2012 International Conference on Information and Computer Applications (ICICA 2012)
IPCSIT vol. 24 (2012) © (2012) IACSIT Press, Singapore

Location-Aware Dynamic Resource Allocation Model for Cloud Computing Environment

Gihun Jung[+] and Kwang Mong Sim

Gwangju Institute of Science and Technology (GIST)

Abstract. Allocating virtual machine (VM) to an appropriate physical machine (PM) is important to enhance the performance of cloud computing environment. In this paper, we propose a dynamic resource allocation model based on the utilization level of PMs in data centers, and the location of user and data center on cloud computing environments. In addition, this paper also proposes a resource management architecture to perform 1) location-aware VM placement and 2) dynamic resource utilization management. Through experimental simulations, the results show that the proposed model guarantees to allocate a VM to an appropriate PM that has proper utilization level for the data center, which is not affecting the performance of each allocated VM, and better response time of each VM due to close location to user.

Keywords: Location-Aware, Agent-Based Cloud Computing, Dynamic Resource Allocation

1. Introduction

One of the most significant benefits of cloud computing is reducing the operating cost of data center through virtualization [1]. To support cost reduction, the resource of physical machines (PM) in data center should be efficiently utilized. However, if the provider only considers maximizing the utilization of data centers (i.e., maximizing the utilization level of physical machines), eventually, it has a bad influence upon the performance of virtual machines (VM) in data center due to high workload of each associated physical machine. To prevent such a performance degradation, an appropriate VM allocation scheme is needed for the overall performance of cloud computing. In addition, the provider needs to set an appropriate threshold of utilization level that does not affect to the performance degradation of virtual machines in the data center.

In this paper, for better performance of VMs in terms of response time, we consider the location of each VM. To provide worldwide services, the provider should have several data centers according to geographically locations. Since cloud computing services are delivered over the public internet [1], [2], which does not guaranteed reliability in general, there may be undesirable performance degradations such as slow response time. Although the provider can designate the allocation for new VMs to a low utilized PM that guarantees no performance degradation due to utilization level, a performance degrade is still possible to occur. If the location of a PM that is prvoding the user's VM is far from the location of a user, the geographical distance between the PM and the user affect to the response time of the VM. Therefore, a cloud provider needs to consider not only the utilization level of PMs, but also the location of a PM to allocate the user request as a VM. To address these issues, this paper proposes a dynamic resource allocation model that extends the model in [3]. The new model considers 1) the location of PMs, and 2) the dynamic utilization level of PMs. The contribution of the paper is as follows. This paper proposes a hybrid resource management architecture to perform 1) location aware VM placement and 2) dynamic resource utilization management so that the model allows a provider to place a new VM in an appropriate PM that shows the best performance and guarantees the maximized utilization level, which prevents performance degradation of the data center.

[+] Corresponding author
E-mail address: ghjung@gist.ac.kr, prof_sim_2002@yahoo.com

The remainder of this paper is organized as follows: Section 2 presents an overview of agent-based cloud computing system for demonstrating the idea of the proposed model. Section 3 describes the proposed location aware dynamic resource allocation model. Section 4 describes the experimental environment and shows the performance evaluation of the proposed model in terms of the allocation outcomes (i.e. response time of user's VM). Finally, Section 5 concludes this paper with a list of future works.

2. System Architecture

An agent-based cloud computing environment is simulated to demonstrate ideas of the proposed resource allocation model. Like Amazon EC2 [4], the proposed system also provides its resources as an infrastructure service in a form of VM instances. To provide flexible and on-demand service, the proposed system allows to user to request an arbitrary amount of resources at any time and from anywhere. For managing flexible user request, the proposed system is designed as a hybrid architecture that is combined with centralized and distributed resource management architectures. As shown in Fig. 1, there are two types of components in a data center: 1) Data Center Super Node (DCSN) and 2) sets of PMs. DCSN is responsible to manage resource utilization reports from PMs in the data center and for searching an appropriate PM to allocate (or migrate) VMs based on its utilization level and its location. DCSN has two subcomponents: 1) Report Repository (RR), and 2) Decision Making Engine (DME). RR stores current resource utilization report from sets of PMs. DME finds a proper PM to allocate new VMs with reduced number of times for the migration while the PMs are running based on these reports. The details to decide an appropriate PMs for new VM or migrated VM are described in Section 4 that include a decision making model for VM placement and a decision making model for VM migration.

