Journal of Computing and Electronic Information Management

ISSN: 2413-1660

Study on Load Balancing Algorithm Based on Metabolic GM (1,1) Prediction Model in Cloud Computing

Jianhua He, Yan Peng *, Xia Long, Hui Huang
College of Computer and Science, Sichuan University of Science and Engineering, Zigong 643000, China

Abstract: Because of the difference of computing resources required by different cloud computing tasks, Cloud computing providers always cannot offer the resources for customers accurately. The problems existing in cloud data centers such as waste and shortage of computing resources result in excessive energy saving and high Service-Level Agreement (SLA) violation rate, which seriously affects the service quality of cloud computing providers. The traditional virtual machine (VM) scheduling algorithm is passive scheduling for the overloaded physical host. In this case, the computing resources are seriously insufficient, which will affect the migration and cause longer migration time. Considering the above issues, this paper proposes a load prediction and migration algorithm based on the metabolic GM (1,1) model, which realizes load prediction of the physical host in advance, and performs the early migration operation according to the workload prediction. The experimental results show that the use of the prediction algorithm for scheduling can effectively reduce data center's energy consumption, VM migration time and SLA violation rate.

Keywords: Cloud computing; metabolic GM (1,1) model; load prediction; VM migration; load balancing.

1. Introduction

Infrastructure as a Service[1] (IaaS) is a service model in cloud computing. Customers can lease software and hardware resources including servers, storage, networks, firewalls etc. from cloud computing offers to build server platforms that meet their needs. IaaS virtualizes hardware resources to achieve dynamic demands of different virtual machines (VM), and improves service quality and reduces service costs by dynamically adjusting VM served in physical hosts. Its dynamic adjustable resource service model determines that service providers need to build VM on physical machines to serve customers. Therefore, the VM migration and other operations can meet the

requirement of cloud computing physical servers, such as load balancing, green energy saving and service quality.

Service quality and energy consumption in a cloud computing environment are important ways to demonstrate cloud environment service capabilities. VM migration is an important way to achieve dynamic resource allocation and reduce energy consumption[2]. Compared with static migration, dynamic migration becomes the mainstream migration method due to its shorter downtime and more flexible resource allocation. In recent years, the research hotspots are providing high-quality services while reducing energy consumption, reducing the number of migrations, and maximizing the use of computing resources.Wu Xiaodong[3] et al. proposed a dynamic prediction algorithm based on static threshold, it analyzes the running data of the VM to estimate the CPU computing performance required by the VM, and combines the static threshold to predict the load of the host and the migration opportunity of the VM, then achieved the VM dynamic migration. Gmach D[4] proposed a strategy for predicting demand at the next moment by using resource pool management, and effectively utilizing resources when realized a large number of services through the tracking-based capacity management approach. Xu Jing[5] et al. modeled VM allocation in the data center as a multi-objective optimization problem, and used group genetic algorithm to achieve efficient use of multi-dimensional resources and solved possible conflicts in resource mapping. Guo Zhenghong[6] et al. proposed an online migration strategy based on the analytic hierarchy process weights and host load grey prediction. This strategy used different resource weighting methods to determine the load of the VM and implements the online migration strategy of the VM under different cloud tasks. All of the above work studied migration strategies of VM in data center, but most of the studies focused on energy consumption model, which could not balance the relationship between migrated VM and cloud computing service quality, and it was difficult to ensure migration under the premise of optimal service quality. In addition, although the existing researches consider the dynamic adjustment of computing power allocated by VM through load forecasting, most of them only study how to achieve prediction, without in-depth consideration of the prediction algorithm applicable to the actual load, not considered the decline of service quality caused by actual migration too.

