Acta Technica *62*, No. 3B/2017, 75–84

The prediction algorithm in server resource utilization [1]

Dan Liu[2], Jilong Gong[2], Xin Sui[2,3], Yan Zeng[2], Zetian Zhang[2], Li Li[2,4]

Abstract. Cloud computing center is usually based on resource utilization and optimizing the task assignment, virtual machine migration and server load. So the servers' resource utilization is critical to the normal operation of the entire data center. The paper finds that monitor data combined with forecasting algorithm can predict resource utilization in a short period of time. The lag of the data center caused by real-time monitoring can be improved. During the simulation experiment, the paper selects the appropriate forecasting algorithm, models mathematical formula, gathers server's real-time resource utilization data, analyzes trend to resource utilization by using forecasting algorithm and compares the forecasting data with real-time monitoring data through visual analysis could more intuitively display experiment results and conclusions.

Key words. Cloud computing, virtual machine, resource utilization.

1. Introduction

Using virtualization technology, Cloud service operator could increase the system's physical resource utilization by mapping single physical server to multiple virtual machines and increase profits. The virtual machine scheduling process is divided into two phases:

1) Selecting the virtual machines' accommodation (placement problem) by analyzing the current task submitted by users and the amount of tasks.

2) When server resource is overused or low-used, the resources are integrated and virtual machines are migrated from one node to the other [1].

[1]This work was supported in part by the science and technology project of "13th Five-Year" planning of the Education Department of Jilin Province (JJKH20170631KJ); the key project of "13th Five-Year" planning of the Education Science of Jilin Province (ZD16024); 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

In order to achieve load balancing, the virtual machine resources are evenly distributed into the cluster during the virtual machines' placement. Although this method ensures the SLA protocol and QOS quality, but it reduces the utilization of physical resources [2]. According to literature [3], the super server operators' average physical resource utilization rate is not over 20 % which causes much waste of resources and power consumption. During the virtual machine scheduling, most enterprises take the pre-analytical method. When there are new tasks, the terminal console creates the new virtual machine based on the load balancing state to perform tasks. Although this guarantees the reliability and stability of service quality, but the core problem is physical resource waste and high energy consumption.

In most research works related with reducing energy consumption, scholars monitor the server resource utilization by setting the threshold. When the utilization takes above the threshold or below the threshold, the migration strategy is triggered to dynamically adjust the resource allocation. This scheduling strategy can optimize resource allocation and reduce system energy consumption in a certain degree [4–5], but most these strategies are obviously hysteresis. The paper found that the monitoring data combined with the forecasting algorithm can predict the resource utilization in a short period of time which optimizes the hysteresis caused by Data Center's real-time monitoring and re-processing.

Virtual machines' migration process would take up CPU, network and other resources. Because the memory copy process is in an iterative way during the migration [6]. This would have an impact on the performance of virtual machines and the applications running on them. According to literature [7], the key parameters which influence the migration costs were in-depth analyzed and the cost forecasting model was constructed. In literature [8], the migration evaluation was proposed by analyzing the impact of real-time migration on application performance. Therefore, in order to detect the early overload and trigger the migration, the forecasting approach is more fruitful than making predictions to avoid overloading during the migration process. The former is the need to address the real application workload and later is a temporary problem during the migration process. According to this problem, the paper proposed the forecasting resource utilization for determining the migration.

2. Problem formulation

The server's resource utilization includes processor resource utilization, runtime memory resource utilization, storage device resource utilization and network resource utilization. According to literature [8], the influence of server relation is: CPU>RAM>network>storage. Therefore, the CPU resource utilization forecasting has been main factor during paper's research.

Presently, the several common forecasting algorithms are:

1) Summary average method: a simple time series method which records a period data and estimates the future value by calculating the average of these values. Since this method does not respectively have a weighted average calculation of the near-term and mid-term values, the resource utilization rate forecasting is mainly based

on the recent value, so this method is not suitable for the experimental research.

2) Moving average method: a method which analyzes and predicts future values by collecting recently real-time values.

