ATLANTIS PRESS

Advances in Engineering Research, volume 150

4th International Conference on Machinery, Materials and Computer (MACMC 2017)

An Optimization Method for Reliable Cloud Service Composition with Low Resource Occupancy

YANG Yan, YAO Huaxiong, WANG Sai

Computer School, Central China Normal University, Wuhan, China 430079

Email: ms_yangyan@163.com

Keywords: service composition; optimization; cloud service

Abstract: In this paper, we investigate the optimization problem of cloud service composition in cloud environment. First, the performance indexes for the optimization of cloud service composition is introduced. Then we propose an optimization method of cloud service composition, and establish a constrained multi-objective model for the optimization problem of cloud service composition which requires high reliability and low CPU and bandwidth occupancy. After that, a travel planning scenario is used to illustrate our approach to cloud service composition optimization. The case study shows the proposed method can effectively solve the problem of cloud service composition with high reliability and low resource consumption in cloud computing environment.

1. Introduction

As a new model of service computing, cloud computing has quickly become a hot topic in research and industry because of its advantages of high availability and high scalability. Cloud computing can provide key software for enterprise management, effectively reduce the cost and maintenance cost of IT software and hardware, and adapt to the needs of enterprises at different phases of development. In cloud computing environment, computing task is distributed in resource pool composed of a large number of computing, storage and other units, enabling users to get the computing power, storage space and information service capacity according to the application needs. So that users can dynamically apply the required resources to support the operation of various applications deployed by users, and be able to focus more on their businesses. This service delivery model can greatly improve the utilization of resources and improve the quality of service of the platform, which has attracted wide attention from the industry.

Cloud service refers to any Internet, application software, system platform, computing resources and other services with large scale, no need for pre investment, on-demand acquisition and easy expansion. From the perspective of end users, cloud services can be divided into infrastructure as a service layer (IaaS), platform as a service layer (PaaS) and software as a service layer (SaaS) [1]. IaaS can be considered as service encapsulation of the hardware / operating system level; PaaS can be seen as service encapsulation of components and middleware; SaaS can be regarded as service encapsulation of application software level. In this work, more attention is given to the cloud services which are the applications or services provided by the SaaS layer as well as the basic capabilities provided by the PaaS layer. However, with the increasing complexity of computing tasks and the increasing variety and amount of cloud services, a single cloud service is becoming increasingly difficult to meet the needs of users. Therefore, it is quite significant to meet the complex computing needs through the composition of cloud services [2] [3]. Under such context, it is necessary to optimize the cloud service compositions in a scientific and reasonable way to provide users and enterprises with the most appropriate cloud services with high reliability and low resource occupancy.

In this paper, we investigate the optimization problem of cloud service composition in cloud environment. First, the performance indexes for the optimization of cloud service composition is introduced. Then we propose an optimization method of cloud service composition, and establish a constrained multi-objective model for the optimization problem of cloud service composition which

requires high reliability and low CPU and bandwidth occupancy. After that, a travel planning scenario is used to illustrate our approach to cloud service composition optimization. The case study shows the proposed method can effectively solve the problem of cloud service composition with high reliability and low resource consumption in cloud computing environment.

The remainder of this paper is organized as follows. Section 2 provides an overview of the related work. In section 3, we introduce the performance indexes for the optimization of cloud service composition. In section 4, an optimization method of cloud service composition is proposed, and a constrained multi-objective mathematical model is established for the optimization problem of cloud service composition which requires high reliability and low CPU and bandwidth occupancy. In section 5, a case study is presented. Section 6 concludes the paper.

2. Related work

Considering that cloud computing is still at the stage of exploration, there are few researches on the optimization of cloud service compositions. S.L. Wang et al. [4] established a multi-target optimization model for the allocation of manufacturing resources in the context of cloud manufacturing, where the maximum succession method was used to solve the model to verify the stability and feasibility of the model. W.N. Liu et al. [5] proposed a solution algorithm based on adaptive particle swarm for the multi-target problem in cloud manufacturing service composition. C.F. Lin et al. [6] proposed a relaxable QoS-based service selection algorithm for composite web services, which can recommend prospective service candidates to users by relaxing QoS constraints if no suitable or available web service could exactly fulfill user requirements. These works have certain theoretical significance and practical value for establishing a cloud service composition optimization model and promoting the application and promotion of cloud computing. However, less attention is given to those situations with a huge number of candidate cloud services as well as the low reliability and load balancing in service composition.

