SLA-Aware Resource Management for Cloud based Multimedia Service

모하메드 사비르 하산, 모하메드 마타하리 이슬람, 박준영, 허의남
경희대학교 컴퓨터공학과
e-mail : {sabbir, motahar, parkhans,johnhuh} @khu.ac.kr

SLA-Aware Resource Management for Cloud based Multimedia Service

Md. Sabbir Hasan, Md. Motaharul Islam, Jun Young Park and Eui-Nam Huh
Dept. of Computer Engineering, Kyung Hee University

ABSTRACT

Virtualization technology opened a new era in the field of Data intensive, Grid and Cloud Computing. Today's Data centers are smarter than ever leveraging the Virtualization technology. In response to that, Dynamic consolidations of Virtual Machines (VMs) allow efficient resource management by live migration of VMs in the hosts. Moreover, each client typically has a service level agreement (SLA), leads to stipulation in dealing with energy-performance trade-off as aggressive consolidation may lead to performance degradation beyond the negotiation. In this paper we propose a Cloud Based CDN approach for allocation of VM that aims to maximize the client-level SLA. Our experiment result demonstrates significant enhancement of SLA at certain level.

1. Introduction

Cloud Computing has received significant attention recent times as the hype is created and responded largely by the companies like Amazon, IBM, Google, Yahoo!, Microsoft, Sun, NASA and RackSpace by making their own cloud platforms for consumers and enterprises to access the cloud resources through services. With the rapid development of Virtualization technology including the advantage of isolation, consolidation and multiplexing of resources, became key role to deploy in modern data centers [1]. Due to virtualization, numerous tasks are seen as a single entity in a virtual machine. It opens new capabilities such as Live migration [2], that attracted substantial attention in recent years to respond to the challenges for cloud computing by providing Load balancing, Power Efficiency and Transparent Infrastructure maintenance to the Virtual Machines, that requires new management logic. On the other hand, Content Delivery Networks (CDN) [3] is a collaborative framework of network on Internet that replicates content over several mirrored Web servers strategically placed at various locations in order to deal with flash crowds [4] or SlashDot. CDN has been used to distribute contents all over the world. We want to bridge Cloud computing and CDN together to maximize our goal by efficient resource management

Virtualization technology leverages Live Migrations of VMs with changing workload in the Physical Host to reduce the number of Physical Hosts or server. To eliminate the static power, idle servers are switched to sleep mode to reduce the energy consumption whereas any idle Hosts can be reactivated when the resource demand increases. However, Infrastructure providers often end up over provisioning of resources to maximize the Quality of Service results High Energy Consumption, poor resource management and Large Operational cost. QoS requirement can be formalized in Service Level Agreement that serves as the foundation for the expected level of service between the Cloud consumer and the Service or Cloud provider. So, the purpose of the optimal resource provisioning by VM consolidation is to make an energy-performance trade-off. The problem of optimal resource allocation and management is challenging due to the diversity present in the clients application and Dynamic workload that have to process by Hosts in the data center.

As Cloud Computing concept evolves and more consumers adopting this technology, the quality and reliability of the Multimedia Cloud services (VoD, IPTV, Live Television etc.) become important aspects. However the demands of the service consumers vary significantly. It is not possible to fulfill all consumer expectations from the service provider perspective and hence a balance needs to be made via a negotiation process. At the end of the negotiation process, provider and consumer commit to an agreement. By managing resource efficiently, IaaS Cloud Provider can ensure better Quality of Service to the Consumer, led to satisfactory results towards SLA. Our contribution can be summarized as follows:

- VM resource allocation scheme to ensure target SLA By Cloud Based CDN
- VM Allocation procedure by server consolidation ensuring low SLA violation

- Comparison of our approach with existing literature.

2. Related Work

Our work is based on Dynamic Consolidation of Virtual Machine's retaining strict Service Level Agreement to consumers. There are several research groups in both academia and industry, working on resource allocation and management by performing static and dynamic consolidation of VM's and servers. Cardosa et al. [5] proposed a solution for VM placement and power-efficient consolidation of VMs in modern data centers where it runs heterogeneous applications. They have adopted min, max, share parameters of XEN and VMware, that represents the Utilization limit of upper and lower of CPU allocation and sharing same resources by different VMs. They also considered a priority based approach for peak load of enterprise environment. As a result it does not support strict SLA and VM allocation is static.