3) Exponential smoothing method: after analyzing Summary average method's disadvantages of equally use of all monitored values and Moving average method's ignoring the use of mid-term values, the paper gives a greater weight to recent-term value and takes a lower weight to the mid-term value. When the monitor values keep advancing, the weights of the previous values are infinitely approaching to zero [9]. Therefore, the paper chooses the third forecasting method as the basis during the simulation experiment process.

Nowadays, the virtual machine's migration is one of the core problems in resource allocation and management, At the same time, it is a contradictory process either. When the resource utilization exceeds the threshold, it helps to distribute the jobs on multiple physical machines to reduce the load. The tasks running on the virtual machine would halt during virtual machine's migration and would take up much network bandwidth. So the virtual machine's frequent migration would cause adverse consequences. Therefore, the forecasting algorithm researched in this paper caters the virtual machine migration technology concerned with the application of server resource utilization.

2.1. Forecasting algorithm

The Exponential smoothing method is proposed during the research in order to improve algorithm accuracy. The predicted value R_{t+1} at time t is obtained by one smoothing and the formula of second smoothing exponent is obtained by $R_{t+1}^{(2)}$. Then the predicted value can be obtained by forecasting formula (6). The specific process is as follows.

First step is the establishment of mathematical model and the basic formula of one smooth exponential function is

$$R_{t+1} = R_t + \lambda(W_t - R_t),\ (0 < \lambda < 1)\ . \tag{1}$$

Here, W_t is the actual monitored value at time t and R_t is the predicted value at time t. Value R_{t+1} is the predicted value at time $t+1$. $W_t - R_t$ is the prediction error. Quantity $\lambda(W_t - R_t)$ indicates that the prediction error is adjusted. λ is the smoothing factor. It is clear that the value of λ will affect the adjustment of prediction error. When λ has a value closer to 1, it reduces the smooth level, but when λ is closer to 0, the smooth level is increased. The value of λ will affect the fluctuation of the forecasting function.

The rewrite formula is

$$R_{t+1} = \lambda W_t + (1-\lambda)R_t\,. \tag{2}$$

Replaced R_t:

$$R_{t+1} = \lambda W_t + (1-\lambda)[\lambda W_{t-1} + (1-\lambda)R_{t-1}] = \lambda W_t + \lambda(1-\lambda)W_{t-1} + (1-\lambda)^2 R_{t-1}\,. \tag{3}$$

The hypothetical replacement of R_{t-1} components value:

$$\begin{aligned} R_{t+1} &= \lambda W_t + \lambda(1-\lambda)W_{t-1} + (1-\lambda)^2 R_{t-1} + \\ &+ \ldots + \lambda(1-\lambda)^{t-1} W_1 + (1-\lambda)^t R_1 \,. \end{aligned} \tag{4}$$

There are two important parameters in forecasting function. The first parameter is the initial value of forecasting system R_1. Since R_1 is unknown, the W_1 can be used as the initial forecasting value. The second parameter is λ and its choice takes a great impact on forecasting. This is the one smoothing exponent function finishing process.

The basic formula of secondary smoothing exponential function is

$$\begin{cases} R_{t+1} = \lambda W_t + (1-\lambda) R_t, \\ \\ R_{t+1}^{(2)} = \lambda R_{t+1} + (1-\lambda) R_t^{(2)}. \end{cases} \tag{5}$$

The value of $R_t^{(2)}$ is $R_t^{(2)} = R_1$ at the beginning of forecasting and the $R_t^{(2)}$ is obtained by the one exponential smoothing basis formula. This is an iterative process.

The formula of the second exponential smoothing forecasting algorithm is

$$X_{t+T} = x_t + y_t \cdot T \,. \tag{6}$$

The quantity X_{t+T} is the forecasting value of the future T period:

$$\begin{cases} x_t = 2R_{t+1} - R_{t+1}^{(2)}, \\ \\ y_t = \frac{\lambda}{1-\lambda} \cdot (R_{t+1} - R_{t+1}^{(2)}). \end{cases} \tag{7}$$

The algorithm implementation flows as follows:

1) Initialize the array $S1, S2$ (used to store real-time monitoring CPU resource utilization) and obtain the CPU resource utilization monitor data into array $S1, S2$.

