Cloud Computing System Scheduling Model Based On Bayesian Network

YangHong[1, a], YangLi[1, a], LvFeng[1, a*]

[1]Yunnan Traditional Chinese Medicine College, Kunming, 650500, China

[a]email: lvfeng_yn@163.com

Keywords: Cloud Computing; Architecture; Elasticity; Copy Service; Optimal Scheduling

Abstract. Cloud computing platform has also posed challenges to resource management and service scheduling using virtualization technology while allowing software applications to become more efficient. On the basis of making study on difference in scheduling between "SaaS" and "IaaS", this paper focuses on resource scheduling of architecture layer, proposing scheduling model based on stochastic theory, and describes the scheduling of the layer as a problem of multi-objective optimization. In addition to service quality requirements, this paper also considers the important feature of cloud service of flexibility, and provides policy for matching task scheduling and elasticity copy service. Experiments show that the design of the scheduling mechanism has optimized the overall performance of the cloud platform, which has achieved a better load balancing and resource utilization.

Introduction

Cloud computing platform allows users to access computing resources in the form of cloud services, and according to the type of resources or services available to users (tenants), cloud computing architecture is divided into different levels. In IaaS layer, virtualization technology is applied to enable many virtual machines (VM) to run on a single physical server and provides flexible resource sharing. In layer SaaS, through the use of the virtual resources provided by layer IaaS model, the cloud computing platform can produce a series of virtual environments to meet different application needs. In addition, flexible virtual resources can form corresponding package to provide services with elasticity for the cloud, so that all applications deployed on the cloud can be more efficient.

Cloud service providers provide users with resources and services, but with the increase in size of cloud, the complexity of resource management is improving, which brings challenges to traditional scheduling mechanism. In the layer IaaS, a single physical machine can run multiple VMs, which makes the acquisition of accurate information about shared resources (CPU, memory, etc.) on physical server become more difficult. In addition, dynamic migration [5] and the elastic characteristics of cloud have also increased dynamics of algorithm design. In this paper, it is intended to design a set of scheduling mechanism to ensure QoS (Quality of Service) requirements for cloud computing, so as to improve the overall performance of cloud platforms.

Design of Optimal Scheduling Scheme of Cloud Services

For services deployed in the cloud platform, by designing appropriate scheduling policy, planning matching relation between service request and request copy, the execution order of service, and occupancy of the underlying resource for service, the quality of cloud resource management services is further enhanced.

Service scheduling scenario in cloud platform is shown in Figure 1. When the request of network manager service arrives at cloud platform, it will firstly be stored in the scheduling queue, and later, based on scheduling strategy, the service scheduling module of cloud platform, aimed to optimize service execution efficiency, selects the most appropriate request for processing in the queue, and assigns the request to the most appropriate copy service for processing, to complete the functions of network resource management. Among them, Mapping policy analyzer is responsible for assigning

tasks to the appropriate service copy to optimize the average response time, while Elasticity controller is responsible for the elastic distribution of cloud services to optimize resource utilization.

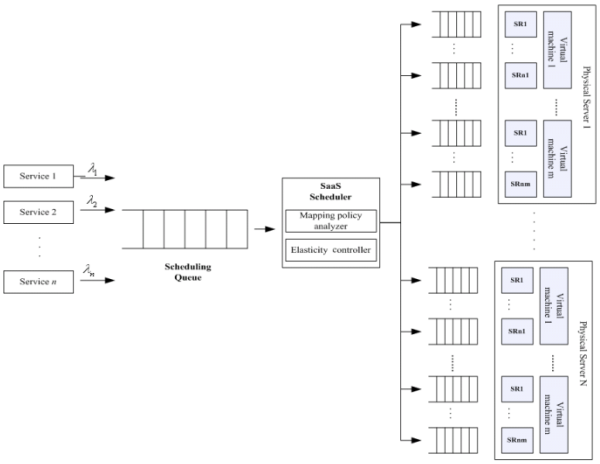

Fig.1. Service scheduling scenario in cloud platform

Specifically, in the scheduling scheme, it requires successively improving optimized queuing strategy for service requests, matching strategy between request and service copy as well as the implementation of allocation strategy for service elasticity. Using stochastic model based on queuing theory, the model design is conducted when service performance of scheduling proceeds to the steady state.