Verma et al [6] described a power aware application placement framework in which at each time frame the placement of VMs is optimized to minimize the power consumption and maximize the performance at certain level. The main difference with his work is that Our proposed algorithm doesn't violate Strict SLA requirement when workload is varied and unpredictable.

Stilwell et al [7] proposed a formulation of the resource allocation problem in shared hosting platform for static workloads with servers that provide multiple types of resources. Their algorithm runs faster in large systems and fulfill QoS requirement but it lack dynamicity when workload in unpredictable and dynamic. Like him other researchers [8] [9] also studied VM resource management techniques to maintain QoS requirement when workload is static in Cloud Computing.

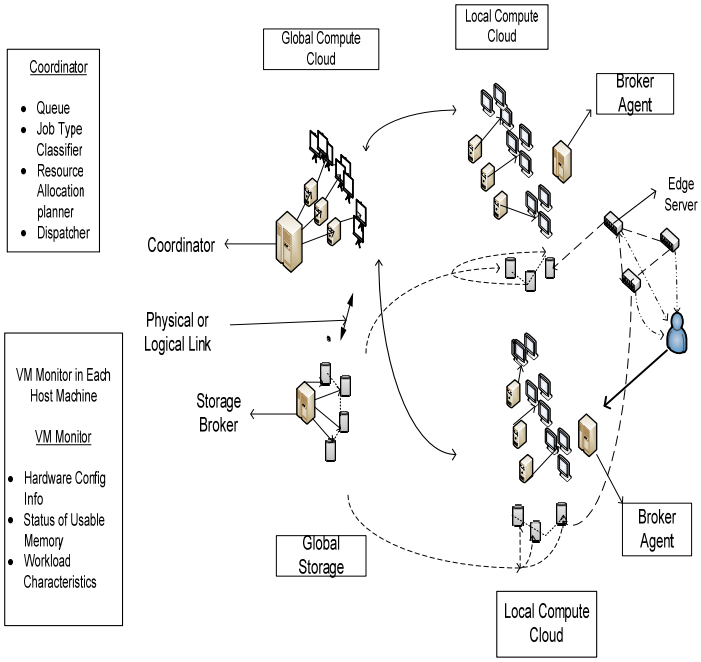
3. Proposed Scheme

The Figure-1 depicted overall Cloud Based scenario for Multimedia Service. We have three components in our architecture. Those are described as follows:

1. *Global Compute Cloud*: It can be referred as Primary datacenter or IaaS provider who has all the raw resources to satisfy consumers or end User's. When a resource demand is submitted from User/Consumer, it handles all necessary tasks to provide the required resources. It has a coordinator module which maintains a proper Job Queue. Queue length longer means serving time will be higher, lead to SLA violation. Job classifier is another component which differentiates and sets priority of resource to assign for specific applications. Moreover, Resource Allocation planner-based on the previous epoch, it predicts demands for future events and continuously do that at run time. After finishing the calculation, Global Compute Module dispatches the resources via dispatcher through local compute cloud or Cloud Consumers. Primary storage is linked physically and logically with the Global Compute Cloud.

2. *Local Compute Cloud:* All the Local Compute Cloud is connected to Global Compute Cloud. It has Broker agent. End User's submit their resource requirement to Broker Agent. If resource is not sufficient it communicates with other compute cloud otherwise redirects to the User. It contains all meta data information. If User request for popular contents it can serve User request by nearest Edge Server's like Youtube does.

3. *Edge Server:* Edge server's normally doesn't have embedded intelligence. Normally it cashes popular and recent contents viewed by user. It also can communicate to neighbor edge servers to stream contents efficiently to users. Edge servers are gaining much popularity due to the era of Cloud Based CDN as it reduces the downtime and improve Quality of Service.