2) Bring the parameter $\lambda W_t R_t$ into one of the exponential smoothing formulae (4) and obtain the exponent value R_{t+1}, then bring R_{t+1} into the secondary exponential smoothing formula(5) and obtain the second smoothing exponential value $R_{t+1}^{(2)}$.

3) Take R_{t+1} and $R_{t+1}^{(2)}$ into (7) and obtain X_t and Y_t. The forecasting value is obtained by bringing X_t and Y_t into formula (6).

The flow of the algorithm is shown in Fig. 1.

3. Results

In order to verify the migration process based on the predicted resource utilization, which can effectively reduce the energy consumption of the system, the simulation experiments are carried out in this paper. The experimental parame-

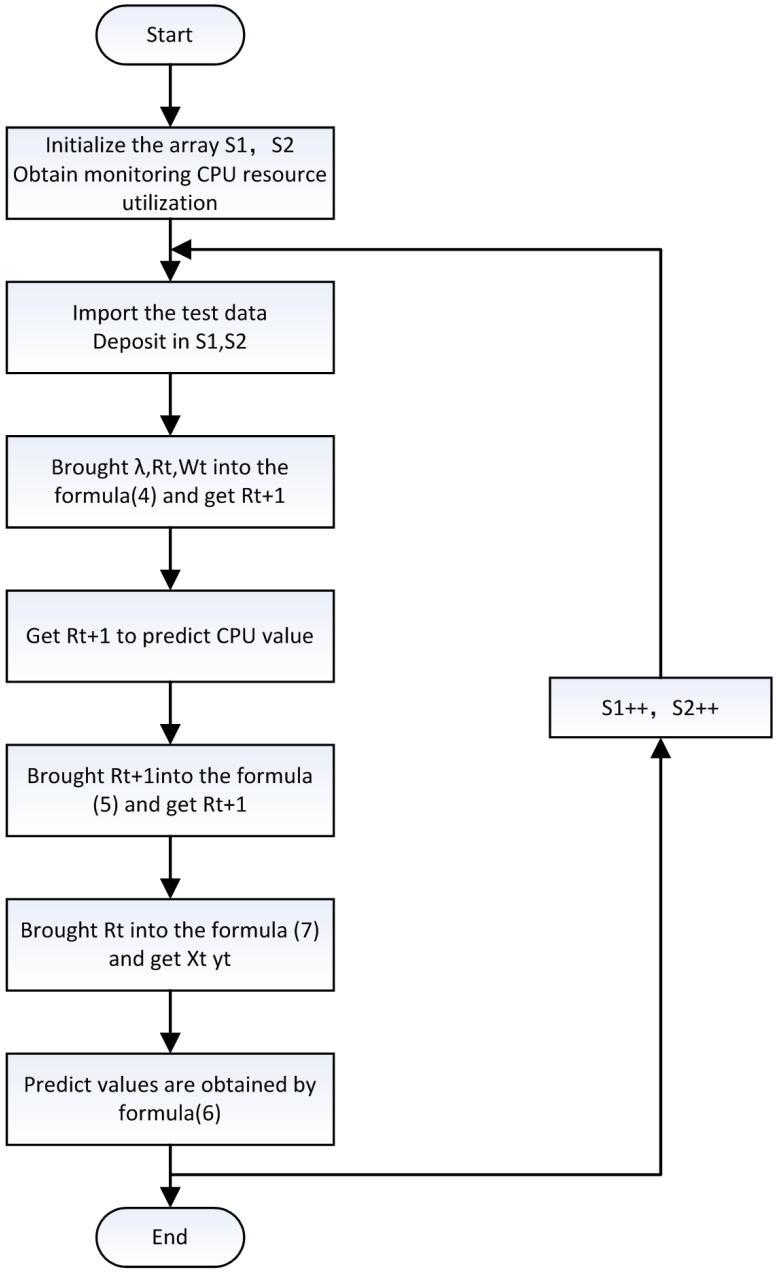

Fig. 1. The process of prediction algorithm

ters are shown in Table 1. The experimental configuration process: we installed Vm-Ware virtualization software to create three Linux virtual machine in the Windows10 operating system, then set them as ControllerNode, ComputeNode, and

NetworkNode, respectively. We install the calculation module, network module, security module and the graphical user interface Dashboard on the ControllerNode. In order to simulate cloud services platform environment, we can assign task request through Dashboard and divide virtual machines on Linux physical machines. Then we can install zabbix (a currently popular server monitoring alarm system) on the ControllerNode node to monitor the resource utilization of the server.

