Acta Technica *61*, No. 4B/2016, 261–270

A loading balance model of virtual machine live migration in cloud computing environment[1]

Xin Sui[2,3], Li Li[2,4], Dan Liu[2], Hong Wei Yang[3], Xu Di[2]

Abstract. Virtual machine living migration technique provided a method for load balancing in cloud computing environments. In order to reduce the SLAV rate and improve server utilization rate, the virtual machines of overload server were migrated to other server and integrated low load server. The model proposed in this paper put the number of over-lowload servers, migration cost of virtual machine and power consumption of servers as the main target. In order to achieve these objectives, the paper defined the power consumption function of the servers, the migration cost function of virtual machines and the SLAV rate function, putting forward an ELBMLM model, and genetic algorithm was applied in the model. The experiment results showed that the proposed model was better than the others.

Key words. Cloud computing, virtual machine migration, load balancing, energy consumption.

1. Introduction

As a service available to the user, the stability of service resource and utilization rate is the key of influencing the intuitive feeling of users and economic benefits of service provider. So how to meet the needs of users and effectively manage the service resources becomes the key to the development of cloud computing [1].

[1] This work was supported in part by the project of "12th Five-Year" planning of the Education Science of Jilin Province (GH14049), the science and technology project of "13th Five-Year" planning of the Education Department of Jilin Province (2016), the key project of "13th Five-Year" planning of the Education Science of Jilin Province (ZD16024), and Key Science and Technology Project of Jilin Province (20160204019GX).

[2] Changchun University of Science and Technology, College of Computer Science and Technology, 130022, China

[3] Jilin Provincial Institute of Education, 130022, China

[4] Corresponding Author, `ll@cust.edu.cn`

Verma proposed an algorithm to solve the problems of virtual machine se lection (MADLVF), which would select the appropriate virtual machine from the overloaded host, migrate it to other server and put the load of the entire data center to tend into balance [2]. Jiao Zhang studied how much bandwidth was required to ensure the migration time and downtime during the migration of virtual machines [3]. In order to guarantee the service quality and reduce the cost, Mohamed Mohamed proposed an autonomic management model, which could effectively control the cloud resources, see [4].

2. Methodology

2.1. Problem formulation

2.1.1. Power consumption. From the document [5], the power consumption value of the physical server had a relationship with CPU usage rate at a certain time, which was presented as an approximate linear relationship. So the power consumption of the physical server could be calculated according to the CPU usage rate. The formula reads

$$P_i(u) = r_i * P_i^{\max} + (1 - r_i) * P_i^{\max} * u_i \,, \tag{1}$$

where $P_i(u)$ is the total power consumption of the physical server i, constant r_i stands for the power consumption ratio of the physical server i when it is idle and peak, $P_i^{\max}$ represents the maximum power consumption by fully utilized server, and u_i denotes the CPU usage rate of the physical server i.

The total power consumption of the physical server i between times t_1 and t_n is then given as

$$E_i = \sum_{t_1}^{t_n} P(u_i(t_j)) \,, \tag{2}$$

where $u_i(t_j)$ is the CPU usage rate of the server i at time t_j and $P(u_i(t_j))$ is the power consumption of the server i at time t_j.

The document research showed that the energy consumption of the network equipment was only related to the configuration of network devices. The calculation formula of network equipment energy consumption is shown as follows

$$P(C) = F(C) + A * X \,, \tag{3}$$

where $P(C)$ stands for the power consumption of the network equipment, C is the configuration parameters of the network equipment, $F(C)$ is the sum of standard and network line card energy consumption of network equipment. Symbol A denotes the power consumption of the network interface and X represents the number of the network interfaces.

The energy consumption of cloud data center is

$$E = \sum_{1}^{N} E_i + \sum_{t_1}^{t_n} P(C)\,, \tag{4}$$

where $N(i)$ is the number of servers.