Table 1 describes the main parameters and related definitions used in scheduling model.

Tab.1. Main parameters and related definitions

Parameters	Definitions
λ_m	Poisson strength arrived by the M-th service request
S_{VM}	Assemble of available virtual resources in cloud platform
S_{SR}	Collection of all service copies in cloud platform
p_{im}	The probability of the i-th service request assigned to the m-th service copy
μ_{im}	The execution efficiency of the i-th service request in the m-th service copy
ρ_{SRm}	Resource utilization of the m-th service copy
T_{SRm}	Elapsed time in the performance of the m-th service copy
L_{total}	Scheduling queue length
I_{total}	Workload processed by unit of resources in cloud platform
T_{total}	The average time of processing a single service in cloud platform at steady state
M	The number and types of cloud services

Bayesian optimization of execution efficiency of cloud services

Assumed that arrival of service requests follows Poissonian distribution, and when the i-th service request arrives waiting queue the strength is λ_i, and then the strength of the service reaching specific copy m is as follows:

$$\lambda_{SRm} = \sum_{i=1}^{M} \lambda_i p_{im}$$

Using queuing theory model, it can be obtained that when the scheduling is in a stable state, the expression of resource utilization of copy m is:

$$\rho_{SRm} = \frac{\lambda_{SRm}}{\mu_{SRm}} = \sum_{i=1}^{M} \frac{\lambda_i p_{im}}{\mu_{im}}$$

Later, according to resource usage of copy, the average time $T_{service}$ in the execution of a service

for copy m and the average time T_{wait} of request queue at steady state are:

$$T_{service} = \frac{1}{\mu_{SRm}} = \frac{\rho_{SRm}}{\lambda_{SRm}} = \frac{1}{\sum_{i=1}^{M} \lambda_i p_{im}} \cdot \sum_{i=1}^{M} \frac{\lambda_i p_{im}}{\mu_{im}}$$

$$\sigma_{service} = \sqrt{\frac{1}{M} \sum_{i=1}^{M} (\frac{1}{\mu_{im}} - T_{service})^2}$$

$$T_{wait} = \frac{\rho_{SRm} T_{service} [1 + (\frac{\sigma_{service}}{T_{service}})^2]}{2(1 - \rho_{SRm})} = \frac{\sum_{i=1}^{M} (\frac{\lambda_i p_{im}}{\mu_{im}}) T_{service} [1 + (\frac{\sigma_{service}}{T_{service}})^2]}{2(1 - \sum_{i=1}^{M} \frac{\lambda_i p_{im}}{\mu_{im}})}$$

When reaching steady state, the total time in performance of a service for copy m is the sum of time cost in queuing and service processing, thus service response time from positioning to copy m is expected to be:

$$T_{SRm} = T_{service} + T_{wait} = \frac{1}{\sum_{i=1}^{M} \lambda_i p_{im}} \cdot \sum_{i=1}^{M} \frac{\lambda_i p_{im}}{\mu_{im}} + \frac{\sum_{i=1}^{M} (\frac{\lambda_i p_{im}}{\mu_{im}}) T_{service} [1 + (\frac{\sigma_{service}}{T_{service}})^2]}{2(1 - \sum_{i=1}^{M} \frac{\lambda_i p_{im}}{\mu_{im}})}$$

The overall response time of the entire services processing in cloud platform is obtained further, also one of the optimization objectives of the model:

$$T_{total} = \sum_{i=1}^{M} (\lambda_i \cdot \frac{\sum_{m=1}^{|S_{SR}|} p_{im} T_{SRm}}{\sum_{m=1}^{|S_{SR}|} \lambda_{SRm}}) \tag{1}$$

