Resource Prediction Based on Double Exponential Smoothing in Cloud Computing

Jinhui Huang[1], Chunlin Li[2], Jie Yu[2]
Department of Computer Science and Technology
Wuhan University of Technology
Wuhan, CHINA
Email: {hjh251515@126.com[1],chunlin74@yahoo.com.cn[2],yuguijie@126.com[2]}

***Abstract*—With the development of cloud computing, customers are more and more concerned with cost on the resources which are not free in the cloud. Cloud resource providers can offer users two payment plans, i.e., reservation and on-demand plans for resource provision. In general, cost on resources gained by reservation plan is cheaper than on-demand plan. So the accuracy of resource prediction is of importance. In this paper, we present a resource prediction model based on double exponential smoothing, which considers not only the current state of resources but also the history records. Experiments performed on CloudSim cloud simulator show that the proposed method has a better performance on prediction accuracy.**

Keywords-resource provision;resource prediction model;double exponential smoothing;CloudSim

I. INTRODUCTION

Cloud computing is an emerging paradigm that aims at streamlining the on-demand provisioning of software, hard-ware, and data as services, providing end-users with flexible and scalable services accessible through the Internet [1].The basic function of cloud computing resource management system(CCRMS) is to response the request from users, and then assign resources which the user imply. Scheduling the resources accurately to make sure the job scheduling executed correctly. Cloud computing resource management includes resource provision and virtual machine placement.

Cloud providers can offer customers two payment plans, i.e., reservation plan and on-demand plan[2]. In general, cost of resources in reservation is cheaper than that in on-demand plan. However, customers need to subscribe a certain amount of resources in reservation plan in advance for future usage. As a result, it is very important to predict the amount of resources the customers need in the future. Due to the uncertain of the customers' need and the price of the resource, we should consider not only the current state of resource but also the history records. This will reduce the resource idle rate and cost in significant measure.

The arbitrary arrival of jobs and the autonomy behaviors of resource and its owner lead to the dynamicity of the resource state, while the resource information center can only store static information about resource(for example, CPU frequency, memory size, disk capacity etc).So prediction models which deduce possible system state in the future by learning from the current and history statistics is needed to improve the efficiency of scheduling and resource reliability.

In this paper ,we propose a double exponential smoothing[3] based resource prediction model, which considers not only the current state of resource but also the history data. We design a resource evaluating method to convert prediction result to the basis of scheduling. Experiments performed on CloudSim cloud simulator show that the proposed method has a better performance on prediction accuracy and resource idle rate.

The rest of the paper is organized as follows: Section 2 presents the related work on resource prediction. Section 3 mainly describe the background of the double exponential smoothing. The prediction model and the scheduling algorithm are described in section 4.Experiment results are described and analyzed in section 5.We draw a conclusion and discuss some future work in section 6.

II. RELATED WORKS

Some researchers about resources prediction have been done. A Markov chain based resource prediction is proposed in [4],which comprehensively considers the rate of CPU usage, level of network load, and resource failure rate to forecast resource future state for getting better job scheduling results. But Markov chain model only consider information of current state, and neglect the history records of resources.

Forecasting for grid and cloud computing on-demand resources based on pattern matching is presented in paper [5].It propose an approach to the problem of workload prediction based on identifying similar past occurrences for the current short-term workload history. It also present in detail the cloud client resource auto-scaling algorithm that uses the approach to help when scaling decisions are made. This considers the history records while it has bad performance when there is no pattern matched for the coming customer.

978-1-4577-1415-3/12/$26.00 ©2012 IEEE

Considering the unique features of long-connectivity applications which are increasingly popular nowadays, paper [6] proposes an algorithm －Exponential Smoothing forecast based on the Weighted Least Connection(ESBWLC). ESBWLC optimizes the number of connections and static weights to actual load and service capability, and adds single exponential smoothing forecasting mechanism. This algorithm can improve the load of real servers effectively, but has a bad performance on prediction accuracy.

Paper [7] introduce an architecture which include a dual-purpose predictor that allows users to negotiate with providers in service-level terms and provides a mean for the scheduler to perform smart resource allocation using these predictions. The evaluation shows that fast algorithms are able to make prediction with an 11% and 17% of relative error for the CPU and memory respectively and the potential of using accurate predictions in the scheduling compared to simple yet well-know schedulers.