Table 1. Simulation experiment parameters

	Specific environmental parameters
Hardware environment	I7 CORE-4770, 8 G RAM, 1 T Hard Disk, 100 MB network bandwidth
Software environment	Vm-Ware, Ubantu, Openstack

3.1. Experimental procedure

1) In the actual operation process, the server mostly takes hours as the monitoring unit. But in the course of experiment, in order to ensure a certain amount of data, we use 30s as the monitoring unit to simulate the resource utilization load status of the server and record the sample.

2) Initially letting $R1 = W1$, we can get a set of predictive data. At the same time,we add a new round of monitoring values to continue the algorithm.

3) Since the choice of λ has a considerable impact on the forecast, two groups of comparative experiments were carried out in the course of the experiment. The first group of λ is 0.3, the second group of λ is 0.8.

4) Finally, the visualization data analysis software Tableau is used to analyze the experimental data and display.

3.2. Analysis of experimental results

The forecast algorithm has a certain hysteresis, the size of λ also has a certain effect on the hysteresis. The closer the value of λ is to 1, the larger the change range of the forecast function is, the more sensitive the response is, and the worse the smoothness is. The smaller the value of λ is, the smaller the change range is, the more hysteresis the response is, the higher the smoothness is. It follows from Fig. 2.

In the course of the experiment, 22, 23, 24 are the sampling point respectively, the real-time monitoring CPU utilization rate was corresponding to 25, 30, 36, as shown in Fig. 3.

In order to verify the validity of the forecast model, the smoothing factors are 0.3 and 0.8, respectively. When $\lambda = 0.3$, the CPU utilization is predicted as 23, 26, 32, as shown in Fig. 4. It can be seen that there is a slight lag when the value of λ is too small.

When $\lambda = 0.8$, the CPU utilization estimates 24, 32, 40, as shown in Fig. 5. Although the predicted value is more sensitive than the previous experimental results (i.e., $\lambda = 0.3$). However, the forecast results exhibit relatively large fluctuations,

