International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Dynamic Ask for Redirection and Asset Provisioning for Cloud-Based Video Administrations under Heterogeneous Condition

Nandini BK[1], Veerappa BN[2]

[1] PG Student, University BDT college of Engineering, Visveswaraya Technological University, Hadadi Road, Davangere, Karnataka, India

[2]Associate Professor and Head of the Dept CS &E, University BDT college of Engineering, Hadadi Road, Davangere, Karnataka, India

Abstract: ***Cloud computing provides a new opportunity for Video Service Providers (VSP) to running compute-intensive video applications in a cost effective manner. Under this paradigm, a VSP may rent virtual machines (VMs) from multiple geo-distributed datacenters that are close to video requestors to run their services. As user demands are difficult to predict and the prices of the VMs vary in different time and region, optimizing the number of VMs of each type rented from datacenters located in different regions in a given time frame becomes essential to achieve cost effectiveness for VSPs. Meanwhile, it is equally important to guarantee users' Quality of Experience (QoE) with rented VMs. In this paper, We formulate the problem as a stochastic optimization problem and design a Lyapunov optimization framework based online algorithm to solve it. Our method is able to minimize the long-term time average cost of renting cloud resources while maintaining the user QoE. Theoretical analysis shows that our online algorithm can produce a solution within an upper bound to the optimal solution achieved through offline computing.***

Keywords: Cloud computing, cloud-based video service, request redirection, resource provision, Lyapunov optimization

1. Introduction

The cloud computing paradigm offers a convenient way for a VSP to dynamically adjust its computing resources rented from cloud service providers (CSPs) according to the demand in a Pay-As-You-Go (PAYG) manner. Compared with traditional approaches, the cloud computing paradigm eliminates VSPs' costs of purchasing and maintaining their own infrastructures. The optimization goal of a VSP is therefore to minimize the monetary cost of renting VMs and guarantee its users' Quality of Experience (QoE) in order to maintain its competitive advantage in the market. However, it is challenging for VSPs to dynamically rent computing resources in the cloud in a cost-effective manner to provide users with adequate level of QoE.

Firstly, the user request arrivals are dynamic and bursty user demands are difficult to predict. With different QoE requirements associated with these user requests, it is difficult to find an optimal way to map them to a variety of resource types in the cloud. Secondly, balancing the cost of cloud resource renting and QoE of users is a difficult decision making problem itself, e.g., higher QoE may cost a VSP more in short term but reward it in long term. Thirdly, a single CSP may not have servers located in geographically different regions that sufficiently cover the users of a VSP. In this case, the VSP may need to use multiple CSPs with different geographically located servers to provide satisfactory QOE to its users. The difference in CSPs' resource pricing in different regions and time slots further complicates the resource renting and user request scheduling for VSPs.

There are some existing works in this area. Most of them consider the resource renting and request scheduling problem separately. For example, [6], [7] deal with the resource provisioning problem by optimizing the cost of renting computing resources from the cloud. They assume request arrival time and service time follow certain distributions. Some work focuses on finding optimal request dispatching strategies [8].

The use of CDNs often requires the negotiation of contracts and incurs a relatively high setup cost. P2P systems require minimal dedicated infrastructure for video content delivery but suffer from problems such as long video start-up delay caused by excessively video data prefetching in a unstable environment. Cloud datacenters provide a dedicated infrastructure as well as a convenient Pay-As-You-Go model of running video services on them, which makes them increasingly popular for video content delivery.

2. Proposed System

Our goal is to give an optimal cloud resource renting and user request scheduling strategy to deal with these challenges. The strategy intends to minimize the long term VM renting cost of resources from multiple CSPs for a VSP while maintaining certain level of user QoE. To achieve this goal, we first formulate the problem into a jointly stochastic optimization problem, and then, apply the Lyapunov Optimization framework to solve the problem. Such a stochastic system does not require predicting the future system states and makes decisions only based on current system state [11]. Based on drift-plus-penalty function transformation, we propose an online algorithm that is able to schedule user requests from multiple regions to distributed datacenters and dynamically compute the near optimal number needed to satisfy user requirements for serving their workloads. The major contributions of this work are summarized as below:

Licensed Under Creative Commons Attribution CC BY

- We propose an algorithm to solve the jointed stochastic problem to balance the cost saving and QoE using Lyapunov optimization framework. The algorithm approximates the optimal solution within provable bounds. Moreover, the algorithm can deal with the case that rental period is various over different datacenters and have a distributed implementation.
- We evaluate the algorithm using both real and synthetic datasets. Our extensive experiments show its effectiveness. Furthermore, the experiments also reveal that the heterogeneity of QoE requirements provides an opportunity to reduce the operational cost of VSPs.

