

Dynamic Request Redirection and Resource Provisioning for Cloud Based Video Services under Heterogeneous Environment

Nandini BK¹ and Veerappa BN²

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

¹nandinibk06@gmail.com

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

²veerappabn@gmail.com

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.

1. INTRODUCTION

INTERNET video is both bandwidth and CPU cycle demanding. According to a report [1] of Cisco Systems Inc. in 2013, the global Internet video traffic will contribute to 69 percent of the Internet traffic in 2017, up from 57 percent in 2012. The growth of Internet video traffic is at an annual rate of 34 percent. Video data processing also demands significant amount of CPU cycles.

Video applications often involve in pre-processing steps such as transcoding [2], encoding/decoding, abstraction [3], adaption [4], rendering [5], etc. to satisfy different requirements. As an example, scenes in an online game are rendered dynamically to follow the actions of players. The difference in the screens of various devices of different players often requires different video codecs. Data processing involved in these steps is computeintensive and normally done on the Video Service Provider (VSP) side. It poses significant challenges for VSPs to efficiently plan and manage their

computing capacity in order to satisfy user requests in a timely manner, particularly when requests may have bursty arrival patterns.

2. EXISTING SYSTEM

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].

- Major platforms for video content delivery over the Internet include large content delivery networks, or CDNs, such as Akamai, P2P systems such as BitTorrent [12] and PPLive [13] and Cloud datacenters.

- 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 pre fetching 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.

3. PROPOSED SYSTEM

We propose a framework that systematically handles resource renting from multiple CSPs and schedules user requests to these resources in a nearly optimal manner. In particular, the framework is capable of handling heterogeneous types of user requests, workloads and QoE requirements. VMs in the cloud have different types and are priced dynamically.

- 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.

4. SYSTEM MODELING

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.

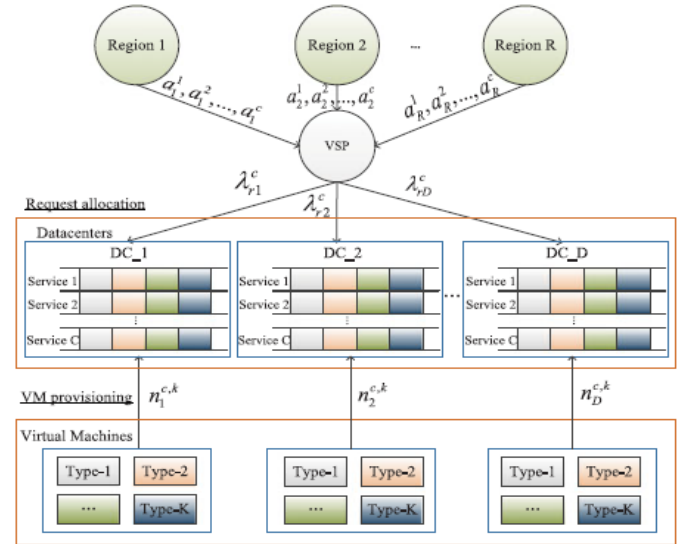


Figure 1: System Architecture.

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. PROBLEM FORMULATION

In this subsection, we formulate VM provisioning cost and user QoE respectively and define the objective.

1) VM COST. From the perspective of users, requests are dynamically generated over different regions in each time slot. the total number of the type- c service requests generated by users from the r th region at time slot t . And denote the number of the type- c service requests from region r allocated to datacenter d at time slot t , λ_{rd}^c as the max number of type- c request generated in region r . Then, we have:

$$a_r^c(t) \leq A_{rc}^{\max}, \forall r, \forall c, t \in [1, T], \quad (1)$$

$$a_r^c(t) = \sum_{d \in D} \lambda_{rd}^c(t), \forall r, \forall c, t \in [1, T]. \quad (2)$$

2) QOE DEFINITION. On the user side, QoE is a major metric to evaluate the service level. We also consider QoE factor when making the resource procurement decision. Usually, in the network systems, QoE is sensitive to both queuing delay and network delay.