3. Performance indexes for the optimization of cloud service composition

In our proposal, the cloud computing service platform implements a specific planning algorithm and can generate service composition schemes to complete the required task according to the given information [7]. The platform selects a collection of cloud services with the same function for each sub-task in the composition scheme. Since it will produce a large number of alternative service composition paths which have the same topology as the composition scheme, a service filtrating process is performed to reduce the number of alternative service composition paths [8]. The sub-task structure in the scheme often contains a variety of basic topologies (such as sequential, parallel, etc.), so the alternative service composition paths often cannot be processed directly by the service composition optimization algorithm [9]. This paper introduces a mechanism to transform the basic topologies into an aggregate topology, so that the topology of a candidate service composition path can be transformed to a simple sequential topology after multiple transforms. So the composition optimization algorithm can help to choose the appropriate service composition path that optimizes multiple performance targets at the same time from all possible ones according to the performance constraints and load balancing factors. This problem is called a constrained multi-objective service composition optimization problem.

The performance indexes for the optimization of cloud service composition include:

(1) Time: Time (for short *T*) refers to the total duration of the execution process of a cloud service.

(2) Cost: Cost (for short *C*) refers to the price payed by the requestor for a specific cloud service.

(3) Availability: Availability (for short *A*) refers to the possibility that a cloud service can provide a specific service normally. Availability is expressed as $A=T_{avail}/T_{total}$, where T_{avail} represents the total running time available for cloud service in the time period T_{total}.

(4) Reliability: Reliability (for short *R*) reflects the possibility or capability of a cloud service completing the assigned service within the specified time under certain conditions, which can be

expressed by successful execution rate of the service, i.e. $R=R_{exe}(t)/B(t)$, where $R_{exe}(t)$ is the number of tasks executed by the cloud service successfully in time period t; $B(t)$ refers to the times of the service being awakened in time period t.

(5) CPU occupation rate: CPU occupation rate (for short *Cor*) refers to the proportion of the computing power needed by a cloud service to the total available computing power of the service node. It can be expressed as $Cor=CPU_{req}/CPU_{avail}$. The *Cor* value of a service composition path is defined as the largest one in the *Cor* values of all cloud services in the service composition path.

(6) Bandwidth occupation rate: CPU occupation rate (for short *Bor*), refers to the bandwidth utilization of overlay connection between candidate cloud services. It can be expressed as $Bor=BW_{req}/BW_{avail}$, where BW_{req} refers to the overlay connection bandwidth required between two candidate services; BW_{avail} refers to the available bandwidth. The *Bor* value of a service composition path is defined as the largest one in the *Bor* values of all the connections in the service composition path.

4. Optimization of cloud service composition

The ultimate goal of the optimization of cloud service composition is to select a collection of cloud services based on the user's demands to minimize the CPU occupation rate and the bandwidth occupation rate, and maximize the reliability of the cloud service composition. Meanwhile, users propose the requirements and restrictions on the tasks. The main constraints include the total constraint for time, cost and availability. According to the above goals and constraints, a multi-target planning model for the optimization of cloud service composition is established as below:

$$\text{Min } Cor(p_d) \quad \text{Min } Bor(p_d) \quad \text{Max } R(p_d) \tag{1}$$

$$\text{subject to} \quad T(p_d) \le T_{max} \tag{2}$$

$$C(p_d) \le C_{max} \tag{3}$$

$$A(p_d) \ge A_{min} \tag{4}$$

Eq. (1) represents the three target functions of the whole planning model. $Cor(p_d)$, $Bor(p_d)$ and $R(p_d)$ represent the CPU occupation rate, bandwidth occupation rate and the reliability of one possible cloud service composition p_d, respectively ($d=1,\ldots,V$; V represents the total number of possible cloud service composition after service filtering). Eq. (2) indicates that the delivery time for completing the task by the cloud service composition cannot exceed the maximum delivery time T_{max}. Eq. (3) shows that the total cost of completing the task by the cloud service composition cannot exceed the maximum cost that the user can pay C_{max}. Eq. (4) indicates the lowest limit to the availability of the cloud service composition is A_{min}.

For the process of a cloud service composition, each overall index in the service composition needs to be calculated according to the structural features of the partial processes in cloud service composition. The basic structures of the service composition flow involve sequential structure, parallel structure, selective structure and loop structure.

Here, we assume CSS^i represents the filtered candidate cloud service collection of the i^{th} subtask. The six optimization indexes of a cloud service CS_j^i ($CS_j^i \in CSS^i$, $j = 1, \ldots, N_i$, and N_i denotes the number of candidate cloud services completing the i^{th} subtask after the filtering) are denoted as $T(CS_j^i)$, $C(CS_j^i)$, $A(CS_j^i)$, $R(CS_j^i)$, $Cor(CS_j^i)$ and $Bor(CS_j^i)$. The following calculation method is used to synthesize the service composition evaluations under the four fundamental composite structures.

