Virtual Machine Migration Strategy Based on the Load Evaluation and Prediction under the Cloud Environment

Zi Yueqi[1,a], Sun Yongqiang[2] , *Sun Yi[b]

[1]Beijing University of Posts and Telecommunications, Beijing 100876, China

[2]China Waterborne Transport Research Institute

[a]email:2013213317@bupt.edu.cn, [b] email:sunyisse@bupt.edu.cn

Keywords: Cloud Computing; Load Balancing; Prediction; Threshold; Virtual Machine Migration

Abstract. Concerning the centralized communication overloading of the virtual machine (VM) in the cloud computing center, this paper puts forward an adjustment strategy, that is, to build a model through the load evaluation of three threshold values in consideration of the load balancing and to achieve the objective of dynamic migration of the VM. First, load evaluation algorithm for four statuses is designed by increasing the division grade of the physical machine (PM). The exponential smoothing model based on time series trend prediction is introduced. Through the prediction of the load in the next moment, the overloaded PM is certified. The VMs to be migrated and the optimal objective target hosts are found out through the calculation of the utilization rate and the utilization trend of various VMs, CPUs, internal storage and bandwidth of target hosts. Based on that, the model is established and the VMs are migrated.

Introduction

With the coming of information era, the demand of processing massive data resource is more scientific and systematical [1]. In addition, cloud resources are stored in heterogeneity and the user requirements is more and more diversified. For example, ships product huge amounts of AIS data during shipping, and need to analysis the high risk area of water through data. To deal with it, high performance of system are required. Thus, the number of VM in data center rises perpendicularly and the way of VM migration becomes complex. If the modulation is unreasonable, system is bound to lead to a waste of resources and even load overweight.

Nowadays, because of lacking of considering about the change of user request, the traditional static VM migration strategy cannot schedule resource rapidly and dynamically. And the manual allocation of resources is obviously lag behind [2]. What's more, others are hardly related to the influence of resource availability, such as CPU, memory and network bandwidth, to the VM migration and cannot guarantee the quality of service [3].

All around these issues, this paper proposes a resource scheduling scheme based on evaluation and prediction of load. It combines a variety of resource usage, realizes load balancing and improves system available by live migration of VM, while achieving the goal of reasonable allocation of resources.

Description of VM migration

The generation, scheduling and destruction of VM are three main steps of VM deployment. Since the processing capacity of each VM is gained and restricted by the host machine, system should evaluate load according to resource usage, and then dynamically schedule VM. When the number of VM is few and the load is low, a migration of specific VM is necessary for ensuring load balancing and the quality of service. The information of VM will be thoroughly destroyed from HM after removal. The flow of VM live migration is made up with the following several steps.

（1）Evaluate the load of each PM. The use of resources of PM should be detected regularly and compared to thresholds. Then dividing PM into four level in accordance with the result. They are light, normal, danger and overload.

（2）Predict the load of specified PM. If the level of PM is dangerous, it need a prediction for its load in next moment. The PM will be added to overload classification as the predictions are greater than the highest threshold.

（3）Elect the appropriate VM to migrate from overweight classification.

（4）Search the optimal PM for migratory VM, according to the available resource and the energy consumption level of PM.

（5）Migrate PM and destroy the information from host machine

The system structure

Compute the loading situation of each PM

At this stage, the load of VM and PM are computed synthetically. Meanwhile, defining the values of three thresholds and comparing with the load to grade PM. Different grades are corresponding to different treatment methods of VM.

Definition the load of VM

System evaluates the state of load of each VM under dynamic weighted load algorithm. In order to better reflect the performance difference, three variables are set concerning the influence of three factors, namely CPU, internal storage and bandwidth, on the load of centralized cluster nodes. The comprehensive loading situation of the normalized VM is worked out, and expressed as $Load_{vm}[i]$. Below is the equation to show the relationship of the above parameters.

$$Load_{vm}[i] = \begin{cases} \alpha \dfrac{CPU_{used}}{CPU_{\max}} + \beta \dfrac{MEM_{used}}{MEM_{\max}} + \gamma \dfrac{NET_{used}}{NET_{\max}}, & \begin{matrix} CPU_{used} < CPU_{\max} \\ MEM_{used} < MEM_{\max} \\ NET_{used} < NET_{\max} \end{matrix} \\ 1, \ others & \end{cases} \tag{1}$$

Where, α, β and γ stand for the weight of various parameter indexes, $\alpha+\beta+\gamma=1$; CPU_{used} and CPU_{max} stand for the current usage amount and the maximum usage amount of CPU resources; MEM_{used} and MEM_{max} respectively stand for the current usage amount and the maximum usage amount of internal storage resources; NET_{used} and NET_{max} respectively stand for the current usage amount and the maximum usage amount of the bandwidth resources.

