher than δ_h and stopped until utilization is lower or equal to δ_l ($\delta_l < \delta_h$). In our simulations, we will compare the proposed resource allocation schemes with the medium utilization scheme with (δ_l, δ_h) as (0.4, 0.5) and the heavy utilization scheme with (δ_l, δ_h) as (0.7, 0.8).

We emulate the requests of user swarms over a 12-hour period, and the arrivals of requests are presented in Fig. 3. From Fig. 3, we can find that the request arrival rate in each class varies dynamically over time scales. We first evaluate the performance of the proposed resource allocation scheme in FCFS scenario. Fig. 4 shows the comparison of resource cost between the proposed resource allocation scheme and the utilization schemes [13] in FCFS scenario. We can see that the proposed resource allocation scheme can achieve a lower resource cost. The reason is as follows. The utilization scheme [13] only guarantees the resource utilization boundaries, but fails to consider the diverse requirements on response time. Thus, if privileged requests and normal requests have the same arrival rate, the utilization scheme will allocate the same amount of resources, leading to the unnecessary cost. Next, we compare the resource cost between the proposed resource allocation scheme and the utilization schemes in priority scenario, which is shown in Fig. 5. Observing Fig. 5, we can find that the proposed resource allocation scheme can provide services at a lower resource cost compared to the utilization schemes. Moreover, we compare Fig. 4 and Fig. 5 and find that the resource cost in priority scenario is lower than that in FCFS scenario. The reason is that resources in priority scenario are concentrated to serve requests, which enables better resource utilization than the isolated resources in FCFS scenario. As a result, the MSP should employ priority service strategy when deploying differentiated multimedia services in a practical cloud environment.

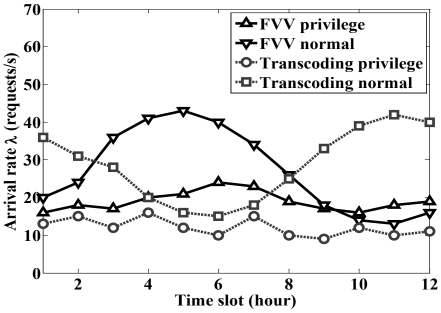


Fig. 3 Request arrival rate for each class of service in cloud over a 12-hour period

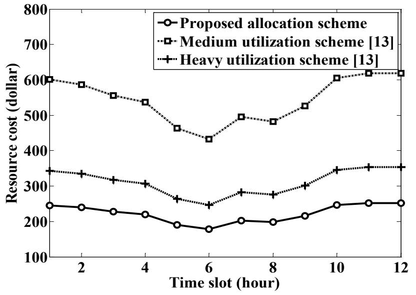


Fig. 4 Comparison of resource cost in FCFS scenario over the 12-hour period

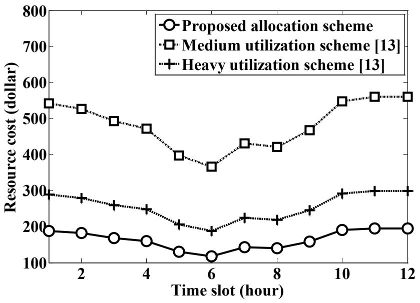


Fig. 5 Comparison of resource cost in priority scenario over the 12-hour period

6. CONCLUSIONS

In this paper, we study resource allocation problems for differentiated multimedia services. We first propose a queueing model to characterize the service process in cloud. Based on the proposed queuing model, we investigate resource allocation in FCFS scenario and priority scenario, respectively. In each scenario, we formulate and solve the optimal resource allocation problem to minimize resource cost under the response time constraints. Simulation results demonstrate that the proposed resource allocation schemes can optimally utilize cloud resources to provide satisfactory services for different classes of requests at the minimal resource cost.

ACKNOWLEDGEMENTS

This work was supported in part by the Canada Research Chair Program, NSERC Discovery Grant, and NSFC Grant 61210005.

7. REFERENCES

[1] Four Reasons We Choose Amazon's Cloud as Our Computing Platform,[Online]. Available by October, 2013: http://techblog.netflix.com/2010/12/four-reasons-we-choose-amazons-cloud-as.html.

[2] Worldwide and regional public cloud services 2012-2016 forecast. [Online]. Available by October 2013: http://www.idc.com/.

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

[4] Y. Wu, C. Wu, B. Li, X. Qiu, and F. Lau, "Cloudmedia: When cloud on demand meets video on demand," in *Proc. of International Conference on Distributed Computing Systems (ICDCS)*, 2011, pp. 268–277.

[5] F. Wang, J. Liu, and M. Chen, "Calms: Cloud-assisted live media streaming for globalized demands with time/region diversities," in *Proc. of IEEE INFOCOM*, 2012, pp. 199–207.

[6] X. Nan, Y. He, and L. Guan, "Optimal resource allocation for multimedia cloud in priority service scheme," in *Proc. of IEEE International Symposium on Circuits and Systems*, 2012, pp.143-146.

[7] H. Wen, Z. Hai-ying, L. Chuang, and Y. Yang, "Effective load balancing for cloud-based multimedia system," in *Proc. of IEEE International Conference on Electronic and Mechanical Engineering and Information Technology*, 2011, pp. 165–168.

[8] Amazon Elastic Compute Cloud. [Online]. Available by November 2013: http://aws.amazon.com/ec2/

[9] Microsoft windows azure. [Online]. Available by November 2013: http://www.microsoft.com/windowsazure/

[10] D. Gross, Fundamentals of queueing theory. Wiley-India, 2008.

[11] S. Boyd and L. Vandenberghe, Convex optimization. Cambridge University Press, 2004.

[12] Amazon EC2 Performance Benchmark, [Online], Available by November 2013: http://www.phoronix.com/scan.php?page=article&item=amazon_ec2_sep13&num=1

[13] A. Wolke and G. Meixner, "Twospot: A cloud platform for scaling out web applications dynamically," *Springer Journal of Towards a Service-Based Internet*, pp. 13–24, 2010.

[14] AWS Elastic Beanstalk. [Online]. Available by November 2013: http://aws.amazon.com/elasticbeanstalk/