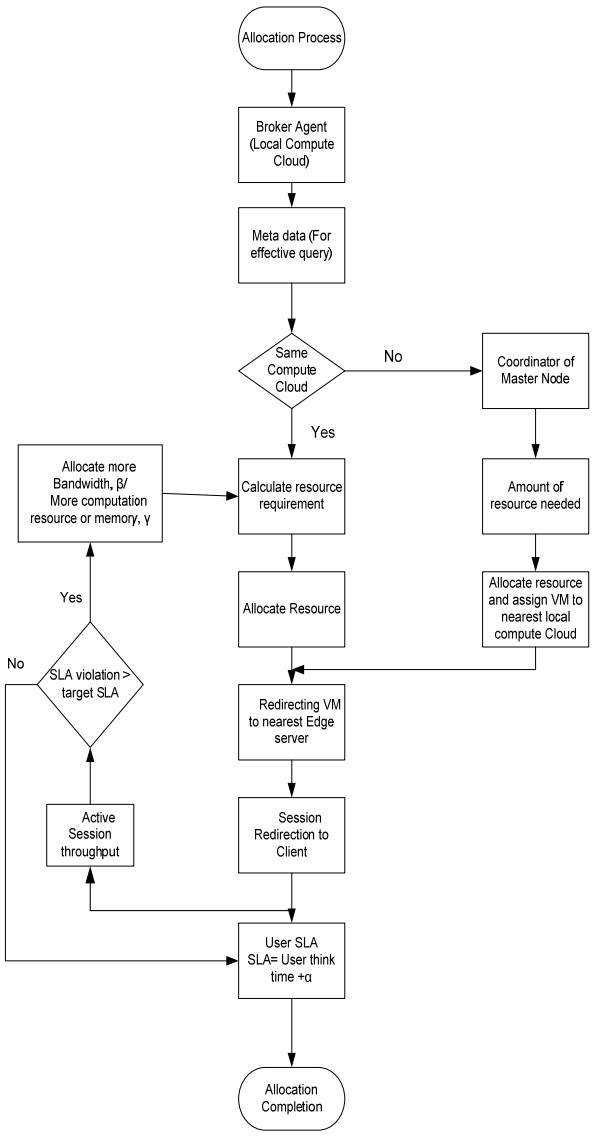
(Figure 1) Cloud Based CDN Scenario for Multimedia Cloud Service

Figure-2 depicts the Resource Allocation process in our proposed scheme. The whole approach largely depends on how we can maximize SLA. So depending on the target SLA, the flow chart allows us option to add more or release resources when it's required and provides optimal allocation of Virtual Machines.

4. VM allocation

The problem of VM allocation can be divided in two ways. Those can be mentioned as- admission of new requirements for VM provisioning and enlisting the VMs in the host, and optimization of current VM allocation. The first part can be

seen as a bin packing problem with variable bin size and prices, whereas bins represent the physical hosts, bin sizes are the available CPU capacity and bin process are the energy consumption. We solved the first problem by using Best Fit Decreasing Algorithm that is shown to use no more than bins [10].

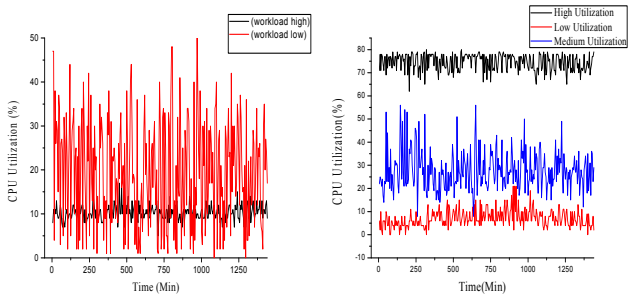


(Figure 2) Resource Allocation Process

However, for solving the second problem, We consider power consumption of Hosts and CPU utilization of Hosts. The pseudo-code of the VM placement is presented in the Algorithm 2. We calculate the other available host's current CPU Utilization and increase of Power consumption if the selected VM had have migrated. By using $\sqrt{x^2 + y^2}$, we get a point where it indicates the tradeoff point of CPU Utilization and incremental of Power Consumption of Hosts and find the suitable Host for VM. Otherwise it will find a Host which Utilization is higher calculating that the Host doesn't get overloaded if the VM migrate to that Host.