2.1.2. Live migration cost estimation. The living migration of virtual machine would lead to the performance degradation of application at the pre-copy stage and down time at the copy phase. The migration time of virtual machine is calculated by the formula

$$t_{V_i} = \frac{M_{V_i}}{B_{V_i}} \tag{5}$$

and the performance decline of the application is given by the formula

$$P_{V_i} = a \cdot \sum_{t_0}^{t_0+t_{V_i}} U_{V_i} + b \cdot M'_{V_i} + c \cdot B_{V_i}\,. \tag{6}$$

Here, t_{V_i}, represented the migration time, M_{V_i} denotes the memory size of the virtual machine, B_{V_i} stands for the network bandwidth, P_{V_i} shows the performance decline of the physical server, t_0 is the starting migration time, U_{V_i} denotes the CPU usage rate, a, b and c are the parameters of the performance degradation caused by the CPU utilization, memory occupancy and network band occupancy, and, finally, M'_{V_i} represents the memory occupied by the live migration.

2.1.3. Service-level agreement (SLA) violation. The violation of the SLA percent is given by the formula

$$\mathrm{SLAVP} = \frac{N_{\mathrm{SLAViolation}}}{N_{\mathrm{total}}}\,, \tag{7}$$

where SLAVP denotes the percent of servers that were in violation of the SLA, $N_{\mathrm{SLAViolation}}$ stands for the number of servers in violation of the service level agreement and N_{total} is the total number of servers.

2.2. Live migration algorithm

2.2.1. Load status of server. The status of the ith server may be described as follows

$$\left\{ \begin{array}{ccc} P_i^{\mathrm{o}} = 1 & | & P_i^{\mathrm{f}} = P_i^{l} = 0, U_i > a \\ P_i^{\mathrm{f}} = 1 & | & P_i^{\mathrm{o}} = P_i^{l} = 0, b \le U_i \le a \\ P_i^{l} = 1 & | & P_i^{\mathrm{o}} = P_i^{\mathrm{f}} = 0, U_i < b \end{array} \right\}. \tag{8}$$

In (8), P_i^{o}, P_i^{f}, P_i^{l} express that the ith server is on overload, full load, low load state or not, respectively, and meets the condition $P_i^{\mathrm{o}} + P_i^{\mathrm{f}} + P_i^{l} = 1$. Symbol U_i indicates CPU utilization, and a and b are the CPU utilization bounds of the full load state.

2.2.2. Virtual machine migration model. The objective functions are

$$F_1 = \min \sum_{i=1}^{n} (P_i^{\mathrm{o}} + P_i^{l}) , \tag{9}$$

$$F_2 = \min \sum_{i=1}^{m} P_{V_i} , \tag{10}$$

$$F_3 = \min \sum_{i=1}^{n} E_i . \tag{11}$$

The constraint conditions of migrating virtual machines are

$$\left\{ \begin{array}{c} \sum_{j=1}^{m} V_j^{\mathrm{cpu}} \cdot P_{i,j} < C_i^{\mathrm{cpu}} \\ \sum_{j=1}^{m} V_j^{\mathrm{mem}} \cdot P_{i,j} < C_i^{\mathrm{mem}} \\ \sum_{j=1}^{m} V_j^{\mathrm{store}} \cdot P_{i,j} < C_i^{\mathrm{store}} \end{array} \right\} , \tag{12}$$

$$\sum_{i=1}^{n} P_{i,j} = 1, P_{i,j} \in \{0,1\} . \tag{13}$$

Formula (12) suggests that the sums of CPU, memory and hard disk for all virtual machines on one server, respectively, are smaller than the capacity of the server's CPU, memory and hard disk, where m is the number of virtual machine migration and n is the number of server integration. Symbols V_j^{cpu}, V_j^{mem} and V_j^{store} are, respectively, the need of CPU, memory and hard disk. Similarly, C_i^{cpu}, C_i^{mem} and C_i^{store} are the corresponding capacities. From formula (13), $P_{i,j}$ indicates whether the jth virtual machine is migrated to the ith server or not, and its value is 0 or 1, accordingly.

