A Cost-Effective VM Offloading Scheme in Hybrid Cloud Environment

Myeongseok Hyeon(✉), Heejae Kim, and Chan-Hyun Youn

School of Electrical Engineering, KAIST, Daejeon, Korea
{tidusqoop, kim881019, chyoun}@kaist.ac.kr

Abstract. Video streaming service is offered by various content provider with cloud content delivery network like a Netflix, Youtube. In this environment, for content deliver, cache servers need to be placed properly in cost-effective manner with guaranteeing streaming performance. In this paper, as contents providers, to minimize the cost of using public cloud and to maximize performance of video streaming service, we suggest cost-effective VM offloading algorithm in hybrid cloud environment (CVOH). The CVOH considers performance degradation in internal cloud and cost for public cloud using penalty cost model and learning curve model, respectively. As the result of evaluation for CVOH, we got about twice better performance than a maximal consolidation case, and 9.1 % better than a maximal offloading case.

Keywords: Video streaming · Hybrid cloud · VM placement

1 Introduction

For offering video streaming service effectively, the architecture of video streaming service has developed into a way to place a cache server to a cloud environment for content delivery. In cases of Netflix and Youtube, content providers, they are using Amazon Web Service and Google Cloud respectively for serving their content [1, 2].

It has been reported that the great number of content consumers' request for streaming service can burst in a short time period, and also refer to needs for network resource in server to handle the peak demand which can appear frequently by Aggarwal et al. [3]. If content provider have to deal with such a dynamic demand by using their own computing resources, it is hard to estimate an amount of demand properly and it is inefficient to construct and to manage servers in respect to cost. Hence, the way using a hybrid cloud environment rises, that a private computing resource is used in dealing with universal demands and in dealing in other specific demands such a peak demand resource of the public cloud can be used.

To use cost-effectively such a hybrid cloud environment, there are two main challenges. First one is to maximize resource utilization in the internal cloud and the other one is to minimize the cost occurring by using other public cloud resource. To satisfy both objectives to guarantee QoS of streaming service being supported by the internal cloud and to minimize cost of management datacenter, content provider have to manage his computing resource properly by placing the cache server properly. The other reason is the cost occurring by using other public cloud resource increases as the

Y. Zhang et al. (Eds.): CloudComp 2015, LNICST 167, pp. 60–69, 2016.
DOI: 10.1007/978-3-319-38904-2_7

quantity of computing resource that content provider uses increases. As a related issue, Bossche et al. [4] proposes cost-optimal scheduling in hybrid IaaS clouds for deadline constrained workloads.

We present a cost-effective virtual machine offloading algorithm in hybrid cloud environment (CVOH), the optimization problem to find a cost-effective solution of virtual machine (VM) placement as a cache server in video streaming service in hybrid cloud computing environment. To guarantee performance of streaming service and to lower cost with VM placement in hybrid cloud environment, the algorithm CVOH considers both performance degradation in the internal cloud and the cost model in public cloud referred by Amit and Xia [5].

2 Problem Description and Scenario

Streaming content provider needs suitable number of cache server placed as a form of VM in datacenter for offering content to end-user's request with a stable performance. Kim et al. [6] analyze resource performance for inter- and intra- datacenter resource management under cloud CDN environment.

Streaming Performance with Consolidation. Experiment presented by Kim et al. focus on measurement of resource performance in respect to a number of VMs with applications using different computing resource.

The cache servers in form of VMs are placed in physical machines (PMs) and measure performance degradation caused by interference from other VMs within same PM executing application of different computing resource.

Table 1 [6] shows the completion time of video streaming service offered through a cache server placed in one PM of internal cloud, and it is in scale of msec. With decreasing network I/O performance, the completion time of video streaming service increases because of delay, buffering for downloading a content from a cache server. In the Table 1, using a same type of computing resource more can make performance degradation more.