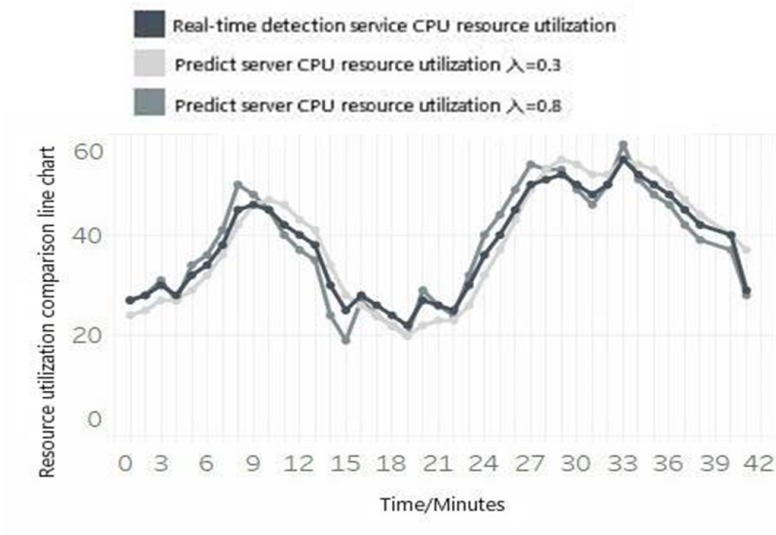

Fig. 2. Comparison of monitoring data and forecast data at the same time

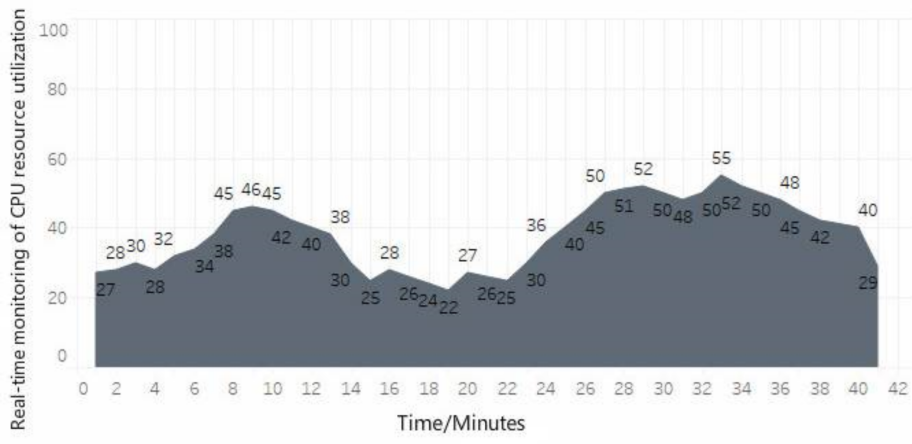

Fig. 3. Monitoring data

which solves the hysteresis and brings the volatility at the same time.

The value of the smoothing factor λ also has some influence on the accuracy of the forecast results. The experimental forecast data are divided into 40 groups. When $\lambda = 0.3$, the accuracy is 79.3 %. When $\lambda = 0.8$, the accuracy is 79.9 %. This is shown in Figs. 6 and 7.

When the number of samples increases, the accuracy of the forecast algorithm with different values of λ will be further revealed. To summarize,according to different needs, to select the appropriate value of λ is essential. In the simulation experiment, the server information is updated in minutes. When the update period