2.2.3. Virtual machine migration algorithm. In order to solve the virtual machine migration model, a global optimization genetic algorithm was designed and the algorithm included the encoding method, crossover operator, mutation operator, correction operator and local search operator.

- Encoding method: Assuming that the number of virtual machines needed to be migrated in the data center is m, the number of servers needed to be consolidated is n, the mapping of virtual machine and server can be expressed by an $n \times m$ matrix

$$P = \begin{bmatrix} P_{1,1} & P_{1,2} & \dots & P_{1,m} \\ P_{2,1} & P_{2,2} & \dots & P_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ P_{n,1} & P_{n,2} & \dots & P_{n,m} \end{bmatrix} .$$

Initial matrix P is the all zero matrix. From the formula (13), there was only one number on one column of the matrix, which was equal to 1, and the other were all 0.

- Crossover operator: First, four integers a, b, c, d were generated randomly: $a, c \in [1, N]$ and $b, d \in [1, M]$. The cross method then was

$$\begin{aligned} P1 &= (P^1_{1,1}, ..., P^1_{a,b-1}|P^1_{a,b}, ..., P^1_{c,d}|P^1_{c,d+1}, ..., P^1_{n,m})\,, \\ P2 &= (P^2_{1,1}, ..., P^2_{a,b-1}|P^2_{a,b}, ..., P^2_{c,d}|P^2_{c,d+1}, ..., P^2_{n,m})\,, \\ P3 &= (P^1_{1,1}, ..., P^1_{a,b-1}|P^2_{a,b}, ..., P^2_{c,d}|P^1_{c,d+1}, ..., P^1_{n,m})\,, \\ P4 &= (P^2_{1,1}, ..., P^2_{a,b-1}|P^1_{a,b}, ..., P^1_{c,d}|P^2_{c,d+1}, ..., P^2_{n,m})\,. \end{aligned}$$

- Crossover operator: First, random generation of a real number $a \in [0, 1]$. If $a \leq P_m$, then jump to the following step. Choice of a column of the matrix, and random generation of integer $b \in [1, N]$, $P_{i,j} = 1, b \neq i$, make $i = b$, $P_{b,j} = 1, P_{i,j} = 0$. Repetition of the previous step from the first to the mth column.

- Correction operator: Let

$$\mathrm{CPU^e} = \sum_{j=1}^{m} (V_j^{\mathrm{cpu}} \cdot P_{i,j}) - C_i^{\mathrm{cpu}}\,,$$

$$\mathrm{MEM^e} = \sum_{j=1}^{m} (V_j^{\mathrm{mem}} \cdot P_{i,j}) - C_i^{\mathrm{mem}}\,,$$

$$\mathrm{Store^e} = \sum_{j=1}^{m} (V_j^{\mathrm{store}} \cdot P_{i,j}) - C_i^{\mathrm{store}}\,,$$

  if $\mathrm{CPU^e} \leq 0 \&\& \mathrm{Mem^e} \leq 0 \&\& \mathrm{Store^e} \leq 0$ return $P^{'}$
  else select a column in the matrix $P_{i,j} = 1$, randomly generate integer $a \in [1, N]$, $a \neq i$, and make $P_{a,j} = 1, P_{i,j} = 0$ satisfying the formula (12)
  endif.
  The last step is repeated from the first to the mth column.

- Local search operator: Find the population individual P, which has the largest F_1 value, from the population after correction. Then choose a population, where $P_{i,j} = 1$, randomly generate an integer $a \in [1, N]$, $a \neq i$, make $P_{a,j} = 1, P_{i,j} = 0$, and generate population individual $P^{'}$.
  if $F_1^{'} < F_1$, go to the previous step; else return $P^{'}$, endif.

3. Results

3.1. Experiment parameter setting

The experiment simulated that the cloud data center consisted of 500 servers of different configurations and two types of virtual machines. Among them, 200 servers were HP Proliant ML110 G4 and 300 servers were HP Proliant ML110 G5. The energy consumption values of the two kinds of servers are shown in Table 1. Configuration parameter of the virtual machine are given in Table 2.

