



DYNAMIC REQUEST REDIRECTION AND RESOURCE PROVISIONING FOR CLOUD-BASED VIDEO SERVICES UNDER HETEROGENEOUS ENVIRONMENT

J. Sabitha*, P. Seline Mestica, R. Usharani*** & S. Geetha Rani******

Department of Information Technology, Dhanalakshmi College of Engineering,
Chennai, Tamilnadu

Abstract:

A Cloud computing provides a new opportunity for Video Service Providers (VSP) to running compute-intensive video applications in a cost effective manner. 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. For optimizing the number of VMs, we give a systematic method called Dynamical Request Redirection and Resource Provisioning to address this problem and formulate as 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 resource while maintaining the user QoE. Extensive experiments show that our method is adaptive to request pattern changes along time and outperforms existing algorithm.

Index Terms: Cloud Based Video Service, Lyapunov Optimization & DYRECEIVE

Introduction:

The main aim is to allocate resources for cloud based video services on user request from multiple regions to distributed data centres and dynamically computes the near and optimal virtual machine. A Cloud computing provides a new opportunity for Video Service Providers (VSP) for 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.

Related Work:

Video Service Providers (VSPs) will dynamically rent computing resources in the cloud in a cost-effective manner to provide users with adequate level of QoE. The three stochastic problems are: Firstly, the user request arrivals are dynamic and bursty. Also demands of user 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.

Methodology and Materials:

The proposed system uses 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 are of different types and are priced dynamically. We propose an algorithm to solve the jointed stochastic problem to balance the cost saving and QoE. We leverage the existence of content delivery network (CDN) to host video services on their various datacenters distributed in various regions. We give a systematic method called Dynamical Request Redirection and Resource Provisioning (DYRECEIVE) to address this problem. With our approach the video service provider is able to provide an efficient, cost effective and quality service to any number of clients.

Algorithm: DYRECEIVE

- 1 Input:
- 2 $s_k, \omega_c, W_{max}, \ell_{c \rightarrow r}, A_{max \rightarrow r}, p_{k \rightarrow d}(\tau), \rho_{k \rightarrow d}, d_{rent}, V, a, b, u, v (\forall c \in C, \forall d \in D, \forall r \in R, \forall k \in K)$;
- 3 Output:
- 4 $n_{c,k \rightarrow d}(\tau), \lambda_{c \rightarrow d}(\tau) (\forall c \in C, \forall d \in D, \forall r \in R, \forall k \in K)$;
- 5 Initialization step: Let $\tau = 0$, $st = cputime$, and set $Q_{c \rightarrow d}(0) = 0, H_{c \rightarrow d}(0) = 0, (\forall c \in C, \forall d \in D), d_{dec}(0) = 0$;
- 6 while the service of VSP is running do
- 7 calculate time slot $\tau, \tau = (curtime - st)/60s$;
- 8 estimate the decision overhead $d_{dec}(\tau)$ based on $d_{dec}(t), t \in [\tau - 5, \tau - 1]$;
- 9 Resource provisioning:
- 10 foreach datacenter $d \in D$ do
- 11 if $(\tau \bmod md) == 0$ then
- 12 Observing the queue backlogs $Q_{c \rightarrow d}(\tau), H_{c \rightarrow d}(\tau)$ and the VM price $p_{k \rightarrow d}(\tau)$ at current time;
- 13 Getting the VM provisioning strategy $(n_{c,k \rightarrow d}(\tau))$ by solving the problem (24) using CVX tool;
- 14 Request redirection:
- 15 if request arrives at system then
- 16 foreach $r \in R, c \in C$ do
- 17 Observing the queue backlogs $Q_{c \rightarrow d}(\tau), H_{c \rightarrow d}(\tau)$, the network delay d_{rd} and estimating the computation delay $d_{comp}(\tau)$ at current time;
- 18 Getting the request redirection strategy $\lambda_{c \rightarrow d}(\tau)$ by solving the problem (21) using (22);
- 19 Update the queues $Q_{c \rightarrow d}(\tau), H_{c \rightarrow d}(\tau)$ according to queue dynamic equation (12)(13) respectively.
- 20 Record the decision-making time consumed at current time slot $d_{dec}(\tau)$.