3. Literature Survey

Weiwen Zhang, Yonggang Wen, Member, IEEE, Jianfei Cai, Senior Member, IEEE, and Dapeng Oliver Wu, Fellow, IEEE. In this paper, we investigate energy-efficient job dispatching algorithm for transcoding as a service (TaaS) in a multimedia cloud. We aim to minimize the energy consumption of service engines in the cloud while achieving low delay for TaaS. We formulate the job dispatching problem as a constrained optimization problem under the framework of Lyapunov optimization. Using the drift-plus-penalty function, we propose an online algorithm that dispatches the transcoding jobs to service engines, with an objective to Reduce Energy consumption while achieving the QUEue STability (REQUEST).

Changick Kim, Member, IEEE, and Jenq-Neng Hwang, Fellow, IEEE. Key frames are the subset of still images which best represent the content of a video sequence in an abstracted manner. In other words, video abstraction transforms an entire video clip to a small number of representative images. In this paper, we present a scheme for object-based video abstraction facilitated by an efficient video-object segmentation (VOS) system

Yu Wu∗, Chuan Wu∗, Bo Li†, Xuanjia Qiu∗, Francis C.M. Lau∗.Internet-based cloud computing is a new computing paradigm aiming to provide agile and scalable resource access in a utility-like fashion. Other than being an ideal platform for computation-intensive tasks, clouds are believed to be also suitable to support large-scale applications with periods of flash crowds by providing elastic amounts of bandwidth and other resources on the fly. The fundamental question is how to configure the cloud utility to meet the highly dynamic demands of such applications at a modest cost.

Chun-Cheng Lin, Member, IEEE, Hui-Hsin Chin, Student Member, IEEE, Der-Jiunn Deng, Member, IEEE. Consider a centralized hierarchical cloud-based multimedia system (CMS) consisting of a resource manager, cluster heads, and server clusters, in which the resource manager assigns clients' requests for multimedia service tasks to server clusters according to the task characteristics, and then each cluster head distributes the assigned task to the servers within its server cluster. For such a complicated CMS, however, it is a research challenge to design an effective load balancing algorithm which spreads the multimedia service task load on servers with the minimal cost for transmitting multimedia data between server clusters and clients, while the maximal load limit of each server cluster is not violated.

4. System Design

We consider such a system scenario: datacenters belong to multiple CSPs that are geographically distributed over Several locations, and run various types of services. Users from different regions can obtain the services from any data center at any time. The system architecture is illustrated in Fig. 1. In the system, users from different regions obtain various of services like video streaming and transcoding from VSPs which do not possess their own datacenters but actually rent the infrastructure (VMs) from CSPs. Once the VSP receives a request, the request should be dynamically redirected to an optimal datacenter according to its QoE requirements and the execution cost, considering the different prices of datacenters over different regions.

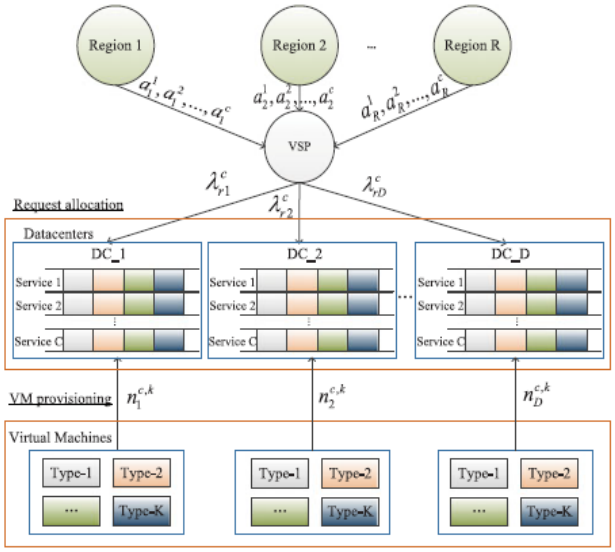


Figure 1: System Architecture

Formally, considering the geo-distributed datacenters set D with size of D=|D| , indexed by d. Each datacenter provides C classes of services denoted by set C (i.e. C=|C|), indexed by c. And a set K of distinct types of VMs, each with specific capacity under different configurations of CPU, memory and storage, are provided in each datacenter. Requests are dynamically generated by users from R=|R| different regions, denoted as set R.