Physical Machine is a real resource (i.e. it is combined different types of resources such CPU, memory, or network bandwidth) that can allocate many VMs. In a PM, there is a specialized layer for virtualization - a hypervisor. The hypervisor generally has the responsibility to allocate VMs and share its resources like traditional operating systems. In this paper, we assume that a hypervisor has a specialized VM called domain zero (Dom 0). To monitor resource utilization, a monitoring agent runs as a module of the Dom 0. The machine monitoring agent (MMA) is used to watch the utilization of each PM. If there are some problems (e.g. exceeding the threshold of utilization level, etc.), the MMA immediately reports to the DCSN to mitigate the problems. For more user dependent monitoring, every user VMs has a small monitoring agent named as a user monitoring agent (UMA). The UMA reports the utilization level of running application (e.g. web server process) to the MMA. Through the proposed system, the provider may use different kinds of resoure allocation models.

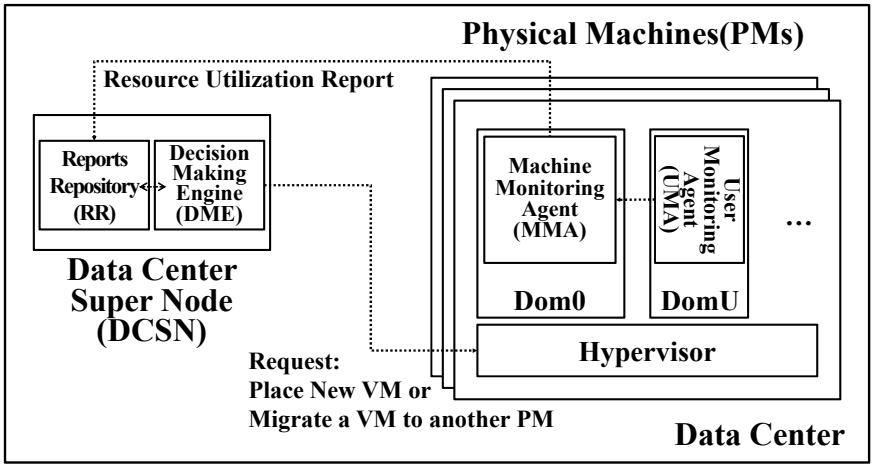

Fig.1 Overview of Agent-Based Cloud Computing System

3. Location Aware Dynamic Resource Allocation Model

To guarantee appropriate utilization of PMs and response time of VMs, the model considers two different factors. Since we assume provider's data centers are geographically distributed, the geometric

distance may affect to the response time of user VM. Hence, the provider should place user VM to the PM in data center as close as the location of user.

3.1. Utility Function

To find out which PM is appropriate for a new VM or migration, the provider evaluates each PM using a utility function as follows:

$$u_m = \alpha * u_m^u + \beta * u_m^T + \gamma * u_m^\sigma \ \left(where, 0 \le u_m \le 1\right) \tag{1}$$

In (2), we have three terms to evaluate the suitability of each data center: 1) the utilization level; 2) expected response time and 3) the location. By (2), cloud computing providers can find appropriate PMs based on higher utility values determined by equation (1).

For utilization level, U_m^U, it is described as follows:

$$u_m^U = \begin{cases} U_{\min} + 1 - \left[\frac{W_\theta - W_c}{W_\gamma - W_c}\right] & W_c \le W_\theta < W_\gamma \\ U_{\min} & W_\gamma \le W_\theta \end{cases} \tag{2}$$

Let W_θ be the expected utilization level of the PM *m* by allocating or migrating user VM *u*, and W_c and W_τ be the current utilization level and pre-defined threshold of the utilization level. In case of placing a new VM (i.e. also migrates from another PM), the provider may consider the utilization increasing cause by allocating a new VM to a PM. In this paper, we assume the provider can predict how much the total utilization level has increased after allocation. Hence, (3) allows the provider to evaluate the suitability of each PM for the user's VM. If the expected utilization level exceeds the threshold level, the provider evaluates the PM as minimum utility (i.e. $U_{min} = 0.1$) to allocate or migrate.