Experimental Work:

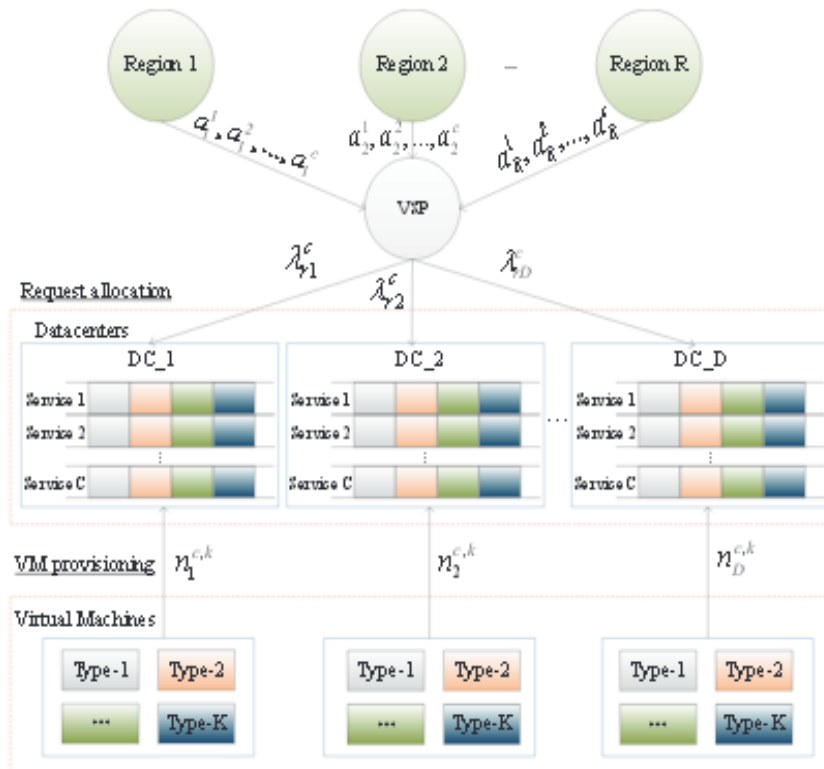
Video Service Provider:

In this module, the video service provider application is implemented. The service provider request for the cloud service provider to host their application in the cloud. The video service provider application has the various types of videos such as the high quality, medium quality and the low quality videos. The video service provides choose the cloud based on the capacity of the virtual machines.

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)

Architecture Diagram:



Cloud Service and Virtual Machine:

Request scheduling and resource allocation in the cloud can be classified based on different perspectives of cloud providers and cloud users. There are many efforts on designing Scheduling strategies for cloud providers. For single datacenters, improving resource utilization and fairness are often the focus. For multiple datacenters, some

work propose scheduling strategies to minimize the cost of electricity use through balancing load among geographically located datacenters. It 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.

Dynamic Redirection and User Response:

In this module, users from different regions obtain several 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.

Flow Diagram:

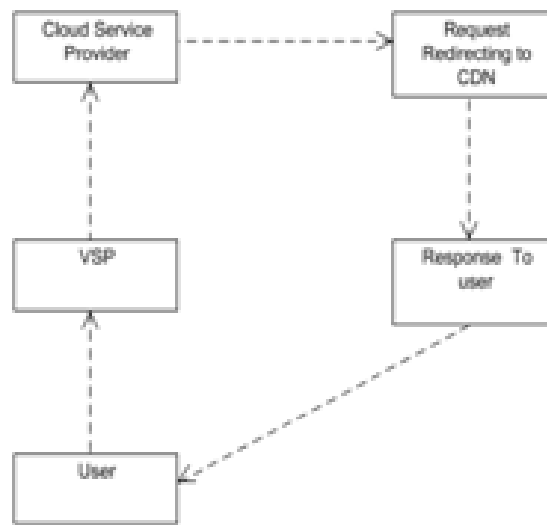


TABLE 2
 AMAZON EC2 VM INSTANCE

name	Number of compute units	price
Small	1	$BP \in [0.05, 0.07]$
Medium	2	$BP \cdot (1 + \log_{2.5}(2))$
Large	4	$BP \cdot (1 + \log_{2.5}(4))$
Extra Large	8	$BP \cdot (1 + \log_{2.5}(8))$

Result and Discussion:

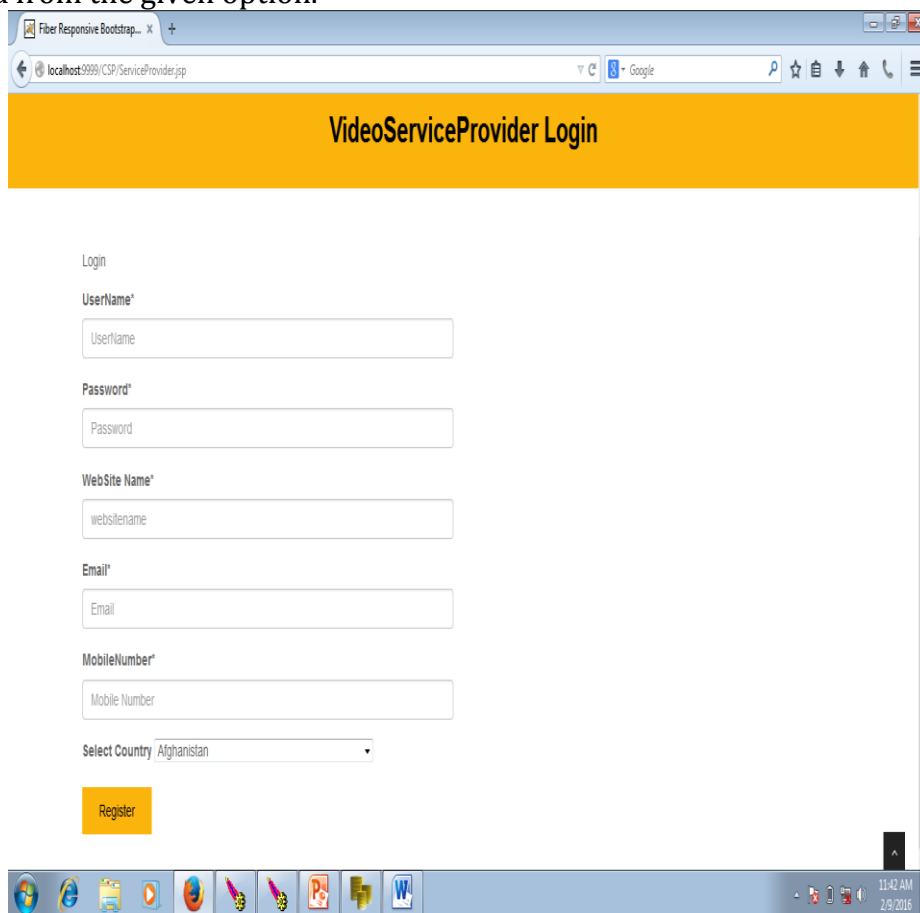
Effectiveness of the Algorithm:

We run our dynamic algorithm for $T = 2, 880$ time slots, with parameter $V = 2 \times 104, m = 10$. Present the cost occurred in each time slot. We observe that the monetary cost curve is fluctuating synchronously with the variation of requests, 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 (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.

Registration Form:

Users from different regions obtain various services like video streaming from CDN by the policy the video service provider already generated. Once the VSP receives a request, the request will be dynamically redirected to an optimal datacenter according to its QoE requirements, geographical location and the execution cost. User should login with the username, password, website name, email, mobile number and country, which is selected from the given option.