Grading of PMs

According to the current strategies, most systems use two threshold values to evaluate the loading of PMs. The shortage of the method is that the threshold value is hard to be correctly quantized. If the definition of the threshold value is too sensitive, many loads might easily stay in the PMs and cause unnecessary migration. However, if the definition of the threshold value is too blunt, it is hard to correctly and efficiently choose the overloaded PMs. Therefore, this paper designs the load evaluation strategy based on three threshold values to solve the problem.

There are three threshold values, namely L_1, L_2 and L_3, using to divide the virtualized resources into four grades according to different loading situations. Four grade are sequenced from small to large, they are respectively light, normal, danger and overweight. Before the calculation of the threshold, the average value of the comprehensive load intensity of all system nodes and the load value of all physical nodes of the system can be worked out first. They respectively expressed as $Load_{ave}$ and $Load_{res}$. Since $Load_{VM}[i]$ allows the existence of errors within the specific range, the load standard deviation, σ, is used to express the error.

$$Load_{ave} = \frac{1}{m} \sum_{i=1}^{m} Load_{vm}[i] \tag{2}$$

$$Load_{res} = \frac{RES_{free}[j]}{\sum_{j=1}^{|v^s|} RES_{free}[j]} Load_{ps} \tag{3}$$

Where, m is the quantity of all resources in the system; RES_{free} stands for the quantity of free resources in the PM; $Loap_{ps}$ stands for the current overall load capacity of the PM. In the practical situation, the load threshold value defined according to the free resources of PM and the threshold defined according to the average of the comprehensive load strength are different. Therefore, L2 is set to be a smaller value, and θ is used to stand for the load difference generated by the PM. The equations of three threshold values are shown below.

$$L_1 = Load_{ave} - \sigma \tag{4}$$

$$L_2 = \min\{(1-\theta)Load_{res}, Load_{ave} + \sigma\} \tag{5}$$

$$L_3 = \max\{(1-\theta)Load_{res}, Load_{ave} + \sigma\} \tag{6}$$

Where, $\left|(1-\theta)Load_{res} - Load_{ave}\right| \leq \sigma$

In order to ensure the accuracy of evaluation, the task can be triggered when there are at least "x" values out of 'n' results meeting the above conditions. When $Load_{VM}[i] < L_1$, it suggests that the load of the PM is too small, and all VMs in the PM should be moved out and the PM should be closed to save energy consumption. When the server meets the condition of $L_1 < Load_{VM}[i] < L_2$, the host machine is regarded to be able to maintain the load balancing state, and the PM with normal load can be added to a set named NS without migration of VMs. When $L_2 <$ LoadVM[i]$<L_3$, the load of the PM is at the dangerous state, and it is vulnerable to be overloaded. Therefore, the PM is added into a set named DS for load prediction and further confirmation of whether it is necessary to migrate VMs. When $Load_{VM}[i] > L_3$, it means the PM is overloaded, and the PM can be added into a set named OS. Then appropriate VMs can be chosen for migration.

Load prediction

Through load prediction, this paper classifies PMs according to their load situations. Physical machines at a dangerous status might operate normally at the moment, but might be overloaded at the next moment. However, the overload might be a temporary danger, and the PM can return to its normal load at the next moment. Therefore, in order to avoid unnecessary migrations or omission of migrations caused by sensitiveness and bluntness, this paper adopts the exponential smoothing model based on the time series trend prediction to conduct load test of the set of PMs at a dangerous situations, and compare the predicted value with the threshold value. If the predicted value exceeds L_2, the VM can be added into the set named OS for migration. The equation of the predicted value at the moment of't+1' is shown below:

$$L(t+1) = RL(t) + R^2 L(t-1) + R^3 L(t-2) + \cdots + R^M L(t-M+1) + C \tag{7}$$

Where, t stands for the current operation moment; parameter RM stands for the influence parameter of the load value at the moment of "t-M" on the predicted value at the moment of 't+1'; C stands for the error. The equation predicts through the exponential decrease method, namely by endowing the load close to the prediction time with a larger weight.