5. Experiment

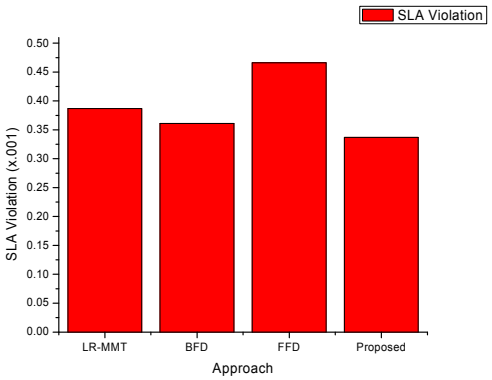
We evaluate our proposed algorithm in CloudSim toolkit. CloudSim is an extensible simulation toolkit that enables modeling and simulation of Cloud Computing systems and application provisioning system created by CLOUDS Lab, University of Melbourne [11]. The *PowerVmAllocationPolicyMigrationAbstract* and *PowerVmSelectionMinimumMigrationTime* classes are modified according to our proposed algorithm.We have conducted experiment with real life data provided by the CoMon project, a monitoring infrastructure for PlanetLab [12]. We used different types of workload to validate our approach. We considered dynamic Environment over static environment.



$$\text{SLA time per active Host} = \frac{1}{N}\sum_{k=1}^{N}\frac{X_{sk}}{X_{ak}}$$

$$\text{Performance degradation due to migration} = \frac{1}{M}\sum_{l=1}^{M}\frac{P_{dl}}{P_{al}}$$

Where N and M are the number of Hosts and VMs respectively; X_{sk} is the total time when host k has experienced 100% CPU utilization caused SLA violation; X_{ak} is the total time host was active; P_{dl} is the approximation of performance degradation due to migration of VM l and we consider it 10% ; P_{al} is the total CPU capacity requested by the VM l during its epoch.



(Figure 3): SLA Violation Comparison

6. Acknowledgement

This work was partly supported by the IT R&D program of MKE (The Ministry of Knowledge Economy)/ KEIT (Korea Evaluation Institute of Industrial Technology) [10035321, Terminal Independent Personal Cloud System]. Professor Eui-Num Huh is corresponding Author.

References

[1] B. Sotomayor, R. Montero, I. Lorente, and I. Foster, “Virtual infrastructure management in private and hybrid clouds”, IEEE Internet Computing, pp. 14–22, 2009.

[2] C. Clark, K. Fraser, S. Hand, J. G. Hansen, E. Jul, C. Limpach, I. Pratt, and A. Warfield, “Live migration of virtual machines”, in Proc. of the Second Symposium on Networked Systems Design and Implementation (NSDI’05), 2005. pp. 273–286,

[3] A. vakali and G. Pallis. “Content Distribution Network-status and trends”, IEEE Internet Computing, pages 68-74, November-December 2003.

[4] M. Robinovich and O. Spatscheck, “Web caching and replication”, Addison Wesley, USA, 2002.

[5] M. Cardosa, M. Korupolu, and A. Singh, “Share and utilities based power consolidation in virtualized server environments,” in Proc of IFIP/IEEE International Symposium on Integrated Network Management IM’09, 2009.

[6] A. Verma, P. Ahuja, A. Neogi, pMapper: Power and migration cost aware application placement in virtualized systems, in Proc of the 9th ACM/IFIP/USENIX International Conference on Middleware, Springer, pp. 243-264, 2008..

[7] M. Stilwell, D. Schanzenbach, F. Vivien, H. Casanova, “Resource allocation algorithms for virtualized service hostings platform,” Journal of Parallel and distributed Computing, vol.70, no. 9 pp. 962-974, 2010.

[8] Jeffrey M. Galloway, Karl L. Smith, Susan S. Vrbsky, “Power aware load balancing for Cloud Computing,” in Proc of the World Congress on Engineering and Computer Science 2011 Vol I October 19-21, WCECS 2011.

[9] Kuo-Qin Yan ; Wen-Pin Liao ; Shun-Sheng Wang , “Towards a Load Balancing in a three-level cloud computing network ”, in Proc of 3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT), 2010, Vol-1, pp.-108-113, 2010.

[10] Yue M. “A simple proof of the inequality FFD (L) < 11/9 OPT (L) + 1, for all 1 for the FFD bin packing algorithm.” Acta Mathematicae Applicatae Sinica (English Series) 1991; 7(4): 321-331.

[11] RN Calheiros, R. Ranjan, A. Beloglazov, CAFD Rose, R. Buyya. “CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms.” Software: Practice and Experience 2011; 41(1): 23-50.

[12]KS Park, VS Pai. “CoMon: a mostly scalable monitoring system for PlanetLab.” In Proc of ACM SIGOPS Operating Systems Review 2006; 40(1):74.