Research on Resource Allocation for Cloud Computing Platform based upon Service Requirement

Qian Zhao¹, Guangsheng Feng², Rui Gao², Ke Han¹

¹ School of Computer and Information Engineering, Harbin University of Commerce, Harbin, China
² College of Computer Science and Technology, Harbin Engineering University, Harbin, China,
   {zhaoqian,fengguangsheng,zhengchen}@hrbeu.edu.cn, hanke@ hrbcu.edu.cn

Abstract. A mixed-prediction based resource allocation approach is proposed in this paper, which is abbreviated as MPRA. This proposed MPRA employs Fast Fourier Transform theory to determine the cyclical attribution. If there is no such attribution existed, the Markov chain is alternatively used to predict the tendency of resource requirement. The experimental results show that the proposed MPRA could predict the future resource requirement more precisely. It could also allocate the virtual machine resource adaptively, decrease the number of occupied physical machines, and reduce the probability of violating the Service-level Agreement.

Keywords: Resource allocation; Fast Fourier Transform; Markov Chain

1 Introduction

The traditional resource allocation and payment according to the configuration of the cloud computing platform cannot satisfy the increasing variety of end users and individual needs, and the payment mode according to the individual requirement has caused extensive concern [1-3]. This traditional way of resource allocation could cause the virtual machine scattered distribution, and physical machine resource occupation increasing and resource wasted easily. As the booming of cloud users especially mobile users, this resource allocation methods not only hinder the improvement of resources utilization rate seriously, and cause the increase of operating costs rapidly [4-6].
To solve these problems, this paper designs an adaptive allocation method, the Mixed Prediction based cloud platform Resource Allocation (MPRA), which is able to distinguish service resources demand characteristics and meanwhile judge periodic characteristics of service demand based on FFT theory. The proposed method could achieve the goal of improving system resource utilization and reducing the violation of SLA simultaneously.

2 MPRA: system model

Let $v = \{v_1, \ldots, v_m\}$ denote the current $m$ virtual machines that cloud platform system has allocated, and $s_i$ denote the amount of resource allocated to virtual machine $v_i$. There are $l$ physical machines present in the cloud platform system, which is expressed by $p = \{p_1, \ldots, p_l\}$, and $R(v, p)$ represents the mapping relationship between the virtual machine set $v$ and the physics machine set $p$.

Suppose $X = \{X(1), \ldots, X(L)\}$ as the $L$ services provided by the cloud platform, and the resources interval is $\{x_{\min}, x_{\max}\}$. Since the cloud platform lacks a priori knowledge and history of the initial access, MPRA will allocate the service resource with $x_{\min}$. After MPRA obtaining $w$ resource usage samples, i.e., $U^{(w)} = \{u_1^{(w)}, \ldots, u_w^{(w)}\}$, the resource demand forecasting process will be triggered: $U^{(w)}$ will be transformed by FFT, where time interval with significant amplitude represents the cycle of request for the resource. However, the minimum one $f_{\min}^{(w)}$ will be chosen as the main period frequency if there are several amplitudes greater than the average. Thus, it follows the maximum period of the resource request $1/f_{\min}^{(w)}$. For a main frequency $f_{\min}^{(w)}$, the cycle window size of $U^{(w)}$ is:

$$Z^{(w)} = (f_{\min}^{(w)})^{-1} \times r$$

where $r$ is the sampling rate.

A similarity test is conducted for any two windows. If the Pearson correlation coefficient of any two windows $U_i$ and $U_j$ closes to 1 and their average values are approximately equal, it is determined that the resource request of the service is in accordance with the cycle $Z(l)$. Then, the DWT (Dynamic Time Warping) technology is used to get the resources required sequence of the service $X(l)$ in the recent cycle, which makes the mapping distance between it and the current requiring resource sequence shortest. Thus, the resource demand of service $X(l)$ at time $t$ is $F^{(l)}(t)$.