For the utility for response time, we assume that it may be affected by both the utilization level of a PM and the geographical distance between the user and the PM. The response time utility is described as follows:

$$u_{\min}^T = \begin{cases} U_{\min} + 1 - \left[\frac{T_e - T_c}{T_{SLA} - T_c}\right] & T_c \le T_e < T_{SLA} \\ U_{\min} & T_{SLA} \le T_e \end{cases} \tag{3}$$

Let T_{SLA} be the Service Level Agreement (SLA) of the response time for a user VM. Let T_c be the current response time in a machine, and T_e be the expected response time after migration or placement to another machine. When T_e exceeds the limit given by an SLA, the value of utility function returns the minimum value (i.e. $U_{min} = 0.1$).

To evaluate the location of each PM and user, the utility of geographical distance $u_{\min}^\delta$ is represented same as previously proposed in [3]:

$$u_{\min}^\delta = 1 - \left[\frac{\Delta_m^u}{\Delta_{range}}\right] \tag{4}$$

To apply the utility function (1), Fig. 2 describes the procedure of two different decision-making in the proposed model: VM placement and migration. The following sub-sections described the detail of each process.

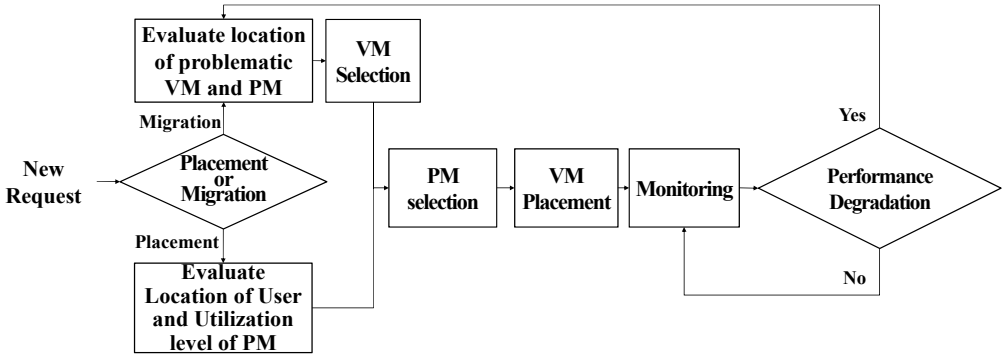

Fig.2 Procedure of the proposed model

3.2. VM Placement Decision Making

When a provider receives a new VM placement for a user, the provider first evaluates the network delay between the user and each data center.

If the provider finds closest data center for the user VM, the provider now evaluates each utilization level of PMs based on utilization report, which is sent from each PM.

Using the utility function, the provider evaluates the feasibility of each PM to place the user VM. Then, the provider chooses the PM that shows highest utility value for placement of the user VM

After VM placement, each PM keeps monitoring its utilization level, and reports to the provider when the utilization is changed.

3.3. VM Migration Decision Making

When a PM's utilization level exceeds a given threshold, it reports to the provider to decide whether VMs are needed to migrate to another PM.

The provider evaluates the feasibility of migration based on the utility function (1). Even if there is a migration overhead (i.e. migration process time), the value of utility function is higher than non-migration situation, the provider decides to migrate problematic VMs to another PM.

When a provider decides to migrate a problematic VM, the provider instructs the hypervisor to migrate the VM to another appropriate PM

After migration, each PM keeps monitoring its utilization level and reports it.

4. SImulation and Empirical Results

To show the feasibility of the proposed resource allocation model, a series of experiments was carried out using the agent-based testbed to evaluate the performance in terms of the response time of user VM by comparing to different types of allocation models.

4.1. Performance Measure

For evaluating the performance of the proposed model, we used the response time of each user's VM as a performance factor. The response time of every VM is different based on the workload and location of PM. In this experiment, the response time of the user's VM is changed during simulation time. The detailed expression is as follows:

$$R_{sp.} = W_c * T_c + W_{vm} * T_{vm} + \Delta_m^u \quad (4)$$

4.2. Experimental Settings

As shown in Tab. 1, there are ten different location zones in this experiment. Each zone represents geographical distributions for both user and data center. Since we assume every data center is homogenous, they have the same amount of resources in terms of numbers of physical CPUs, memory and storages. However, the request of user is randomly generated in certain range of constraints (i.e. the response time of each VM may be varied in certain range). For migration constraints, there is a SLA for response time of user VM. When the response time of a user VM reaches a given level, the provider may consider migrating the VM to another PM.