A M/M/1 queuing model predicting method(MQMPM) is proposed in paper [8]， which is based on continuous-time birth and death process. This predicting method has some shortcomings. One is that this method has bad predicting performance when the sequence continues to increase rapidly. The predicted value curve is delayed some time compared with the real utilization sequence. And another shortcoming is that the resource reservation in flat and smooth period is too much.

This paper proposes a double exponential smoothing method which considers both the current data and the history records. The method is widely used in sales and production forecast, and it can be also used in cloud computing in which existing business relationship.

III. DOUBLE EXPONENTIAL SMOOTHING

A. The conception of double exponential smoothing

Exponential smoothing method is proposed by Robert G.Brown who thinks that the situation of time series is of stability and regularity. That is to say, the previous statements will continue into the near future. Exponential smoothing is one of the most frequently used method in production forecast. Simple mean method use all of the history records. Moving average method do not consider forward data, and weighted moving average method give the recent data greater weight. Exponential smoothing method do not give up the history record while give them weaken influence degree, which has both the advantages of simple mean method and weighted moving average method. The fundamental formula of Exponential smoothing is as follow:

$$S_t = \alpha \cdot y_t + (1-\alpha) S_{t-1} \qquad (1)$$

S_t represents prediction value of t-period. y_t represents actual value of t-period. S_{t-1} represents prediction value of t-1-period. α represents the smoothing factor (0< α <1). Exponential smoothing method can be classified as single exponential smoothing, double exponential smoothing and cubic exponential smoothing according to the smoothing frequency.

- Use single exponential smoothing when time series have no significant trend changes. The predictor formula can be expressed as:

$$y'_{t+1} = \alpha y_t + (1-\alpha) y'_t \qquad (2)$$

y'_{t+1} represents prediction value of t+1-period. y_t represents actual value of t-period. α represents the smoothing factor(0< α <1). y'_t represents prediction value of t-period. It can be also written as:

$$y'_{t+1} = y'_t + \alpha(y_t - y'_t) \qquad (3)$$

We can see from the formula that the next predict value is the sum of current predict value and α discount for subtraction of current actual value and predict value.

- Double exponential smoothing is another smoothing of single exponential smoothing. It is suitable for time series with linear trend. The predict formula express as:

$$y_{t+m} = (2y'_t - y_t) + m(y'_t - y_t)\alpha/(1-\alpha) \qquad (4)$$

In the formula, $y_t = \alpha y'_{t-1} + (1-\alpha) y_{t-1}$,so we can see that double exponential smoothing is a linear equation with the intercept $(2y'_t - y_t)$, slope $(y'_t - y_t)\alpha/(1-\alpha)$ and independent variable predicting times m.

B. Determine the smoothing factor α and the initial value of predictive model

The determination of the smoothing factor α is very important in exponential smoothing, while it is easy influenced by subjective factor. In general, the value of α should be greater when the data is prone to big swings which will increase the impact of the recent data to the result. And inversely, α should be smaller. There are two methods we can choose in theory horizon. The first one is decided by experience. When the time series appear with stable level trend, α should be small(0.05~0.20).When time series are prone to small swings, α can be a little greater(0.1~0.4).When time series are prone to big swings, α should be much greater(0.6~0.8). α should be (0.6~1) when time series are of rising or declining trend.

Another is trial method. First we determine the scope according to specific time series. And then try several α to calculate predict standard error. We choose the one with the smallest standard error.

The determination of the initial value of predictive model is also very important. In general, we choose the first data of the time series as the initial value when the item number is great(n>15).Inversely, we set the initial value as average of the previous few numbers when its item number is small(n<15). There are also some more accurate methods, such as choose several data(3 to 5) at both ends of the series. And then calculate the mean number of them. We get two points at the place of (t_1, x_1) and (t_2, x_2) .According these two points, we can build two linear trend equations:

$$\begin{cases} x_1 = a + bt_1 \\ x_2 = a + bt_2 \end{cases} \quad (5)$$

We can figure out a and b according the above equation set. According to the fundamental formula of exponential smoothing we can conclude that:

$$\begin{cases} a = 2s_t^1 - s_t^2 \\ b = \dfrac{\alpha}{1-\alpha}(s_t^1 - s_t^2) \end{cases} \quad (6)$$

Then we can figure out s_t^1 and s_t^2 which are initial single exponential smoothing value and initial double exponential smoothing value.