Selection of VMs to be migrated

This process refers to selecting some proper VMs for migration from the set named OS to ensure load balancing. However, what is the optimal Virtual Machine? The major criteria to judge the migration quality is using minimum migration times and shortest migration duration and reducing the utilization rate of the PMs. In the current strategies, lots of them migrate the VM whose CPU utilization rate is highest. Since CPU, memory and network bandwidth resources of the VM might impose certain influence on load, load balancing cannot be achieved most efficiently if only the utilization rate of CPU is taken into consideration. Therefore, according to the utilization rate changing trend of the CPU, memory and bandwidth of the VM, this paper calculates the influence of the three resources on the load situation of the PM. The higher the value is, the greater the resource affects the load of the PM.

$$CPU_{trend} = \frac{CPU[t] - CPU[t-1]}{CPU[t]} \tag{8}$$

$$MEM_{trend} = \frac{MEM[t] - MEM[t-1]}{MEM[t]} \tag{9}$$

$$NET_{trend} = \frac{NET[t] - NET[t-1]}{NET[t]} \tag{10}$$

Where, CPU[t], MEM[t] and NET[t] stand for the average usage amount of CPU, memory and network bandwidth of VM at the moment respectively. CPU[t-1], MEM[t-1] and NET[t-1] stand for the average usage amount of CPU, memory and network bandwidth of VM at the previous moment respectively.

Research found that [4], no matter whether the load is triggered by CPU or memory, migration should comprehensively considerate the utilization rate of CPU and memory of the VM that occupying the host. However, when the VM migration triggered by the bandwidth, only VMs with a huge consumption of bandwidth need be considered and chosen for migration [4].Therefore, this paper puts forward a selection strategy taking migration effects and times into comprehensive consideration. The equation is shown below.

$$U = a \cdot \frac{CPU_{trend}}{MEM_{trend}} + (1-a) \cdot \frac{1}{1-CPU_{trend}} \cdot \frac{1}{1-MEM_{trend}} \cdot \frac{1}{1-NET_{trend}} \qquad (11)$$

Where, a stands for the weight factor. When the utilization rate of the CPU of the PM is high or the utilization rate of the memory is low, “a” is endowed with a large value, and the VM with the smallest U value is migrated to quickly achieve the load balance of the PM. Under the other situations, “a” is endowed with a small value. Based on a comprehensive consideration of the influence of the three factors, appropriate VMs can be chosen for migration.

Selection of target hosts

NS, the set of PMs whose load is evaluated to be normal, is the major target host for VMs to be migrated. According to the available free resources, this paper works out the probability of a PM to be chosen, and rank the probability in a descending order. Those with a higher selection probability should be added into the set named MS. Then calculate and rank the energy consumption of the selected PMs in the set named MS. The host with the minimum energy consumption is defined as the target host, and the VMs will be migrated into it.

The probability can be worked out based on the free resources, RES_{free}, of physical nodes. This suggests the higher the probability is, the more the free resources of the target host is, and the easier for the host to meet demands of the VMs to be migrated about different resources. Therefore, higher probability means chances are higher for VMs to be migrated to be selected as target hosts. The probability is expressed as P_j.

$$P_j = \frac{RES_{free}[j]}{\sum_{j=1}^{|v^s|} RES_{free}[j]} \qquad (12)$$

The free resources in the equation can be replaced with resources influencing migration to calculate the probability. For example, in terms of migration triggered by a high utilization rate of CPU, the probability can be worked out by putting the required CPU resources of VMs to be migrated and the CPU free resources of target hosts into the above equation. Physical hosts with a high probability can be selected to ensure the resource supply for hosts after migration.

Calculate the energy consumption of PMs in the set named MS, and define the one with the minimum energy consumption as the target PM.

$$E = \int_t [k \cdot P + (1-k)P \cdot u] \qquad (13)$$

Where, k stands for the percentage of energy consumption when the server is idle in energy consumption at full load; P stands for the power constant of the server at full load, which can be statistically measured; u stands for the utilization rate of CPU in the target host. The equation calculates the linear relationship between energy consumption and the utilization rate of CUP to confirm the energy consumption of the physical cost. The lower the energy consumption is, the better the selected one is. After the selection is confirmed, migration of VMs can be started.

The result and analysis of experience

Since CloudSim Toolkit supports the large-scale experiment simulation, this paper adopts it to build the simulated environment and simulate migration strategies based on the algorithm put forward in this paper, the classic Elastic Load Balancing (ELB) algorithm [5] and Double Threshold algorithm. They are compared to verify the feasibility and advantages of the migration method for VMs put forward in this paper. Table 1 shows the setting of experiment parameters of PMs and virtual hosts. The total number of tasks is 1,000.