(1) For sequential structure,

$$T(p_d)^{seq} = \sum_{i=1}^{n} T(CS_j^i)$$

$$C(p_d)^{seq} = \sum_{i=1}^{n} C(CS_j^i)$$

$$A(p_d)^{seq} = \prod_{i=1}^{n} A(CS_j^i)$$

$$R(p_d)^{seq} = \prod_{i=1}^{n} R(CS_j^i)$$

$$Cor(p_d)^{seq} = \max(Cor(CS_j^i))$$

$$Bor(p_d)^{seq} = \max(Bor(CN_{CS_j^i}^{in}), Bor(CN_{CS_j^i}^{out}))$$

Here, $CN_{CS_j^i}^{in}$ and $CN_{CS_j^i}^{out}$ are the input and output overlay connections of the service node where cloud service CS_j^i resides.

(2) For parallel structure,

$$T(p_d)^{par} = \max(T(CS_j^i))$$
$$C(p_d)^{par} = \sum_{i=1}^{n} C(CS_j^i)$$
$$A(p_d)^{par} = \prod_{i=1}^{n} A(CS_j^i)$$
$$R(p_d)^{par} = \prod_{i=1}^{n} R(CS_j^i)$$
$$Cor(p_d)^{par} = \max(Cor(CS_j^i))$$
$$Bor(p_d)^{par} = \max(Bor(CN_{CS_j^i}^{in}), Bor(CN_{CS_j^i}^{out}))$$

(3) For selective structure,

$$T(p_d)^{sel} = \sum_{i=1}^{n} (T(CS_j^i) \times \lambda_i)$$
$$C(p_d)^{sel} = \sum_{i=1}^{n} (C(CS_j^i) \times \lambda_i)$$
$$A(p_d)^{sel} = \sum_{i=1}^{n} (A(CS_j^i) \times \lambda_i)$$
$$R(p_d)^{sel} = \sum_{i=1}^{n} (R(CS_j^i) \times \lambda_i)$$
$$Cor(p_d)^{sel} = \sum_{i=1}^{n} (Cor(CS_j^i) \times \lambda_i)$$
$$Bor(p_d)^{sel} = \sum_{i=1}^{n} (\max(Bor(CN_{CS_j^i}^{in}), Bor(CN_{CS_j^i}^{out})) \times \lambda_i)$$

Here, λ_i is the probability that the cloud service CS_j^i is selected, which satisfies $\sum_{i=1}^{n} \lambda_i = 1$.

(4) For loop structure,

$$T(p_d)^{loop} = \theta \times \sum_{i=1}^{n} T(CS_j^i)$$
$$C(p_d)^{loop} = \theta \times \sum_{i=1}^{n} C(CS_j^i)$$
$$A(p_d)^{loop} = \prod_{i=1}^{n} A(CS_j^i)$$
$$R(p_d)^{loop} = \prod_{i=1}^{n} R(CS_j^i)$$
$$Cor(p_d)^{loop} = \max(Cor(CS_j^i))$$
$$Bor(p_d)^{loop} = \max(Bor(CN_{CS_j^i}^{in}), Bor(CN_{CS_j^i}^{out}), Bor(CN_{CS_{j_n}^n \to CS_{j_1}^1}))$$

Here, θ represents the times the cloud service is executed in a loop; $CN_{CS_{j_n}^n \to CS_{j_1}^1}$ represents the overlay connection from $CS_{j_n}^n$ to $CS_{j_1}^1$ in the loop.

Based on the above model, the process of cloud service composition optimization is given as below.

Step 1 Normalize the data V_{ij} of cloud service CS_j^i on the six optimization indexes.

The value of negative-type index is normalized as $Q_{ij} = \begin{cases} \frac{V_j^{\max} - V_{ij}}{V_j^{\max} - V_j^{\min}}, & V_j^{\max} - V_j^{\min} \neq 0 \\ 1, & V_j^{\max} - V_j^{\min} = 0 \end{cases}$

The value of positive-type index is normalized as $Q_{ij} = \begin{cases} \frac{V_{ij} - V_j^{\min}}{V_j^{\max} - V_j^{\min}}, & V_j^{\max} - V_j^{\min} \neq 0 \\ 1, & V_j^{\max} - V_j^{\min} = 0 \end{cases}$, where $V_j^{\max} = Max(V_{ij})$, $V_j^{\min} = Min(V_{ij})$, $1 \le i \le k$, and k represents the number of cloud services.