Table 1. Measured network I/O performance. In the PM where a cache server for streaming service is placed, VMs using different computing resource is placed with a different number and Measuring the completion time of streaming service. [6]

	Video streaming completion time (msec)			
n	0	1	2	3
Compress-7zip [7] (CPU intensive)	220595	223811	222687	223674
Cachebench [8] (memory intensive)		221652	223008	220740
Bonnie ++ [9] (Disk I/O)		231943	253414	258503
Video streaming (Network I/O)		**230158**	**234090**	**268448**

The right side graph in Fig. 1. [6] shows the measurement of bandwidth with increasing a number of cache servers using same computing resource, network I/O, in a same PM. In this figure, network bandwidth converges with time, and it achieves more slowly as n is larger. That is performance degradation in streaming service appearing as a delay in a settling time. By the result showing, tendency of streaming performance degradation can be figured out when streaming cache servers are placed in consolidation manner in internal cloud.

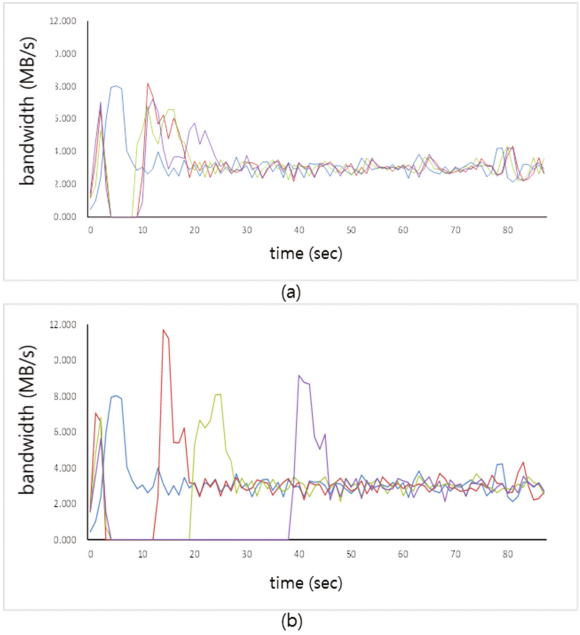

Fig. 1. Bandwidth of video streaming cache server measured with placing other VMs in same PM. Blue, red, green, and purple represent the number of other VMs n = 0, 1, 2, 3 respectively with Compress-7zip (a) (CPU intensive), streaming cache server (b) (network I/O intensive) [6] (Color figure online).

2.1 Video Streaming in Hybrid Cloud Environment

In this paper, we consider hybrid cloud Environment to decrease performance degradation of streaming service when computing resource in content provider's internal cloud is insufficient due to large scale of end-users' requests.

Figure 2 show architecture and flow of video streaming in hybrid cloud computing environment. In this environment, content provider send cache server request to both internal cloud and public cloud he uses after considering end-users' requests and a number of cache server needed. The cache server is placed in each PM in datacenter or public cloud as a VM. In this procedure, the VM placement module make a request set of proper VM placements with considering performance degradation profiling data in internal cloud and cost occurring by using public cloud resource.

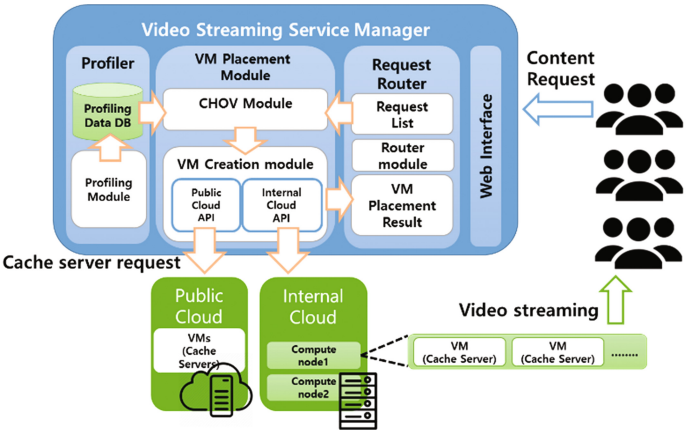

Fig. 2. Video streaming flow in hybrid cloud computing environment

The profiling data used in the VM placement module contains information of videos which content provider offers and about specification of internal cloud, performance degradation with cache server consolidation. The profiling data of videos are bit rate, size, video length and etc. In respect to an internal cloud, it contains specification of nodes in a datacenter, power consumption, performance degradation tendency of VM and etc.