Cloud service distribution based on bayes

In addition to optimizing the execution efficiency of service, and improving the overall response time of working in cloud platform, this scheduling policy has further taken into account the elasticity distribution of cloud services, elevating efficient use of resources, to avoid excessive and insufficient allocation of resources. Specific modeling method as to the elasticity assignment of cloud services is as follows:

With queuing model, request volume of backlog of copy m in steady state is:

$$L_{SRm} = \lambda_{SRm} T_{SRm} = \sum_{i=1}^{M} \lambda_i p_{im} (T_{wait} + T_{service})$$

With resource utilization, the request volume for copy m processing unit of resources in steady state is further obtained as follows:

$$I_{SRm} = \frac{L_{SRm}}{\rho_{SRm}} = \frac{\sum_{i=1}^{M} \lambda_i p_{im} (T_{wait} + T_{service})}{\rho_{SRm}}$$

Furthermore, the workload of unit of resources processed by the entire cloud platform is obtained, and this is also another optimization goal of the model:

$$I_{total} = \frac{L_{total}}{\rho_{SRm} |S_{SR}|} = \frac{\sum_{i=1}^{M} \sum_{m=1}^{|S_{SR}|} \frac{\lambda_i p_{im} T_{SRm}}{\rho_{SRm}}}{|S_{SR}|} \tag{2}$$

Integrated with (1) and (2), the optimal scheduling problem of cloud services is abstracted as a multi-objective optimization model: find the optimal probability distribution p_{im}, thus minimizing the response time T_{total} of service; at the same time, maximize the average execution workload I_{total} according to elasticity of services.

$$\text{Min} \quad T_{total} = \sum_{i=1}^{M} (\lambda_i \cdot \frac{\sum_{m=1}^{|S_{SR}|} p_{im} T_{SRm}}{\sum_{m=1}^{|S_{SR}|} \lambda_{SRm}})$$

$$\text{Max} \quad I_{total} = \frac{\sum_{i=1}^{M} \sum_{m=1}^{|S_{SR}|} \frac{\lambda_i p_{im} T_{SRm}}{\rho_{SRm}}}{|S_{SR}|}$$

The main constraints of the optimization strategy is: the sum of the probability distribution is 1; the flexibility of single service and the load of service is positively correlated; the total allocated amount of service copy cannot exceed the total virtual resources provided by cloud platform; all service copy sets in cloud platform cannot be less than the total amount of the type of service request.

s.t.

$$\begin{cases} \sum_{m=1}^{|S_{SR}|} p_{im} = 1, \forall 1 \le i \le M, 1 \le m \le |S_{SR}| \\ \forall m \in S_{SR}, \forall 1 \le i \le M, \sum_{m}^{p_{im}>0} \mu_{SRm} \propto \sum_{m}^{p_{im}>0} L_{SRm} \\ \sum_{m}^{p_{im}>0} |S_{SRm}| \le |S_{VM}| \\ |S_{SR}| \ge M \end{cases}$$

The core idea of scheduling algorithm is to rationalize allocation policy of service queue, service matching, and service elasticity, so that the above-mentioned optimization model can achieve optimal values, to improve service response time, while maximizing working efficiency of unit resources.

Experiment and Results Analysis

To validate the effectiveness of the algorithm, five different software applications and services are deployed in the cloud, so that the arrival of every service request can satisfy Poisson distribution of different intensities. According to the policy of algorithm 1, while considering the strategy of elastic cloud services, every service request can be optimally allocated to the service copy.10 virtual machines with the same virtual resources are created on 6 physical machines, and three different services can be created at most on each virtual machine, so that there are at most 30 service copies in the cloud, and elastic increment of a particular software application is not better than 10 copies.