IV. PREDICTION MODEL AND SCHEDULING ALGORITHM

A. Prediction model

Two assumptions about prediction model are proposed to simplify the building of modeling.

a) Each resource performance characteristic can be measured quantifiably,and described by a stream of measurements. On the basis,more precise description of resources can lead to more accurate.

b) We can monitor and gather the performance measurements non-intrusively.Information monitoring and gathering can make use of the operating system utility,the load of which can be neglected.

In this paper, we consider two factors of the performance, CPU and memory. The utilization of CPU and the utilization of memory are calculated by the following formula.

$$C_{util} = \frac{c_{user} + c_{sys}}{c_{total}} \times 100 \quad (7)$$

$$M_{util} = \frac{M_{user} - M_{cache}}{M_{total}} \times 100 \quad (8)$$

The subscript ‘util’ denotes the utilization rate of a resource, ‘user’ denotes the utilization rate of a user, ‘sys’ denotes the utilization rate of a system, ‘cache’ denotes the utilization rate of cache memory, ‘total’ denotes the maximum available utilization rate.

In general, we can use SSE(Summary of the Squared Errors) or MSE(Mean of the Squared Errors) to evaluate the accuracy of the prediction model. In this paper, we choose SSE. It can be calculated by the following formula.

$$SSE = \sum_{i=1}^{n} (S_i - y_i)^2 \quad (9)$$

S_i denotes the prediction value of time i-period. y_i denotes the actual value of time i-period.

B. Scheduling algorithm description

In the description of scheduling algorithm, we introduce the Resource Management Log(RML),which keep the resource informations the customer used before. Once the customer completes a job scheduling, the resources used this time will be added to RML. Here the resources we take CPU and memory for instance. Then we can get the smoothing factor α and the initial exponential smoothing value according to the time series. After this, the resource that the customer needed next time can be predicted, while the customer just need to reserve it.Fig.1 describes a prediction algorithm based on prediction.

```
1   main(){
2    init();
3    for(i=1;i<TotalNum_customers ;i++){
4     while(job j_n completed){
5     //communicates with RML to get resource list
6     //add current resource used log to RML
7     rml_list(i)=get_rml_list(i);
8     newest_list(i)=add( j_n ,rml_list(i));
9     //determine factor α and initial smoothing value
10    //according to the newest_list(i)
11    if(n>15){
12       s'_1 = newest_list(1) ;
13       s''_1 = newest_list(1) ;
14    }else{
15       s'_1 =average( newest_list(1) ,
16            newest_list(2) , newest_list(3) );
17       s''_1 = s'_1 ;
18    }
19    if(avsd<10){
20          α =0.1;
21    }else if(10<=avsd<100){
22          α =0.3;}
23     else{
24          α =0.7;
25    }
26    s''_{n+1} ← (s'_1, s''_1, α) ;
27    } //end while
28   } //end for
29  } //end main
```

Figure 1. prediction algorithm

Function average(a, b, c) at line 15 means the average value of a, b and c. Symbol 'avsd' at line 19 and line 21 means the standard deviation of actual value adjacent of the resource. s'_1 denotes the single exponential smoothing value and s''_1 denotes the double exponential smoothing value.

V. THE EXPERIMENT

The experiment is on the CloudSim cloud simulator which is a framework for modeling and simulating the cloud computing infrastructures and services[9]. In order to confirm the accuracy of double exponential smoothing prediction algorithm, we compare the predicting value with the actual value. On the other hand, we compare this method with simple mean based method and weighted moving average method, which show the advantage of double exponential smoothing method.

A. Compare with the actual value

The first experiment is comparing the prediction value with the actual value on the amounts of CPU used. The result is shown in Fig.2.