Cloud computing data center should reduce energy consumption on the premise of ensuring the quality of service, so data center must prioritize the provision of sufficient computing resources. The cloud computing tasks in the data center have the characteristics of short execution period and fast calculation task request, the load of the host machine changes frequently, showing small amount of data and fast iterative features. Therefore, there is a need for a load prediction algorithm that can quickly and

accurately predict the load of the next cycle using a small amount of data, thereby providing a reliable decision basis for migration scheduling. The grey system theory proposed by Deng Julong took "small sample" and "poor information" uncertain information system with "partial information known and partial information unknown" as research objects. The theory realized the prediction and description of system evolution by analyzing valuable information of "partial" known information[7]. The metabolic GM (1,1) prediction model continuously corrects the error caused by the original GM (1,1) prediction model using the latest actual data generated continuously. Combined with the load change characteristics of the data center, the metabolic GM(1,1) prediction model has better prediction effect than other prediction models. According to the above analysis, this paper proposes a load prediction algorithm for cloud computing host based on metabolic GM(1,1). Based on the static migration trigger threshold, this paper analyzes the host historical load and predicts the load condition of the host for the next moment. In this way, low-load virtual machines in host machines that may run above and below the threshold can be migrated in advance, leaving computing resources to high-load computing tasks, reducing energy consumption and SLA violation rate while guaranteeing service quality.

2. 2. Grey prediction model

2.1 GM(1,1) prediction model

The raw series is $X^{(0)}=(x^{(0)}(1),x^{(0)}(2),\cdots x^{(0)}(\mathrm{n}))$, $k=1,2,3,\cdots \mathrm{n}$; let $X^{(1)}$ be 1-AGO series of $X^{(0)}$, $X^{(1)}=\left(x^{(1)}(1),x^{(1)}(2),\cdots,x^{(1)}(n)\right)$, $x^{(1)}(k)=\sum_{i=1}^{k}x^{(0)}(i),k=1,2,\cdots,n$; $Z^{(1)}$ is said to be MEAN series of $X^{(1)}$, $Z^{(1)}=\left(z^{(1)}(2),z^{(1)}(3),\cdots,z^{(1)}(n)\right)$, $z^{(1)}(k)=\frac{1}{2}\left(x^{(1)}(k-1)+x^{(1)}(k)\right),k=2,3,\cdots,n$

Definition 2.1.1 The raw series is $X^{(0)}$, let $X^{(1)}$ be 1-AGO series of $X^{(0)}$, $Z^{(1)}$ is said to be MEAN series of $X^{(1)}$, so

$$x^{(0)}(k)+az^{(1)}(k)=b \tag{1}$$

is to be called the grey GM(1,1) model.

Definition 2.1.2 The raw series is $X^{(0)}$, let $X^{(1)}$ be 1-AGO series of $X^{(0)}$,

$$\frac{dx^{(1)}}{dt}+ax^{(1)}=b \tag{2}$$

is to be called the white equation of grey GM(1,1) model.

Theorem 2.1.1 The raw series is $X^{(0)}$, let $X^{(1)}$ be 1-AGO series of $X^{(0)}$, $Z^{(1)}$ is said to be MEAN series of $X^{(1)}$, $\alpha=(a,b)^T$ are parameters, in which

$$B=\begin{bmatrix} -z^{(1)}(2) & 1 \\ -z^{(1)}(3) & 1 \\ \vdots & \vdots \\ -z^{(1)}(n) & 1 \end{bmatrix} \quad Y=\begin{bmatrix} x^{(0)}(2) \\ x^{(0)}(3) \\ \vdots \\ x^{(0)}(n) \end{bmatrix} \tag{3}$$

Then:1) Using the least squares method to solve the parameters

$$a=(B^TB)^{-1}B^TY \tag{4}$$

2) The solution of white equation is

$$x^{(1)}(\mathrm{t})=(x^{(0)}(1)\text{-}\frac{b}{a})e^{-at}+\frac{b}{a} \tag{5}$$

3) The time response function of GM (1, 1) model is

$$x^{(1)}(k+1)=(x^{(0)}(1)-\frac{b}{a})e^{-ak}+\frac{b}{a} \tag{6}$$

4) Restore value is:

$$\begin{aligned} \hat{x}^{(0)}(k+1)&=a^{(1)}\hat{x}^{(1)}(k+1) \\ &=\hat{x}^{(1)}(k+1)-\hat{x}^{(1)}(k) \\ &=\left(1-e^{a}\right)\left(x^{(0)}(1)-\frac{b}{a}\right)e^{-ak},\quad k=1,2,\cdots,n \end{aligned} \tag{7}$$

2.2 Metabolic GM (1,1) prediction model

As time goes by, in the development of any grey system, there will be some random disturbances or driving factors entering the system, which will affect the development of the system. Therefore, with the GM(1,1) model, only one or two cycles of data can be accurately predicted. The farther away from the time origin, the weaker the prediction significance of GM(1,1). In practical applications, it is necessary to constantly consider the disturbances or driving factors that successively enter the system over time, and put each newly obtained data into the time to establish a new information model for dynamic prediction. Along with the development of the system, the old data information will gradually reduce, at the same time of updating information, get rid of old information in time, the sequence modeling can reflect the characteristics of the system in the present. In addition, continuous information renewal can effectively avoid the difficulties of increasing the information, expanding computer memory, and increasing the number of modeling operations.