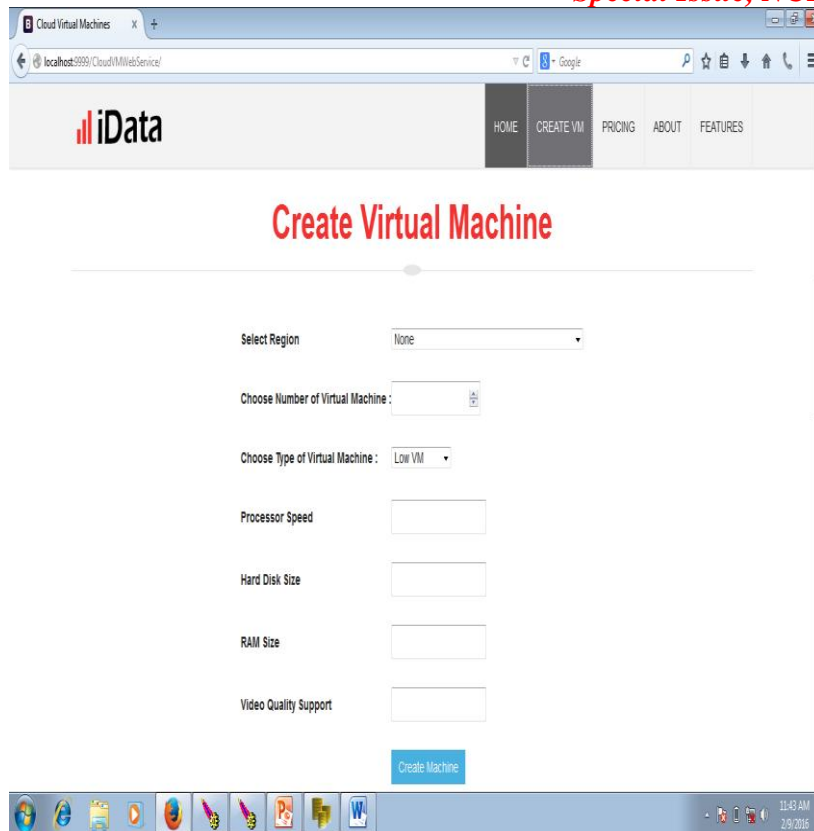


The screenshot shows a web browser window with the address bar displaying 'localhost:3099/CSP/ServiceProvider.jsp'. The page title is 'VideoServiceProvider Login'. The form contains the following fields:

- Username* (text input)
- Password* (password input)
- WebSite Name* (text input with placeholder 'websitename')
- Email* (text input with placeholder 'Email')
- MobileNumber* (text input with placeholder 'Mobile Number')
- Select Country (dropdown menu with 'Afghanistan' selected)
- Register (yellow button)

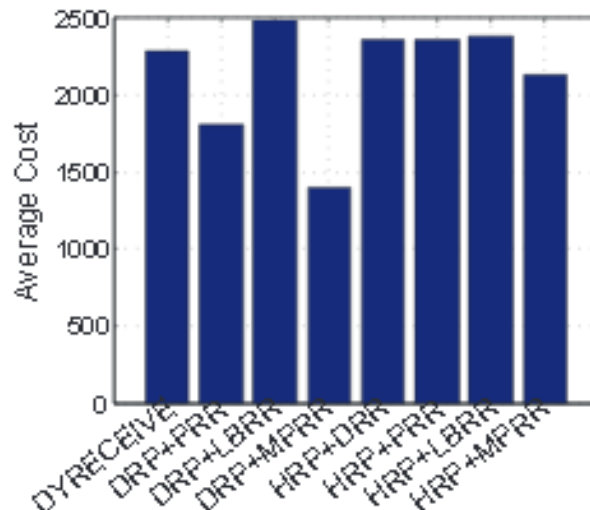
Virtual Machine Creation:

The Video service provider request for the CDN to host their application in the cloud. The video service provider application has the various type of videos such as the high quality, medium quality and the low quality videos. The video service provides choose the Virtual instances on various data centers and request the CDN to host their Services. The rent for data center usage will be calculated by CDN and offered to video service provider.



Impact of V and m:

For parameter V, as can be seen 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 declines since the variation of queue backlogs (i.e., $H_c d(\tau) + Q_c d(\tau)$) improves with the increase of V, which is consistent with Lemma 4. Therefore the parameter V control the tradeoff between the cost and user QoE in the system, which also verifies the Lemma4.



Conclusion:

Thus we allocated resources for cloud based video services on user request from multiple regions to distributed data centers and dynamically computed the near and optimal virtual machine. The video service application deployment is done on various data centers.

Future Enhancement:

In the future, this can be developed as banking transaction whether the VSP is satisfied with the bill generation process he can proceed with the banking process. The banking gateway is connected when transaction is initialized and OTP will be generated and send to VSP mail ID which he can validate it in upcoming process to complete the transaction. If the transaction is made successfully he can get access to various data center and virtual instances.

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 Transactions on Vehicular Technology*, vol. 63, no. 5, pp. 2002–2012, Jun 2014.
3. B. Günsel and A. Tekalp, "Content-based video abstraction," in *Proceedings of International Conference on Image Processing*, Oct 1998, pp. 128–132.
4. S.-F. Chang and A. Vetro, "Video adaptation: Concepts, technologies, and open issues," *Proceedings of the 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 *Proceedings of the 19th ACM International Conference on Multimedia(MM'11)*, 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 *Proceedings of the International Conference on Distributed Computing Systems(ICDCS'11)*, June 2011, pp. 268–277.
7. X. Nan, Y. He, and L. Guan, "Optimal resource allocation for multimedia cloud based on queuing model," in *Proceedings of the IEEE 13th International Workshop on Multimedia Signal Processing (MMSP'11)*, Oct 2011, pp. 1–6.
8. H. Wen, Z. Hai-, Lying. Chuang, and Y. Yang, "Effective load balancing for cloud-based multimedia system," in *Proceedings of the International Conference on Electronic and Mechanical Engineering and Information Technology(EMEIT'11)*, vol. 1, Aug 2011, pp. 165–168.