Table 1. Energy consumption of two types on different servers

Server	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
G4	86	89.4	92.6	96	99.5	102	106	108	112	114	117
G5	93.7	97	101	105	110	116	121	125	129	133	135

Table 2. Virtual machine parameters

Parameter	VM1	VM2
Number of CPU	1	1
Frequency of CPU	0.5GHz	1GHz
Memory	1GB	2GB
Disk	50GB	50GB

3.2. Experiment comparison results

3.2.1. Number of over-lowload servers. In the simulation process, when the number of requests for virtual machines increased, the number of servers in the over-lowload increased. The experimental result showed that the migration scheme proposed by this model could reduce the number of overloads and low loads to the minimum and the rate of growth to the slowest, which was compared with the other three models. The results are shown in Fig. 1.

3.2.2. Migration cost of virtual machines. Figure 2 shows the migration cost of virtual machines under four models with variable number of virtual machine requests. The experiment result indicates that the migration cost of ELBMLM model increased 6.6 times when the number of requests for virtual machines was increased 10 times. However, as the maximum migration cost model—the FRAware model—increased 7.34 times, and it showed that the proposed scheme could effectively reduce the live migration cost of virtual machines, which the model was compared with other three models, and it could reduce the total energy consumption of cloud data center.

3.2.3. Power consumption of cloud data center. The experiment result reflects the power consumption condition by increasing the number of requests for virtual machines. The total power consumption of the proposed scheme, which was com-

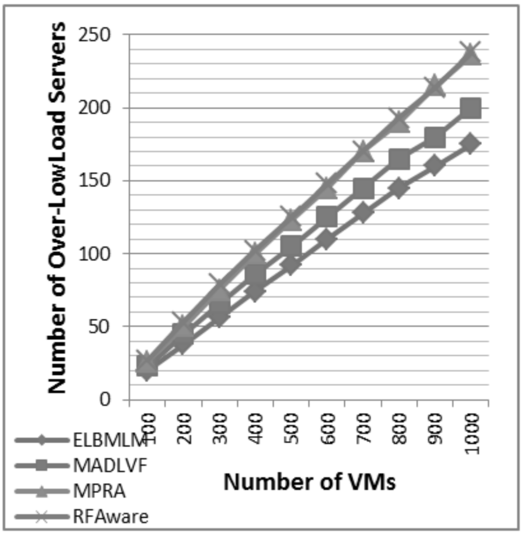

Fig. 1. Number of over-lowload servers

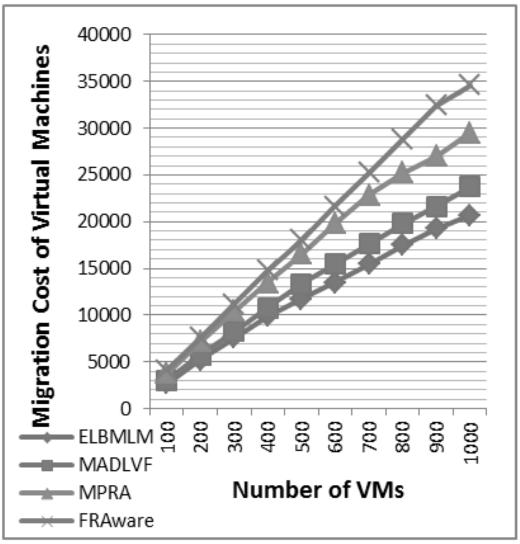

Fig. 2. Migration cost of virtual machines

pared with the other three models', was not the lowest but the second. In order to realize the minimum over-lowload servers' number and the cost of virtual machines' live migration, the power consumption of cloud data center must be influenced. The result is depicted as Fig. 3.