We focus on an algorithm deciding how many VM has to be placed in public cloud and PMs in datacenter in VMP module, namely cost-effective VM offload in hybrid cloud environment. Main objective of the algorithm in this paper is to find an optimal solution when there exits tradeoff between minimizing cost of using public cloud and performance degradation in internal cloud.

3 CHOV Model

In this section, we present the CHOV used in VMP module for finding a solution of optimal VM placement. As an important consideration, we introduce two models, one is penalty cost model about performance degradation in datacenter and the other one is cost model occurring from using resource in public cloud. The CHOV is expressed by sum of those two models, and the solution of VM placement is the point that minimizes the sum of two models.

$VM = \{VM_0, VM_1, \ldots VM_n\}$ is a set of whole cache servers which needed to deal with end-users' requests, each cache server VM_i is placed in PM of internal cloud or public cloud as a VM. Equation (1) is a decision variable and **VM** depicts a definition of matrix of VMs.

$$vm_{ij} = \begin{cases} 1, & \textit{when } VM_i \textit{ is placed in } PM_j \\ 0, & \textit{when } VM_i \textit{ is placed in } PM_k \textit{ and } k \neq j \end{cases} \quad (1)$$
$$i \in \{0, 1 \ldots \mathrm{n}\}, j \in \{0, 1 \ldots \mathrm{m}\}$$

In Eq. (1), $PM = \{PM_0(= PC), PM_1, PM_2, \ldots PM_m\}$ is a set of PMs in internal cloud which content provider can manage and public cloud depicted as PC. PM_0 is a public cloud PC.

Cost model for public cloud. The cost model occurring from using resource in public cloud is Eq. (2) which Amit and Xia suggest. $\boldsymbol{l}$ denotes a number of VMs placed in public cloud, $\boldsymbol{K}$ is the cost of the first unit. $\boldsymbol{\alpha} \in (0, 1)$ is the learning factor of public cloud the content provider uses.

$$\text{Cost}(l) = \frac{K \cdot l^{1 + \log_2 \alpha}}{1 + \log_2 \alpha} \tag{2}$$

$$l = \sum_{i=1}^{n} vm_{i0} \tag{3}$$

The cost model Eq. (2) is based on the learning curve model. It assumes that as the number of production units are doubled the marginal cost of production decreases by a learning factor. It has been reported that for a typical Cloud provider like a Amazon EC2, the learning factors has typically value in range (0.75, 0.9).

Penalty Cost for performance degradation. Equation (4) denotes the penalty cost for performance degradation in internal **cloud** which content provider can manage.

$$\text{PenaltyCost}(\mathbf{VM}) = \beta \cdot Pd(\mathbf{VM}) = \beta \cdot \left(\frac{\sum_{\boldsymbol{VM}} Completion\ time}{\sum_{\boldsymbol{VM}} Video\ Length} - 1\right), \quad \beta > 0 \tag{4}$$

Video Length is the time from start to end of videos which content provider offers through each cache server, *Completion time* is the time really took from start to end when it is offered to end-user by streaming. As lower performance of streaming service make a buffering, delay happen, more completion time increases.

Thus, optimal problem considering penalty cost and cost for public cloud in hybrid cloud environment is denoted by Eqs. (5), (6) and (7).

$$\begin{aligned} \text{minimize} \quad \text{totalCost}(\mathbf{VM}) &= \text{PenaltyCost}(\mathbf{VM}) + \text{Cost}(l) \\ &= \beta \cdot Pd(\mathbf{VM}) + \frac{K \cdot l^{1 + \log_2 \alpha}}{1 + \log_2 \alpha} \end{aligned} \tag{5}$$

$$\text{Subject to} \sum_{j=0}^{m} \sum_{i=1}^{n} vm_{ij} = \text{n} \tag{6}$$

$$vm_{ij} \in \{0, 1\}, \forall i \in \{1, 2 \ldots n\}, \forall j \in \{0, 1 \ldots m\} \text{ and } vm_{ij} = 1 \; if \; VM_i \; is \; placed \; in \; PM_j. \tag{7}$$

To find a solution for this problem stated, implementation is under.