5. Proposed Algorithm

In response to the challenges of problem P1, we take advantage of Lyapunov optimization techniques [11] to design an online control framework, which is able to concurrently make request redirection and resource procurement decision. In particular, our control algorithm does not require future information about user requests, which also can be proved to approach a time averaged cost that is arbitrarily close to optimum, while still maintaining system stability.

Licensed Under Creative Commons Attribution CC BY

Table 1: Important Notations

$\mathcal{D}$	set of datacenters distributed over multiple regions
$\mathcal{C}$	set of all services classes
$\mathcal{R}$	set of user regions
$\mathcal{K}$	set of VM types
m	time interval to decide resource provisioning
ρ_d^k	the availability of the type-k VM in datacenter-d
ω_c	workload of type-c request
W_{max}	max workload of each type request
ℓ_c	tolerable delay of type-c service
$a_r^c(t)$	number of the requests of type-c from region r at t
$\lambda_{rd}^c(t)$	number of requests of type-c allocated to d in region r at t
N_d^k	number of VMs of type-k in datacenter d
N_{max}	max number of VMs of each type over all datacenters
$A_{rc}^{\max}$	max number of request for type-c in region-r
$n_d^{c,k}(t)$	number of type-k VM for type-c request in d at t
$p_d^k(t)$	price to provision a type-k VM in d at t
s_k	compute capacity of type-k VM
Q_0	the minimal QoE level should be guaranteed for users
Q_{max}	the max QoE level users can achieve
$H_d^c(t)$	unprocessed workload of type-c request in d at t
$Q_d^c(t)$	Virtual queue to satisfy the constraint (11)

5.1 Lyapunov Optimization

According to the standard optimization framework theory[11], to minimize the time-averaged objective function, we the original stochastic optimization problem into a problem of minimizing the Lyapunov drift-plus-penalty.

$$H_d^c(t+1) = \max\left[H_d^c(t) - \sum_{k\in\mathcal{K}} \rho_d^k n_d^{c,k}(t) s_k, 0\right] + \sum_{r\in\mathcal{R}} \lambda_{rd}^c(t) w_c.$$

Which means that VM fault availability of type k VM fault tolerance are considered in the model.

5.2 Online control Algorithm

Fortunately, a careful investigation of the R.H.S of inequality (19) reveals that the optimization problem can be equivalently decoupled into two sub problems: 1) request redirection and 2) resource procurement. The details of solving the two sub problems are presented as follows.

1. Request redirection: By observing the relationship among variables, the part related to request redirection can be extracted from the R.H.S of as:

$$\mathbb{E}\left\{\sum_{\tau=t}^{t+m-1}\sum_{d\in\mathcal{D}}\sum_{c\in\mathcal{C}}\left\{\sum_{r\in\mathcal{R}}\lambda_{rd}^c(\tau)(H_d^c(t)w_c + Q_d^c(t)A_c d_{rd})\right\}|\Theta(t)\right\}.$$

Furthermore, it should be noted that requests of each type generated from each region are independent. The centralized minimization can be implemented independently and distributedly.