Definition 2.2.1 Take the new information $x^{(0)}(n+1)$ into the raw data series $X^{(0)}=\left(x^{(0)}(1),x^{(0)}(2),\cdots,x^{(0)}(\mathrm{n})\right)$,While removes the oldest information $x^{(0)}(1)$.Then, the new data series is $x^{(0)}(2),x^{(0)}(3),\cdots,x^{(0)}(\mathrm{n}+1)$. $GM(1,1)$ model is established based on definition 2.1.1. So according to the method, the prediction target is finished. The model is called the metabolic $GM(1,1)$ model.

2.3 Load prediction algorithm based on metabolic GM(1,1) model

In order to predict the load of the physical host through historical load data, the virtual machine monitor (VMM) needs to save the historical load data of the first n cycles. For all physical hosts in the data center, the following steps are used to perform load prediction for the next cycle.

step1: The VMM needs to save the historical load data of the first n cycles as the raw series $X^{(0)}$, $X^{(0)}=\left(x^{(0)}(1),x^{(0)}(2),\cdots,x^{(0)}(\mathrm{n})\right)$

step 2: According to the equation (7), the load of the physical host is calculated at the n+1 point;

step 3:Loop step 2 until getting all physical hosts' load data for n cycles.

Step 4: According to definition 2.2.1, metabolic GM(1,1) model is established;

Step 5: Continue to run step 4 until the predict load of all hosts are obtained;

Step 6: Send the predicted result to the VMM to determine the migration physical host and destination physical host.

3. Load balancing prediction migration algorithm based on grey model

The resource allocation of a virtual machine has two modes of allocation: pre-allocation and dynamic allocation. The traditional resource pre-allocation method cannot flexibly utilize cloud computing resources, which causes great resource waste for low-load computing tasks, and the high-load cloud task has the problem of insufficient computing power, so the resource pre-allocation method has been gradually eliminated. Dynamically allocate computing resources needed for real-time monitoring of cloud tasks, adjust cloud computing tasks served by the host, migration cloud computing tasks in overloaded physical hosts to other low-load physical hosts, reduce SLA violation rate and guarantee QoS service quality.

3.1 VM load metric evaluation criteria

The VMM detects the resource consumption data such as the CPU utilization, memory usage, and bandwidth occupancy of the VM in each physical host to determine the load of the corresponding host. Use Equation 8 to measure CPU utilization:

$$\mathrm{P}_{cpu}=\frac{\sum_{i=1}^{n}V_{cpu_i}}{Total_{cpu}} \tag{8}$$

Where V_{cpu_i} is CPU computing resource utilization from ith VM, $Total_{cpu}$ is the maximum CPU computing resources from the host, n is the number of virtual machines owned by the host.

3.2 Source server migration judgment

In this paper, the historical load of each host is counted by the VMM program. All source hosts are classified by analyzing historical load condition to predict the load condition of the next time source host.

1) If the host load prediction condition satisfies $U_{cpu} < Thr_l$, the host load is too small, the compute tasks can be moved out, and the host can be shut down to reduce power consumption. Then add this host to the low-load hosts' list.

2) If the host load prediction condition satisfies $Thr_l < U_{cpu} < Thr_m$, the host load is moderate, and the task does not need to be migrated. Then add such hosts from small to large load to moderately loaded hosts' list.

3) If the host load prediction condition satisfies $U_{cpu} > Thr_m$, the host is in a heavy load state, migrating some of the low-weight computing tasks to reduce the host load and ensure the quality of service. Then add this host to the high-load hosts' list.

3.3 Migration algorithm based on metabolic GM(1,1) prediction model

To achieve load balancing of cloud computing , the algorithm predicts the load of each host through the metabolic GM(1,1) prediction model at first, then divides the host into high load, medium load and low load according to the migration trigger threshold. And migrating the high load host and low load host.