However, when the service $X(l)$ does not have cyclical characteristic, we use Markov chain to predict the resource requirement at time $t$ in the paper. In order to use Markov theory, the resources required state of the service $X(l)$ is equally divided into $n = \{(x_{\min} - x_{\max})/I\}$ intervals in advance, which is used to distinguish the state of the resource requirements in different conditions. Correspondingly, each resource demand interval represents a state of the resource or service status, and resource requirement for each state takes the mean value of the interval. Therefore, the changes of the resource requirement of the service can be considered as a time series $X(l)(t) = 1, 2, \ldots, n$. Resource requirement for each time point belongs to the different state of resource requirements. Then, we employ statistical technology to
calculate the transition probability matrix of the Markov process.

\[
p^{(i)} = \begin{pmatrix} p_{i,j}^{(i)} \end{pmatrix} \quad (i = 1, \ldots, n; j = 1, \ldots, n)
\]

(2)

where \( p_{i,j}^{(i)} \) is the transition probability from state \( s_i^{(i)} \) to state \( s_j^{(i)} \). \( m_i^{(i)} \) represents the numbers that the state \( s_i^{(i)} \) appears at different time, and \( m_j^{(i)} \) represents the numbers state \( s_i^{(i)} \) transfer to state \( s_j^{(i)} \). So we can obtain the state transition probability matrix of this service \( p^{(i)} \).

Based on this, MPRA predicts the short-term resource requirements by constructing a discrete finite-state Markov chain model. Provided that the Markov chain is homogeneous, any system status probability in time \( t \) can be calculated by C-K equation, and it follows the possible state prediction of the service:

\[
\pi_i^{(l,t)} = \pi_i^{(l,0)} p_i^{(l)} \quad (i = 1, \ldots, n)
\]

(3)

where \( \pi_i^{(l,0)} \) and \( \pi_i^{(l,t)} \) is the probability distribution at the initial time and at time \( t \) of \( X(l) \) respectively. According to the current state \( \pi_i^{(l,0)} \), the probability in state \( s_i^{(i)} \) after the time \( t \) can be inferred. Therefore, the forecast value of the resource requirements of the service at time \( t \) is \( \pi_i^{(l,t)}(p^{(i)}) \). In summary, the resource requirements of the service \( X(l) \) at time \( t \) can be forecasted as:

\[
R^{(l,t)}(t) = \begin{cases} F^{(l,t)}(t), & \text{if } X^{(l)} \text{ has cyclical characteristics} \\ \pi_i^{(l,t)}(p^{(i)}), & \text{if } X^{(l)} \text{ has Markov property} \end{cases}
\]

(4)

3 Experiment and analysis

We first build two guest virtual machines VM1 and VM2, and each guest virtual machine is assigned a complete physical kernel and 512M memory to simulate physical machine with fixed resources. Then, VM1_sub and VM2_sub are built as the Web Servers. Moreover, aperiodicity application AP_1 and periodic application AP_2 are simulated by RUBis through importing ClusterData2011_10.

As shown in Fig1, in contrast with ARMR[6], the MPRA has a consistent tendency with fixed allocation until it collects a certain number of samples. The dynamic resource allocation mechanism is triggered in in 8 min and 16 min, and CPU utilization of virtual machine remains about 70% after 22 min. Thus, the fixed allocation leads to a great waste of physical CPU resources in contrast with MPRA.

Fig2 shows the memory utilization rate of VM1. The ARMA algorithm adjusts memory resource in 8 min, 14 min and 26 min. Correspondingly, the memory utilization rate increases to 85% in 26 min because of inaccurate prediction, which may lead to a collapse event during the service process. The MPRA has accumulated enough prediction samples in 12 min so that the utilization rate of memory remains ideal state about 70%.
4 Conclusion

In this paper, the MPRA method is proposed to allocate cloud platform resource according to the periodic and aperiodic service requirement. This method first employs FFT theory to judge the periodic cycle and uses Markov process to predict aperiodic service resource requirement. This method obtains a higher resource utilization efficiency than other traditional ones, and the number of violating SLA is deducted obviously. In the future, we will focus our research on the aperiodic service requirements.

References