Therefore, for a request k , we define its delay as follows:

$$d^k = d_{net}^k(\cdot) + d_{que}^k(\cdot), \quad (4)$$

where d_{que} and d_{net} denote the function of network delay and queue delay of the request k . In reality, the two kinds of delay is very hard to estimate due to they depends on some different factors (e.g., network delay depends on queuing delay (at routers), transmission delay and propagation delay of the routing path). For simplicity, we assume queuing delay is determined by the workload status and VM resources assigned to this workload, while network delay is mainly determined by the routing distance between clients and datacenters.

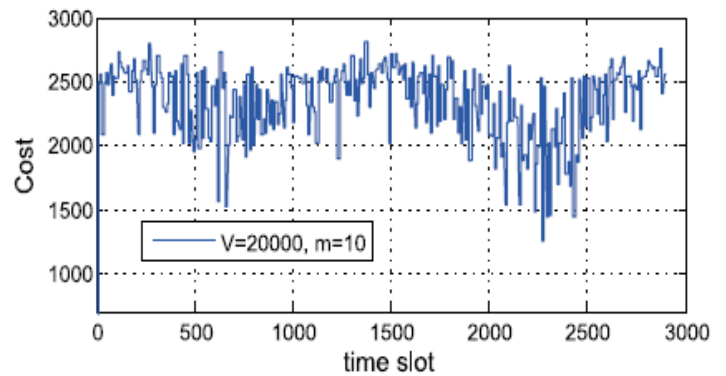
6. RESULT AND ANALYSIS

In order to facilitate the comparison, two metrics are defined in this section. (1) Cost Ratio (CR), which measures the cost proportion of a single case among the total cost obtained by all cases.

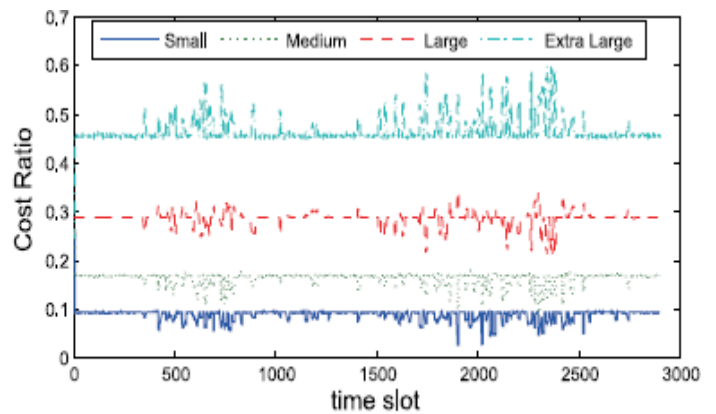
1) Effectiveness of the Algorithm. We run our dynamic algorithm for $T = 2880$ time slots, with parameter $V=20000, m=10$. Fig. a presents the cost occurred in each time slot. We observe that the monetary cost curve is fluctuating synchronously with the variation of requests as shown in Fig. which means that our algorithm can adaptively lease and adjust VMs resources to meet dynamic user demands, without forecasting the future workload information.

In detail, the cost comparison of each type of VM is illustrated in Fig. b, in which we use the metric CR for comparison. It can be observed that, under the variation of workload, the cost ratio of each VM type is relatively stable in the whole sense especially within crowd flash period. It may attribute to the fact that, within crowd period, resources are inadequate to the system and all type of the VMs will be rented to guarantee the user QoE, which cause a stable cost ratio near to the price ratio. Also the Extra Large is shown to have the highest ratio. It is due to that the more capacity of

the VM is the lower the unit price of the VM is, so that the system will prefer to rent VM with more capacity to reduce their cost. However, within the uncrowded period.



(a) Cost incurred by the system over time slots



(b) Cost ratio of each type of VM over time slots

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. Gonsel 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. 19th ACM Int. 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 multimedia 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 streaming of video games from the cloud with low latency,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 8, pp. 1405–1416, Aug. 2014.