In a set of VMs, $VM = \{VM_0, VM_1, \ldots VM_n\}$, place each VM to the PM of the internal cloud in order of decreasing in expected network bandwidth based on profiling data. There are threshold values, $BW_{PM_j}^{thres}$ in each PM based on profiling data. This threshold values refer to range of network bandwidth where the performance degradation is marginal and measured experimentally. In this procedure, place the VMs maximally consolidated but not exceed the threshold in each PM denoted by Eq. (8).

$$\sum_{i=0}^{n} vm_{ij} \cdot Bw(VM_i) < BW_{PM_j}^{thres} \tag{8}$$

$Bw(VM_i)$ is expected network bandwidth in VM_i based on profiling data about content offered by content provider.

Algorithm 1. CHOV

INPUT
1. VM : Set of VMs needed to provide streaming service to end-users' request
2. PM : Set of PMs in internal cloud and Public cloud. PM_0 is a public cloud.

PHASE1. VM Placement only in PMs without degradation.

While ($VM \neq \emptyset$)
$vm = max\, BW\, VM_{\max\_bw}\, in\, VM$
Foreach $PM_j,\ j \in \{1, 2\ \ldots m\}$
If $BW_{PM_j}^{thres} > (\sum_{i=0}^{n} Bw_{PM_j}(VM_i) + vm)$
$PM_j \cup \{vm\}\ and\ VM/\{vm\}$
Break foreach
End if
End foreach
If vm can be placed in $\forall PM_j,\ j \in \{1, 2\ \ldots m\}$
$PM_0 \cup \{vm\}\ and\ VM/\{vm\}$
End if
End while
TotalCost = totalCost(**VM**)
Solution = **VM** /* current placement */

PHASE2. Find optimal placement to minimize totalCost(**VM**)

While ($PM_0 \neq \emptyset$)
$vm = min\, BW\, VM_{\min\_bw}\, in\, PM_0$
$PM_{\min\_bw} = \min BW\ PM_k\ in\ PM, k \neq 0$
$PM_{\min\_bw} \cup \{vm\}\ and\ PM_0/\{vm\}$
If TotalCost > totalCost(**VM**)
TotalCost = totalCost(**VM**) and Solution = **VM**
End if
End While

Fig. 3. The proposed CHOV algorithm.

After that, if VMs needed in streaming service remain, it denotes computing resource of internal cloud is insufficient to deal whole end-users' request without performance degradation. Hence there is need to offload VMs to the public cloud. To find the solution minimizing totalCost(**VM**), initially assume that all remaining VMs placed to public cloud. In order of increasing in delta of PenaltyCost, in other words, place the VM with minimal bandwidth in public cloud to the PM with using minimal bandwidth one by one until no VM is placed in public cloud, maximally consolidated.

Estimated Performance Degradation. To estimate **PenaltyCost**, approximation performance degradation estimation is needed. Equations (9) and (10) denotes performance degradation estimation based on profiling data about datacenter specification. It is found experimentally.

$$Pd(\mathbf{VM}) \approx Pd_{estimated}(\mathbf{VM})\frac{1}{2} \tag{9}$$

$$= \begin{cases} \gamma \cdot \{\sum_{j=1}^{m}\left(\sum_{i=0}^{n} vm_{ij} \cdot Bw(VM_i) > BW_{PM_j}^{thres}\right)\}^2 \\ \qquad \text{where} \sum_{i=0}^{n} vm_{ij} \cdot Bw(VM_i) > BW_{PM_j}^{thres} \\ 0 \text{ where} \sum_{i=0}^{n} vm_{ij} \cdot Bw(VM_i) \leq BW_{PM_j}^{thres} \end{cases} \tag{10}$$

$Pd_{estimated}(\mathbf{VM})$ reflects the fact that performance degradation becomes more severe as difference between estimated bandwidth and threshold in each PM. γ is a control parameter for scaling (Fig. 3).