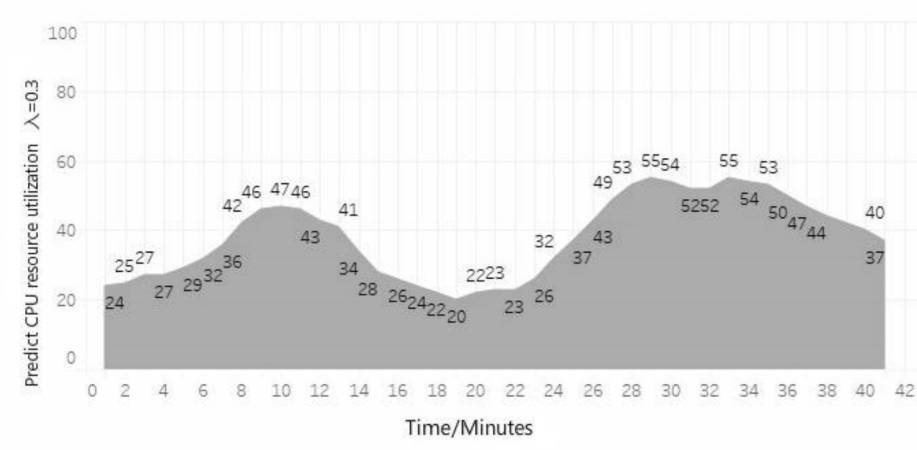

Fig. 4. $\lambda = 0.3$ forecast data charts

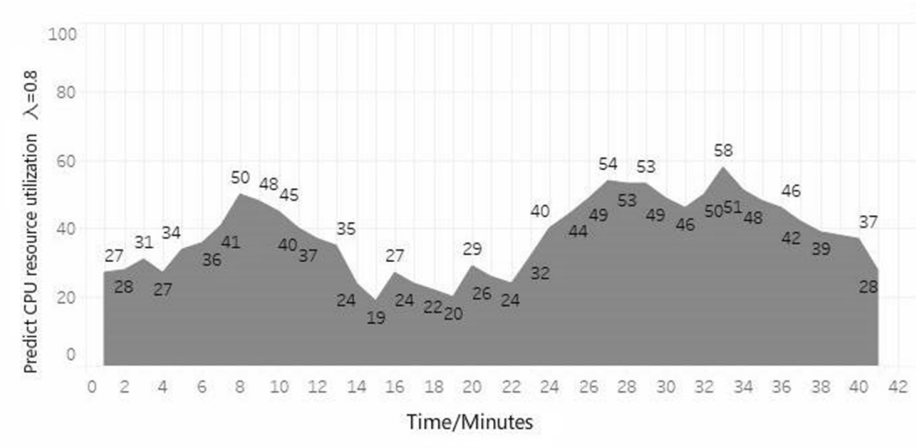

Fig. 5. $\lambda = 0.8$ forecast data charts forecast data charts

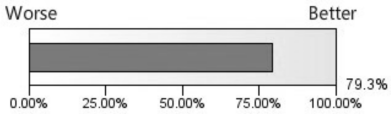

Fig. 6. $\lambda = 0.3$, prediction the accuracy of the algorithm

becomes longer, the hysteresis of the forecast algorithm becomes smaller due to the longer time period. So we propose that value of λ is too small when the server in the normal state, so as to avoid excessive reflection conditions lead to the frequent migration of virtual machines in the server, affecting the quality of QOS, resulting in SLA violations.

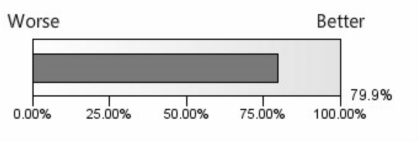

Fig. 7. $\lambda = 0.8$, prediction the accuracy of the algorithm

4. Conclusion

In this paper, the utilization rate of CPU resources in the server can be predicted accurately by the forecast algorithm, and the influence of the value of λ on the forecast trend and the forecast result is discussed emphatically. This paper analyzes the present situation of server resource utilization and puts forward corresponding solutions. The forecast algorithm can find out the possible overload and no load earlier, which can be used to select the source of the virtual machine migration. Through the forecast algorithm with the virtual machine migration can achieve lower energy consumption and QOS upgrade. At the same time, there are some drawbacks in this study. It is found that the forecast results and the actual monitoring results are different and the accuracy of the algorithm needs to be further improved. At the same time, we should further study the choice of virtual machine migration source and migration target, combined with the forecast algorithm, the establishment of model simulation cloud platform cluster scheduling process, observation server status and energy consumption.

References

[1] A. Beloglazov, R. Buyya: *Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers.* concurrency and computation ,Pract. Exper *24* (2012), No. 13, 1397—1420.

[2] T. Setzer, A. Wolke: *Virtual machine re-assignment considering migration overhead.* IEEE Network Operations and Management Symposium (2012), 631–634.

[3] S. Vakilinia, B. Heidarpour, M. Cheriet: *Energy Efficient Resource Allocation in Cloud Computing Environments.* IEEE Access *4* (2016), 8544–8557.

[4] O. Litvinski, A. Gherbi: *Experimental evaluation of openstack compute scheduler.* Procedia Computer Science *19* (2013), 16–123.

[5] J. H. Chen, C. F. Tsai, S. L. Lu, S. L. Luc, F. Abedin: *Resource reallocation based on sla requirement in cloud environment.* IEEE 12th International Conference on e-Business Engineering (2015), 377–381.

[6] H. Liu, C. Z. Xu, H. Jin, J. Gong, X. Liao: *Performance and energy modeling for live migration of virtual machines.* The 20th International Symposium on High Performance Distributed Computing, ser. HPDC '11 (2011), 171–182.

[7] W. Voorsluys, J. Broberg, S. Venugopal, R. Buyya: *Cost of virtual machine live migration in clouds: A performance evaluation.* The 1st International Conference on Cloud Computing, ser. CloudCom '09 (2009), 254–265.

[8] S. R. M. Amarante, A. R. Cardoso, F. M. Roberto, J. C. Jr.: *Using the multiple knapsak problem to model the problem of virtual machine allocation in cloud computing.* Proc. CIT 2013 (2013), 476–483.

[9] 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.

Received April 30, 2017