Tab.1 Experiment constraints

Number of User Request	Number of Data Center	Limit of PM Workload	Location Range	Response Time SLA
250	10	20.0	{zone 0, …, 9}	100ms

4.3. Simulation Results

Our experiments show that the proposed model (LADRA) achieved better performance than the previously proposed model (TDARA). As shown in Fig. 3, the response time of each user's VM was increased generally over time. This is because that the number of user requests have accumulated during

simulation. Consequently, it affects the increase in overall workload of the PMs. In Fig. 3, without using the proposed model, the response time was increased over the limit given in the SLA. However, for the proposed model, even though the response time was increased over time, the response time did not exceed limit given in the SLA. This is because using the proposed model, the provider keeps monitoring the response time of each VM to prevent exceeding the limit given in the SLA. In addition, by the monitoring the utilization level of PMs, the proposed model could migrate a problematic VM to another machine for reducing performance degradation.

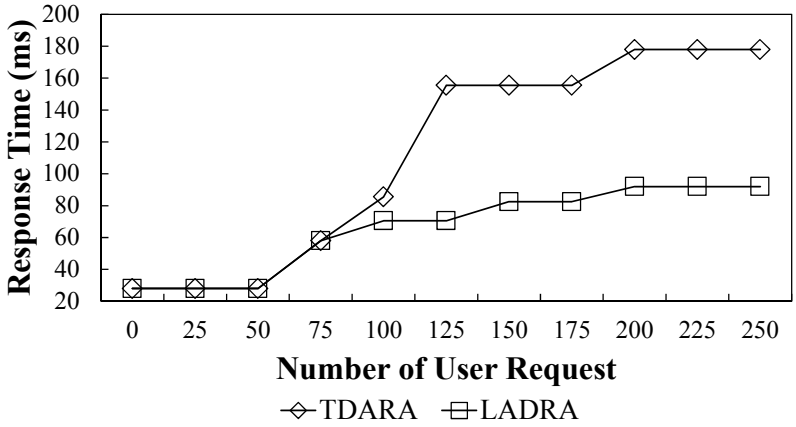

Fig.3 The performance of the proposed model

5. Conculsion and Future Work

We propose a dynamic resource allocation model based on the utilization level of PMs in data center and the location of user and data center on cloud computing environments. We also propose a resource management architecture to perform 1) a location aware VM placement and 2) a dynamic resource utilization management. The proposed model allows a provider to dynamically place a new VM to an appropriate PM that would achieve the best performance, guarantee the maximized utilization level, and prevent performance degradation, of the data center. The simulation results show that the proposed model guarantees an appropriate utilization level for the data center that is not affecting the performance of each allocated VM, and better response time of each VM than traditional models. However, the evaluation is executed in a testbed simplified from the cloud computing environment introduced in Section 2. Consequently, for the future work, we are plan to 1) implement a real cloud computing environment such as Xen Hypervisor [5] and 2) evaluate the proposed model to verify the real performances in a practical situation.

6. Acknowledgement

This work was supported by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korea Government (MEST 2009-0065329) and DASAN project (140316).

7. References

[1] MICHAEL Armbrust; ARMANDO Fox; REAN Griffith; ANTHONY D. Joseph; RANDY Katz; ANDY Konwinski; GUNHO Lee; DAVID Patterson; ARIEL Rabkin; ION Stoica; MATEI Zaharia. A view of cloud computing. *Commun. ACM* 53, 2010, PP50-58.

[2] PETER Mátray; PETER Hága: SANDOR Laki; ISTVAN Csabai; GABOR Vattay., On the network geography of the Internet, INFOCOM, 2011 Proceedings IEEE , 2011, PP126-130

[3] GIHUN Jung; KWANG MONG Sim; PAUL C. K. Kwok; MINJIE Zhang., A TIME-DRIVEN Adaptive Mechanism For Cloud Resource Allocation. Proceedings of IC-BNMT2011, 2011, PP441-446

[4] Amazon EC2, http://aws.amazon.com/ec2/

[5] Xen Hyperviosr, http://xen.org/products/xenhyp.html