4 Evaluation

4.1 Experiment Setting

To evaluate the CHOV, we form the experiment setting as shown in Fig. 4. To construct internal cloud of content provider, Openstack [10] is used. This cloud environment consists of a control node, 2 compute nodes and their specification is denoted as shown in Table 2.

The VMs used as cache servers have a 1 VCPU, 2 GB of memory, 20 GB Disk. For public cloud environment, we choose Amazon EC2 [11], and as cache servers VM instances with a 1 VCPU, 1 GB of memory, EBS only storage.

The videos that a content provider offers are shown in Table 3 with their profiling data.

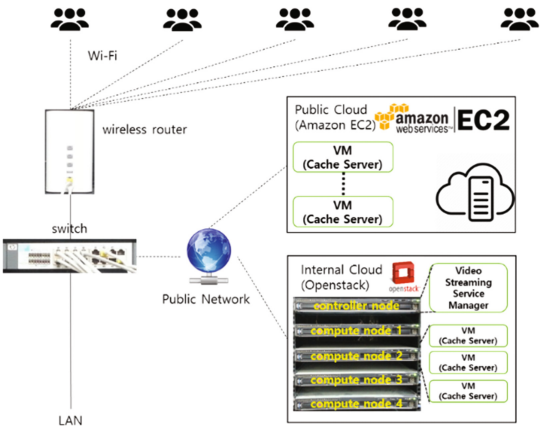

Fig. 4. Experiment setting in CHOV

Table 2. Internal cloud environment setting using Openstack.

	Control node	Compute node 1	Compute node 2
Functions	Cloud controller node, network, volume, API, scheduler, image services, Nova compute	Nova compute	Nova compute
Specification	16 cores (Intel® Xeon® CPU E5-2650 v2, 2.60 GHz),		
	32 GB memory, 242 GB Disk, Ubuntu 14.04.3 LTS	16 GB memory, 258 GB Disk, Ubuntu 14.04.2 LTS	

Table 3. 3 videos used in experiment and their profiling data. Each of them has a different bit rate, size, video length.

	Video 0	Video 1	Video 2
Bit rate	313 Kbyte/s	707 Kbyte/s	3008 Kbyte/s
Video length	231.6 sec	227.0 sec	227.0 sec
Video size	72.5 MB	160.5 MB	683 MB

4.2 Experiment Result

Figure 5 shows severe performance degradation is measured in case of maximal consolidation but comparing CVOH and maximal offloading, the whole performance degradation of VMs have similar aspect. The performance degradation is defined as Eq. (11).

$$\text{performance degradation} = \left(\frac{\sum_{VM} Completion\ time}{\sum_{VM} Video\ Length} - 1 \right) \tag{11}$$

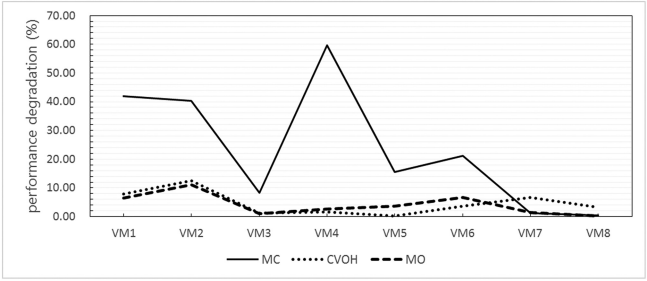

Fig. 5. Performance degradation graph of VMs placed as cache servers. MC, CVOH, MO denote respectively maximal consolidation, cost-effective VM offloading in hybrid cloud environment, maximal offloading cases.

In Fig. 6(a), as the worst case, performance degradation of video streaming results 23.5 % in case of MC, and CVOH is worse than maximal offloading case but there is small difference comparing to MC case as 4.58 %, 4.15 % are shown respectively in CVOH and MO. By the result shown in (b) and (c), CVOH is more cost-effective solution than other cases to content provider.