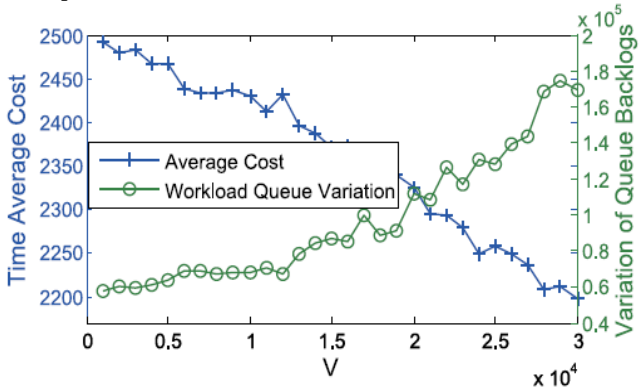
2. VM procurement: Resource procurement problems in each datacenter are independent, can be solved distributedly within each datacenters. For a single datacenter d, the resource procurement problem can be rewritten as:

$$\begin{aligned}\min\ & V\,\mathbb{E}\left\{\sum_{\tau=t}^{t+m-1}\sum_{d\in\mathcal{D}}\sum_{c\in\mathcal{C}}\sum_{k\in\mathcal{K}} n_d^{c,k}(t)p_d^k(t)|\Theta(t)\right\}\\ &+\mathbb{E}\left\{\sum_{\tau=t}^{t+m-1}\sum_{d\in\mathcal{D}}\sum_{c\in\mathcal{C}} Q_d^c(t)\Gamma_d^c(t)|\Theta(t)\right\}\\ &-\mathbb{E}\left\{\sum_{\tau=t}^{t+m-1}\sum_{d\in\mathcal{D}}\sum_{c\in\mathcal{C}} H_d^c(t)\left(\sum_{k\in\mathcal{K}}\rho_d^k n_d^{c,k}(t)s_k\right)|\Theta(t)\right\}\\ \text{s.t}\ & (10).\end{aligned}$$

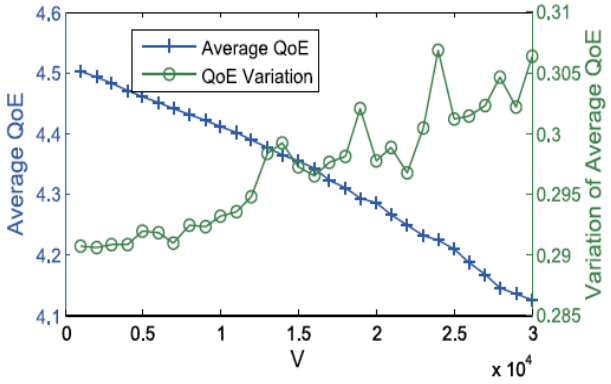
Thus, we can exploit many methods such as interior point method and gradient projection method to deal with it.

6. Result and Analysis

The proposed algorithm is able to deal with the case that the VM provisioning periods vary over different datacenters, to validate the original mathematical derivation result the following experiments mainly consider the case that the periods are the same over datacenters nonetheless, a comparison experiment is conducted to validate the capability of the proposed algorithm in dealing with different rental period for different datacenters.



(a) Impact of V on cost and workload queue



(b) Impact of V on QoE

For parameter V, as can be seen in fig (a) with the increasing of V, the time average cost obtained using our algorithm declines significantly and converges to the minimum level for a larger value of V, However, the stability of the system simultaneously decline since the variation of queue backlogs improves with the increase of V.

Furthermore, the cost reduction is achieved at the cost of

Licensed Under Creative Commons Attribution CC BY

degrading the users QoE. As can be seen in fig (b) user QoE is decreasing with the increase of parameter V. Additionally The variation of QoE is increasingly fluctuating with the increase of V, which means that increasing V degrades the stability user QoE level.

7. Conclusion

This proposed a novel method called a request redirection and resource procurement from the perspective of vsps. We showed that it is capable of reducing the cost of providing video services in the cloud and achieving satisfactory user QoE level simultaneously. The method provided an efficient way to video services in a general and heterogeneous environment consisting of dynamic user workload, dynamic resource price multiple services with heterogeneous QoE requirements, and heterogeneous QoE requirements, and heterogeneous datacenters.