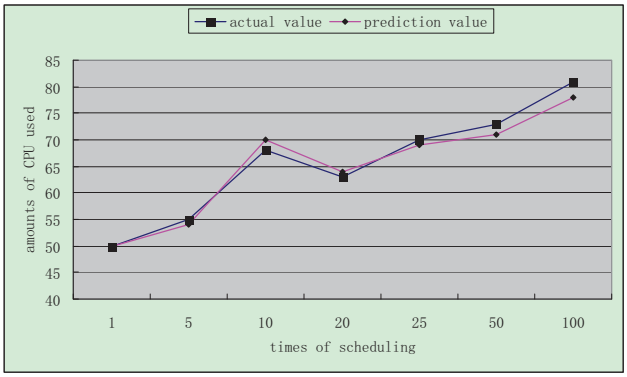

Fig 2.Compare with the actual value

It can be seen from Fig.2 that the prediction value equals the actual value at the first time of scheduling. As the time of scheduling increasing, the prediction value increase or decrease according to the actual value.

B. Comparison on prediction accuracy

The second experiment is comparing double exponential smoothing method with simple mean based method and weighted moving average method. The result is show in Fig.3.

Fig 3.comparison of prediction accuracy

We can see from the Fig.3 that double exponential smoothing method is much better than mean-based method and weighted moving average method. That is because double exponential smoothing method considers not only the current data but also the history records, and give the different data with different influence levels.

VI. CONCLUSION

The double exponential smoothing based resource prediction model fully consider the current data and the history records. We do not neglect the past data while give them weaken influence

degree by introducing the smoothing factor. Experiment proved that the model has a better performance comparing with the other two models. During resource provision in cloud computing, the customers at first predict resources they needed next time ,and then reserve them. Under the premise of good prediction accuracy, customers can save much money on using cloud resource.

In the future works, further research will consider energy efficient which is also conducive to save money for the customers and providers.

REFERENCES

[1] Dikaiakos, M.D.; Katsaros, D.; Mehra, P.; Pallis, G.; Vakali, A.; , "Cloud Computing: Distributed Internet Computing for IT and Scientific Research," *Internet Computing, IEEE* , vol.13, no.5, pp.10-13, Sept.-Oct. 2009

[2] Chaisiri, S.; Lee, B.; Niyato, D.; , "Optimization of Resource Provisioning Cost in Cloud Computing," *Services Computing, IEEE Transactions on* , vol.PP, no.99, pp.1, 0

[3] LaViola, J.J., Jr.; , "An experiment comparing double exponential smoothing and Kalman filter-based predictive tracking algorithms," *Virtual Reality, 2003. Proceedings. IEEE* , vol., no., pp. 283- 284, 22-26 March 2003

[4] Shi Lili; Yang Shoubao; Guo Liangmin; Wu Bin; , "A Markov Chain Based Resource Prediction in Computational Grid," *Frontier of Computer Science and Technology, 2009. FCST '09. Fourth International Conference on* , vol., no., pp.119-124, 17-19 Dec. 2009

[5] Caron, E.; Desprez, F.; Muresan, A.; , "Forecasting for Grid and Cloud Computing On-Demand Resources Based on Pattern Matching," *Cloud Computing Technology and Science (CloudCom), 2010 IEEE Second International Conference on* , vol., no., pp.456-463, Nov. 30 2010-Dec. 3 2010

[6] Xiaona Ren; Rongheng Lin; Hua Zou; , "A dynamic load balancing strategy for cloud computing platform based on exponential smoothing forecast," *Cloud Computing and Intelligence Systems (CCIS), 2011 IEEE International Conference on* , vol., no., pp.220-224, 15-17 Sept. 2011

[7] Reig, G.; Alonso, J.; Guitart, J.; , "Prediction of Job Resource Requirements for Deadline Schedulers to Manage High-Level SLAs on the Cloud," *Network Computing and Applications (NCA), 2010 9th IEEE International Symposium on* , vol., no., pp.162-167, 15-17 July 2010

[8] Yuxiang Shi; Xiaohong Jiang; Kejiang Ye; , "An Energy-Efficient Scheme for Cloud Resource Provisioning Based on CloudSim," *Cluster Computing (CLUSTER), 2011 IEEE International Conference on* , vol., no., pp.595-599, 26-30 Sept. 2011

[9] R. Calheiros, R. Ranjan, A. Beloglazov, C. De Rose, and R. Buyya, "Cloudsim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisoning algorithms," *Software: Pratice and Experience*, vol. 41, no. 1, pp. 23-50, 2011.