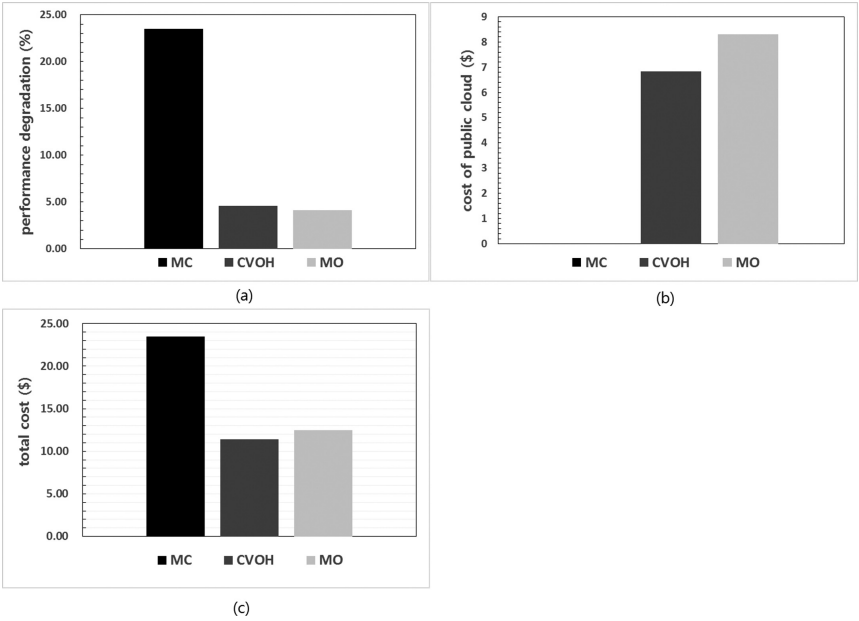

Fig. 6. This graph set shows the results of experiment in each case of MC, CVOH, MO. MC, CVOH, MO denote respectively maximal consolidation, cost-effective VM offloading in hybrid cloud environment, maximal offloading cases. (a) Entire performance degradation graph of each case. (b) Cost of using public cloud instances in each case. (c) Total cost in each case.

5 Conclusion

In this paper, we suggest cost-effective VM offloading algorithm for video streaming services in hybrid cloud environment. The CVOH considers cost model for using public cloud based on the learning curve model, and a penalty cost of performance degradation in internal cloud. From the result of experiments, CVOH shows that it has better performance than maximal consolidation in dealing end-users' request and it is also more cost-effective than maximal offloading. Comparing with a maximal consolidation case, CVOH has about twice better performance in total cost and comparing with a maximal offloading case, it has 0.4 % worse performance, but in total cost it has 9.1 % better.

Acknowledgments. This research was supported by the MSIP under the ITRC (Information Technology Research Center) support program (NIPA-2014(H0301-14-1020)) supervised by the NIPA (National IT Industry Promotion Agency), and 'The Cross-Ministry Giga KOREA Project' grant from the Ministry of Science, ICT and Future Planning, Korea.

References

1. AWS Case Study: Netflix. http://aws.amazon.com/ko/solutions/case-studies/netflix/
2. Google Cloud Boosting YouTube Upload Speeds. http://blogs.wsj.com/digits/2011/02/14/googlecloud-boosting-youtube-upload-speeds/
3. Aggarwal, V., Gopalakrishman, V., Jana, R., Ramakrishnan, K.K., Vaishampayan, V.A.: Optimizing cloud resources for delivering IPTV services through virtualization. In: COMSNETS (2012)
4. Bossche, R.V.D., Vanmechelen, K., Broeckhove, J.: Cost-optimal scheduling in hybrid IaaS clouds for deadline constrained workloads. In: CLOUD 2010
5. Amit, G., Xia, C.H.: Learning curves and stochastic models for pricing and provisioning cloud computing services. Serv. Sci. **3**, 99–109 (2011)
6. Kim, H., Hyeon, M., Jang, H., Youn, C.: An analysis of resource performance for inter- and intra- datacenter resource management under cloud content delivery network environment. In CIAPT (2015)
7. Phoronix Test Suite. http://www.phoronix-test-suite.com/
8. LLCbench. http://icl.cs.utk.edu/projects/llcbench/cachebench.html/
9. Bonnie++. http://www.coker.com.au/bonnie++/
10. OpenStack. http://www.openstack.org/
11. Amazon EC2. http://aws.amazon.com/ec2