References

[1] Cisco System Inc., "Cisco visual networking index: Forecast and methodology, 2012–2017," 2013.
[2] W. Zhang, Y. Wen, J. Cai, and D. Wu, "Toward transcoding as a service in a multimedia cloud: Energy-efficient job-dispatching algorithm," IEEE Trans. Veh. Technol., vol. 63, no. 5, pp. 2002–2012,
[3] B. Gunsel and A. Tekalp, "Content-based video abstraction," in Proc. Int. Conf. Image Process., Oct. 1998, pp. 128–132.
[4] S.-F. Chang and A. Vetro, "Video adaptation: Concepts, technologies, and open issues," Proc. IEEE, vol. 93, no. 1, pp. 148–158, Jan.2005.
[5] D. Miao, W. Zhu, C. Luo, and C. W. Chen, "Resource allocation for cloud-based free viewpoint video rendering for mobile phones," in Proc. 19thACMInt. Conf.Multimedia, 2011, pp. 1237–1240.
[6] Y. Wu, C.Wu, B. Li, X. Qiu, and F. C. M. Lau, "Cloudmedia: When cloud on demand meets video on demand," in Proc. 31st Int. Conf. Distrib. Comput. Syst., Jun. 2011, pp. 268–277.
[7] X. Nan, Y. He, and L. Guan, "Optimal resource allocation for multimed cloud based on queuing model," in Proc. IEEE 13th Int. Workshop Multimedia Signal Process., Oct. 2011, pp. 1–6.
[8] H. Wen, Z. Hai-ying, L. Chuang, and Y. Yang, "Effective load balancing for cloud-based multimedia system," in Proc. Int. Conf. Electron. Mech. Eng. Inf. Technol., Aug. 2011, vol. 1, pp. 165–168.
[9] Y. Yao, L. Huang, A. Sharma, L. Golubchik, and M. Neely, "Power cost reduction in distributed data centers: A two-time-scale approach for delay tolerant workloads," IEEE Trans. Parallel Distrib Syst., vol. 25, no. 1, pp. 200–211, Jan. 2014.
[10] D. Wu, Z. Xue, and J. He, "iCloudAccess: Cost-effective streamingof video games from the cloud with low latency," IEEE Trans. Circuits Syst. Video Technol., vol. 24, no. 8, pp. 1405–1416, Aug. 2014.
[11] M. Neely, Stochastic Network Optimization With Application to Communication and Queueing Systems. San Rafael, CA, USA: Morgan and Claypool, 2010.
[12] B. Cohen, "Incentives build robustness in bit torrent," in Proc.Workshop Econ Peer-to-Peer Syst., 2003, vol. 6, pp. 68–72.
[13] X. Hei, C. Liang, J. Liang, Y. Liu, and K. Ross, "A measurement study of a large-scale P2P IPTV system," IEEE Trans. Multimedia, vol. 9, no. 8, pp. 1672–1687, Dec. 2007.
[14] A. Ghodsi, M. Zaharia, B. Hindman, A. Konwinski, S. Shenker, and I. Stoica, "Dominant resource fairness: Fair allocation of multiple resource types," in Proc. 8th USENIX Conf. Netw. Syst. Des. Implementation 2011, pp. 323–336.
[15] Y. Song, Y. Sun, and W. Shi, "A two-tiered on-demand resource allocation mechanism for VM-based data centers," IEEE Trans. Services Comput., vol. 6, no. 1, pp. 116–129, Jan. 2013.
[16] S. Ren, Y. He, and F. Xu, "Provably-efficient job scheduling for energy and fairness in geographically distributed data centers," in Proc. IEEE 32nd Int. Conf. Distrib. Comput. Syst., Jun. 2012, pp. 22–31.
[17] Z. Liu, M. Lin, A. Wierman, S. H. Low, and L. L. Andrew, "Greening geographical load balancing," in Proc. ACM Joint Int. Conf. Meas. Model. Comput. Syst., 2011, pp. 233–244.
[18]] G. Lee, B.-G. Chun, and H. Katz, "Heterogeneity-aware resource allocation and scheduling in the cloud," in Proc. 3rd USENIX Conf. Hot Topics Cloud Comput., 2011, pp. 1–5.
[19] F. I. Popovici and J. Wilkes, "Profitable services in an uncertain world," in Proc. ACM/IEEE Conf. Supercomput., 2005, p. 36.
[20] J. Tang, W. P. Tay, and Y. Wen, "Dynamic request redirection andelastic service scaling in cloud-centric media networks," IEEE Trans. Multimedia, vol. 16, no. 5, pp. 1434–1445, Aug. 2014.

Licensed Under Creative Commons Attribution CC BY