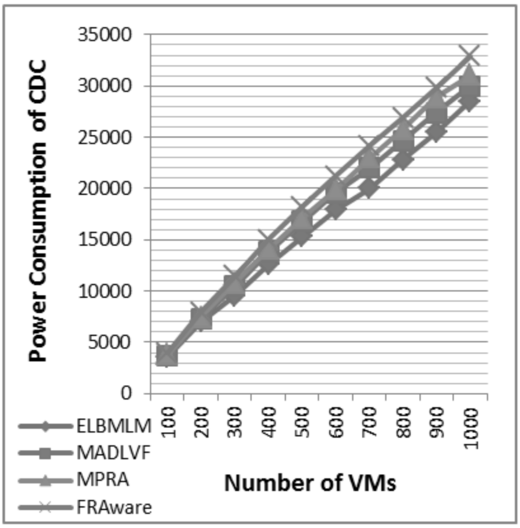

Fig. 3. Power consumption of CDC

3.2.4. Average violation rate of service level agreement. Figure 4 shows the average violation rate of service level agreement to the cloud data center under the variable number of virtual machine requests. The experiment result showed that four models' average violation rate of service level agreement with the increasing requests' number of virtual machine, and the violation rate of this model was among the four models', presenting a relatively stable state and effecting well.

4. Discussion

The proposed model compared with MADLVF, MPRA, and RFAware model in the simulation experiment not only took the cost of power consumption and live migration into consideration but also reduced the number of over-lowload servers and energy consumption of cloud data center to the minima.

It could be seen through the research, that the live migration of virtual machine could effectively regulate load balancing, prevent resource aggregation and finally reduce the energy consumption of data center. During the simulation experiment, this paper used genetic algorithm as optimization algorithm and this algorithm could fix ELBMLM model well.

5. Conclusion

The parameters of computation using the model proposed in the paper were compared with parameters of other three models on the number of over-lowload servers: migration cost of virtual machines, power consumption of servers and SLAV average

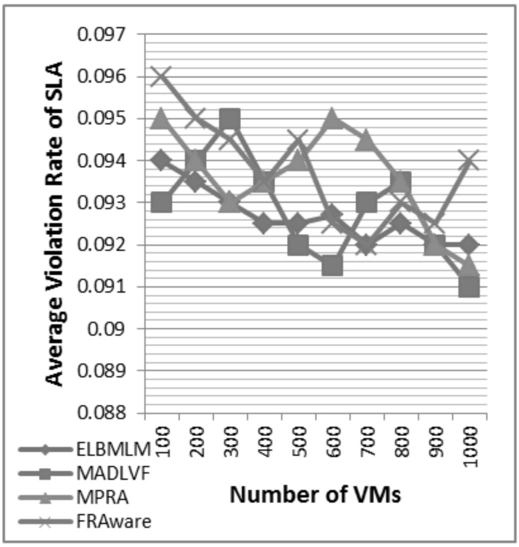

Fig. 4. Average violation rate of SLA

percent of four aspects. The results of experiment show that the proposed model is better than other three models both in the number of over-lowload servers and migration cost of virtual machines, which leads to savings in energy consumption.

References

[1] A. Beloglazov, J. Abawajy, R. Buyya: *Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing.* Future Generation Computer Systems *28* (2012), No. 5, 755–768.

[2] J. K. Verma, C. P. Katti, P. C. Saxena: *MADLVF: An energy efficient resource utilization approach for cloud computing.* IJ Information Technology and Computer Science *6* (2014), No. 7, 56–64.

[3] J. Zhang, F. G. Ren, C. Lin: *Delay guaranteed live migration of virtual machines.* Proc. IEEE Conference on Computer Communications, 27 April–2 May 2014, Toronto, Canada, 574–582.

[4] M. Mohamed, M. Amziani, D. Belaid, S. Tata, T. Melliti: *An autonomic approach to manage elasticity of business processes in the cloud.* Future Generation Computer Systems *50* (2015), 49–61.

[5] G. Chen, W. He, J. Liu, S. Nath, L. Rigas, L. Xiao, F. Zhao: *Energy-aware server provisioning and load dispatching for connection-intensive internet services.* Proc. Symposium on Networked Systems Design and Implementation (USENIX), 16–18 April 2008, San Francisco, California, USA 337–350.

Received November 16, 2016