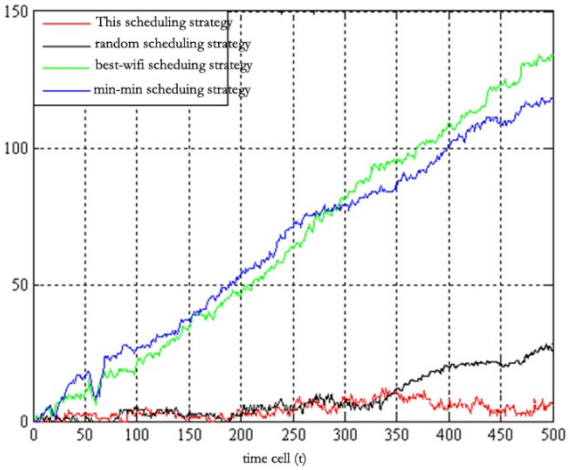

Fig.2. Task backlog curve

A comparison is made among the scheduling mechanism mode, the widely used algebraic modeling scheduling mechanism, as well as the ordinary stochastic scheduling model. Figures 2 and 3 shows change of task backlog and average response time in four scheduling mechanisms.

Bayesian algebraic model (Min-min, Best-fit algorithms) scheduling mechanism is used with high backlog of tasks. The reason is that prior to scheduling of these algorithms, many of them need to prejudge and select the tasks of the earliest completion time and the maximum throughput to perform, which requires consuming time to find "quality" task in the scheduling queue, but if the "quality" task is not performed, other tasks can neither be implemented, which has led to a large backlog of tasks. If there is no reasonable choice for queue length of the waiting queue of each service copy, there would be losses of numerous task requests, resulting in an infinite wait. However, the use of stochastic model makes it possible for each task to be performed, and optimization objectives can be achieved by designing the mapping relation between the most appropriate task and copy services between tasks, rather than by selecting "quality" tasks, thus the backlog of unprocessed tasks is quite small,. Therefore, compared to the algebraic model, stochastic model is a better choice in the design of scheduling mechanism.

As can also be seen from Figure 2, although it is also a stochastic model, compared to the general stochastic model, especially with the increase in the execution time, this algorithm has more obvious effect on reducing the backlog of task. This is the result of the strategy for arrangements of elasticity service, and optimized arrangement is made for the elasticity of each service depending on the load of the cloud services: When the load is heavy, the execution efficiency if improved by increasing the copy of these services, thus the number of the unprocessed backlogging tasks is quickly reduced naturally.

Conclusions

This paper has presented optimum deployment and scheduling mechanism of cloud services in the scene of SaaS, and has proposed a stochastic model describing SaaS scheduling as multi-objective optimization problem. Designed new SaaS scheduling mechanism has made overall consideration of the service performance and elasticity distribution. According to the experimental results, this scheduling mechanism makes effective improvement of the average executive tasks in SaaS platform based on elastic arrangements of individual services, thereby protecting the high performance of cloud platform. In subsequent studies, we will continue to examine more issues existed in cloud services, such as security constraints of applications, expenditure of migration, data sharing among different services and data coupling, and all these require implementing further design and perfection of the existing stochastic model. In addition, the scheduling mechanism of the elastic cloud services also needs continuous improvements.

Reference

[1] Li, Wubin, Johan Tordsson, and Erik Elmroth. An aspect-oriented approach to consistency-preserving caching and compression of web service response messages. In Web Services (ICWS), 2010 IEEE International Conference on, pp. 526-533. IEEE, 2010.

[2] Y. Geng, J. He, K. Pahlavan, Modeling the Effect of Human Body on TOA Based Indoor Human Tracking[J], International Journal of Wireless Information Networks 20(4), 306-317

[3] Lv, Zhihan, and Tianyun Su. 3D seabed modeling and visualization on ubiquitous context. In SIGGRAPH Asia 2014 Posters, p. 33. ACM, 2014.

[4] Yishuang Geng, Kaveh Pahlavan, On the Accuracy of RF and Image Processing Based Hybrid Localization for Wireless Capsule Endoscopy, IEEE Wireless Communications and Networking Conference (WCNC), Mar. 2015

[5] Jie He, Yishuang Geng and Kaveh Pahlavan, Toward Accurate Human Tracking: Modelling Time-of-Arrival for Wireless Wearable Sensors in Multipath Environment, IEEE Sensor Journal, 14(11), 3996-4006, Nov. 2014