Name	Number	CPU capacity/MIPS	Memory/GB	Network bandwidth/MLs^{-1}
PM	50	2000,300, 6000	2,4,8	60,100,150
VM	1000	250,500,1000,1500	0.5,0.75,1,1.5	5,10,20,35

Table 1. Experimental parameters of PM and VM

In order to evaluate the three migration strategies, this paper measures the load balance degree that the three algorithms can achieve through the load balancing factor. In the experiment, the generation period of new tasks is set to be 5s, and the task objectives and the number of tasks are randomly distributed. Below is the common equation for the load balancing factor:

$$\varphi = \sqrt{\frac{\sum_{k=1}^{n}(L_k - \bar{L}_k)^2}{n-1}} \tag{14}$$

Where, L_k stands for the load balancing factor of the VM, namely the weighting value of the load of the VM; $\overline{L}_k$ stands for the average load. In order to ensure the high load balance of the system, the smaller the standard deviation of the loan balance is, the better. In other words, the smaller the load balancing factor is, the better the dynamic adjustment of migration is. The result of experiment is shown as figure 1.

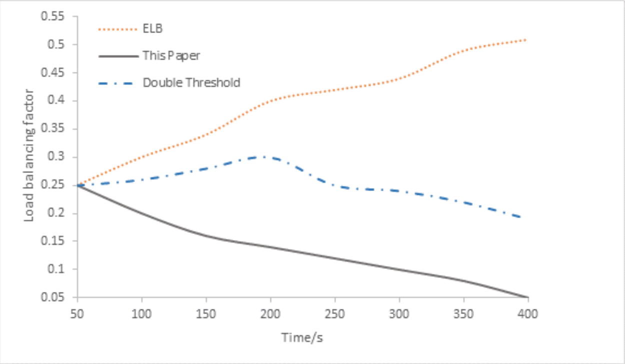

Fig.1. Experimental results

From figure 1, it can be seen that the load balancing factor of ELB algorithm increases along with the passage of time. Besides, during the whole adjustment process, the balancing factor is far higher than that of the other two adjustment methods. This suggests the algorithm's load balancing effect is not ideal. The balancing factor in the other two algorithms decreases along with the passage of time. This states that these two algorithms can efficiently reduce the load of the PM to finally achieve load balance. However, the balancing factor in Double Threshold algorithm reaches the peak around 200s, suggesting that Double Threshold algorithm might cause the decrease of the load balancing capability of the VM for some reasons during the random distribution of initial tasks. However, the adjustment effect of the algorithm put forward in this paper is more significant and takes the least time. During the scheduling process, the load balancing factor keeps on falling, suggesting that the load balancing capability of the PM based on the algorithm put forward in this paper is the best

Conclusion

In order to achieve dynamic migration of VMs, reduce unnecessary migrations and realize the purpose of load balancing, this paper puts forward a VM migration strategy based on load evaluation and prediction. Simulation results suggest that, compared with the previous migration algorithms, the algorithm put forward in this paper are better in terms of its dynamic adjustment capability and capability of maintaining load balance of VMs, so it is more feasible.

Acknowledgement

Think China Waterborne Transport Research Institute for providing the fund. This work was financially supported by the Shanghai Natural Science Foundation (0666666), Innovation Program of Shanghai Municipal Education Commission (060000) and Shanghai Leading Academic Discipline Project of Shanghai Municipal Education Commission (0555555).

References

[1] S Anbazhagan, K Somasundaram. Cloud computing security through symmetric cipher model [J]. International Journal of Computer Science & Information Technology, 2014, 6(3).

[2] Mauro Andreolini, Sara Casolari, Michele Colajanni. Dynamic load management of VMs in a cloud architecture [J]. Department of information Engineering, 2010:201-204 (in Chinese).

[3] Zeng Longhai, Zhang Bofeng. Research on construction technology of Virtual cluster based on cloud environment [J]. Microelectronics and computer. 2010:31-35 (in Chinese).

[4] Hu Liting, Jin Hai, Liao Xiaofei, et al. Magnet: a novel scheduling policy for power reduction in cluster with VMs[C]//Proceedings of 2008 IEEE International Conference on Custer Computing. Tokyo: IEEE Computer Society, 2008: 13-20 (in Chinese)

[5] WU He-Sheng，CJ Wang，JY Xie. TeraPELB-an Algorithm of Prediction-based Elastic Load Balancing in Cloud Computing[J]. Journal of System Simulation, 2013, 25(8):1751-1750