The algorithm is implemented as follows:

Input: metabolic GM (1,1) prediction results, maximum load threshold $Thr_{\max}$, minimum load threshold $Thr_{\min}$.

Ouput: host load prediction list

Step1: Using the predicted data to classify the physical host for each physical host load. The various state hosts join the corresponding queue according to the load value from high to low.

Step2:The maximum load VM in the head of the over-load list will be migrated to the host in the tail of middle-load list, then the host will be removed from the corresponding list.

Step3: Step2 will be performed until the overload queue or medium load queue is empty. If the overload queue is empty and the medium load queue is not, jump to step 4. Otherwise, terminate the migration and turn to step 6.

Step4: All VM in the head of the low-load list will be migrated to the host in the tail of middle-load list, then the host will be removed from the corresponding list and shut down the low load host to reduce power consumption.

Step5: Loop step4, terminate the migration until the overload queue or medium load queue is empty. Turn to step6.

Step6: Wait for the migration signal, then perform step1 for a new migration.

4. Experiment and result analysis

4.1 Experimental environment and parameter description

This paper uses the sampling data of the PlanetLab platform at the sampling interval of the ten days from March 3, 2011 to April 20, 2011, which is provided by the CloudSim[9] simulation tool, as the load input of the cloud computing task CPU. CloudSim is an open source cloud computing simulation program based on Java developed by the University of Melbourne Cloud Computing and Distributed Systems Lab. The program allows the user to customize the scheduling scheme, the number of tasks and other key parameters, which suitable for users to schedule simulation.

Based on the reuse of the CloudSim framework, the key classes such as Power VmAllocation Policy MigrationAbstract and Power Vm Allocation Policy Migration Static Threshold are rewritten, and the function codes of the important function methods such as get Over Utilized Hosts By GM11, getHostOverUtilized, and is Hos t Over Utilized By GM11 are realized. Based on the threshold-based scheduling policies and the minimum utilization selection policies, a pre-migration optimization algorithm for load balancing in cloud computing based on metabolic GM(1,1) prediction model is realized too. At the same time, the sampling data of the PlanetLab platform is used as the virtual machine's CPU load input value. The emulator simulates a cloud computing center consisting of multiple physical hosts for cloud task processing. In order to achieve low SLA default rate and low energy consumption level, all computing resources consolidated by virtual machine monitors, are dynamically allocated to different cloud computing tasks (or "virtual machines") based on the requirements of the tasks (which are the standards for requesting computing resources). The performance of the algorithm was evaluated with SLA default rate, energy consumption level and total number of migrations. The parameters of some simulation experiments are shown in Table 1.

Table 1. The parameters of simulation experiments on April 20, 2011

Experimental parameters	Experimental value
number of physical hosts	800
number of cloud tasks	1033
total length of cloud tasks	216000000
number of virtual machines	1033
number of CPUs	2
CPU computing power of the physical host	{1860，2860}MIPs
CPU computing power of the virtual machine	{2500，2000，1000，500} MIPs
scheduling interval time	300

For the traditional static threshold-based migration triggering policy, in order to reduce

the load of the host, the migration policy is triggered by the physical host after it is overloaded. Such migration is a remedial migration. Because virtual machine migration needs to consume certain computing resources, the migration of physical host under high load can cause problems, which are too long migration time and low QoS service quality. But for metabolic GM (1,1) prediction model, it can predict the historical load of the physical host, perceive the load of the physical host at the next moment in advance, and perform pre-processing operations. So, This predictive approach can leave more computing resources to the virtual machine monitor for migration processing and reduce migration times and SLA default rates. Compared with the traditional static threshold-based migration policy, the improved algorithm proposed in this paper can effectively reduce the number of migrations by more than 80%, reduce the SLA default time by 50%, and reduce energy consumption by about 50%. The experimental results are shown in Figures 1 to 3.

As can be seen from the figure, the prediction-based virtual machine migration algorithm can effectively reduce the key indicators of cloud computing data centers such as migration times, SLA default rate and energy consumption by predicting the load of the physical host at the next moment. By using predictive algorithms, more efficient computing services can be provided by cloud computing services in the data center. Then, the migration time can be effectively reduced because sufficient computing resources are reserved for the high-load physical host to perform the virtual machine migration task.

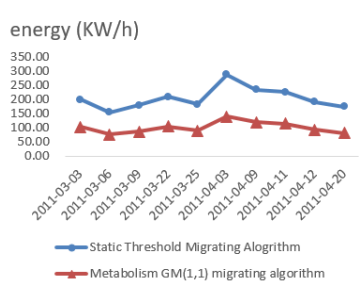

Figure1. Energy saving

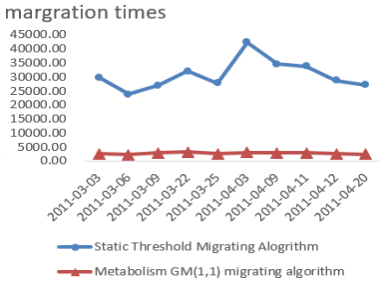

Figure2. Margration times

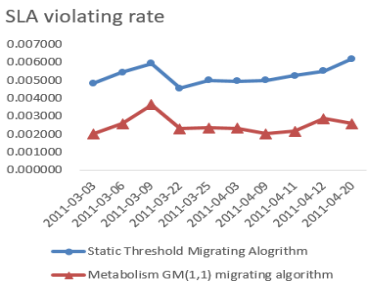

Figure3. SLA violate rate

5. Conclusion

It is an important way to ensure the quality of cloud computing services by the migration of virtual machines in cloud computing. According to the CPU, bandwidth, memory and other resources consumption of the virtual machine, the corresponding load situation of the host is obtained, and then the virtual machine served by each physical host is dynamically adjusted to the load of the physical host, so that the cloud computing provider can better provide computing service to customers.

The prediction-based virtual machine migration policy proposed in this paper can dynamically select the virtual machines that need to be migrated through the historical

load status of the physical host. The physical host that can place the virtual machine selected through the minimum utilization policy, which ensure the service provision capability of the high-load host, reduce unnecessary physical host usage, and reduce energy consumption.

The simulation result shows that compared with the static-based migration method, the method can reduce the number of migrations and reduce the migration times by ensuring the resource cost required for migration, it can ensure the smooth operation of the cloud computing data center.

Acknowledgements

Founding: The work was supported in part by the Science and technology plan of zigong science and technology bureau No.2018GYCX33.

References

[1]Vaquero L M, Rodero-Merino L, Morán D. Locking the sky: a survey on IaaS cloud security[J]. Computing, 2011, 91(1):93-118.

[2]Jin H，Deng L，Wu S，et al. MECOM：Live migration of virtual machines by adaptively compressing memory pages[J]. Future Generation Computer Systems,2014,38(3):23-35.

[3] Xiaodong W, Jianjun H. Threshold-based energy-efficient VM scheduling in cloud datacenters[J]. J. Huazhong Univ. of Sci. & Tech. (Natural Science Edition), 2018(9).

[4]Gmach D,Rolia J,Cherkasova L,et al. Workload analysis and demand prediction of enterprise data center applications[C].Proceedings of the 2007 IEEE 10th International Symposium on Workload Characterization, Washington, 2007:171-180.

[5] Xu J , Fortes J A B . Multi-Objective Virtual Machine Placement in Virtualized Data Center Environments[C]// 2010 IEEE/ACM Int'l Conference on Green Computing and Communications & Int'l Conference on Cyber, Physical and Social Computing. IEEE, 2010.

[6] Zhenghong G, Xinhua M, Anyi L. On-line Migration Strategy in Cloud Computing Based on AHP Weight and Gray SerVer Load Predicting[J]. Computer Measurement & Control, 2015.23(3):1002-1004.

[7]Sifeng L, Yaoguo D, Zhigeng F, et al, Grey system and its application. Science press, Beijing, 2010.

[8]Li H . Virtual machine allocation method based on gray correlation degree in cloud computing. Journal of Computer Applications, 2014, 34(8): 2252-2255.

[9]Calheiros R N,Ranjan R,Beloglazov A,et al. CloudSim:a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms[J]. Software: Practice and Experience(SPE),2010,41(1):23-50

[10]Mignotte M.How to share a secret[C]//LNCS 149：Proceedings of the Workshop on Cryptography, 1983:371-375.