Step 2 Transform the multi-target problem into a single-target problem to solve, that is

$$Min\ Z(p_d) = w_1 Cor(p_d) + w_2 Bor(p_d) + w_3 / R(p_d)$$

Where w_1, w_2 and w_3 represent the weight of CPU occupation rate, bandwidth occupation rate and reliability respectively, which satisfies $w_1+w_2+w_3=1$.

Step 3 Implement the algorithm in [10] to solve the problem and obtain the optimal cloud service composition.

5. Case study

In case study, we take a travel planning scenario as an example. When a user submits a travel reservation request to the cloud computing service platform, the platform implementing an AI planning algorithm generates a service composition scheme containing four sub-tasks, i.e. ST_1, ST_2, ST_3, ST_4, to perform the required task according to the given information. Here, ST_1 and ST_4 are in sequential structure, while ST_2 and ST_3 are running in parallel. Then the platform selects a collection of cloud services with the same function for each sub-task, before a service filtrating process is performed based on the evidential reasoning method, and the number of alternative service composition paths is greatly reduced. After that, with the filtered cloud services, an optimization model for cloud service composition is established, which can be detailed as following.

(1) Extract the performance data of the six indicators, i.e. time, cost, availability, reliability, CPU occupation rate and bandwidth occupation rate, of candidate cloud services from the cloud computing service platform, and normalize these data.

(2) According to the demand of user, determine the weight of the three goals, i.e. $w_1=0.3$, $w_2=0.3$, $w_3=0.4$, as well as the longest duration $T_{max}= 15$, the maximum cost $C_{max}= 200$, and the minimum availability $A_{min}=0.99$.

(3) By solving the composition optimization model in Eq. (1) – Eq. (4) using the algorithm in [10], the optimal cloud service composition is obtained, i.e. $\{CS_5^1, CS_3^2, CS_8^3, CS_{14}^4\}$.

The cloud computing service platform has a very large number and various types of cloud services. With time going by, the scale of cloud service continues to increase, which can easily lead to a decrease in the efficiency of the optimization of cloud service composition as well as bring great challenges to the high-efficient operation of cloud computing service platform. The method presented in this paper has the features of high efficiency and universality in the optimization of cloud service composition. It can effectively solve the problem of cloud service composition with high reliability and low resource consumption in cloud computing environment. Besides, it is also beneficial to improving the service efficiency of cloud computing service platform and helping quickly locate user's service demands, to find the optimal cloud service composition.

6. Conclusion

An optimization method for cloud service composition is proposed in this paper. Firstly, the performance indexes for the optimization of cloud service composition is introduced, including time, cost, availability, reliability, CPU occupation rate and bandwidth occupation rate. Secondly, the evidential reasoning method is adopted to filter the various cloud services based on the user's demands to form the candidate cloud service collection, so that the solution space of the cloud service composition optimization is greatly reduced. Thirdly, a multi-target planning model is established with the properties of the lowest CPU and bandwidth occupation rate and the highest reliability of the cloud service composition. The cloud services will be selected from the candidate cloud service collection to make a composition, and further determine the optimal cloud service composition by solving the model. Finally, the case study in travel planning scenario shows the proposed method can effectively solve the problem of cloud service composition with high reliability and low resource consumption in cloud computing environment.

Furthermore, this method can be applied to other fields by changing the performance indexes and service composition objectives, such as web service composition, cloud computing for education and learning, cloud healthcare and other cloud platforms. The optimization process and model of cloud service composition will be improved and the intelligent solution algorithm will be used to further improve the solving efficiency.

Reference

[1] C. Li, Y.C. Liu, X. Yan, Optimization-based resource allocation for software as a service

application in cloud computing [J]. Journal of Scheduling, 2017, 20 (1), pp.1-11.

[2] S. Wang, A. Zhou, F. Yang, R.N. Chang, Towards Network-Aware Service Composition in the Cloud [J]. IEEE Transactions on Cloud Computing, 2016 (99), pp.1-14.

[3] A. Jula, E. Sundararajan, Z. Othman, Cloud computing service composition: A systematic literature review [J], Expert Systems with Applications, 2014, 41 (8), pp.3809-3824.

[4] S.L. Wang, W.Y. Song, L. Kang, et al. Manufacturing resource allocation based on cloud manufacturing [J]. Computer Integrated Manufacturing Systems, 2012, 18(7), pp. 1396-1405.

[5] W.N. Liu, Y.M. Li, B. Liu, Service composition in cloud manufacturing based on adaptive mutation particle swarm optimization [J]. Journal of Computer Applications, 2012, 32(10), pp. 2869-2874, 2878.

[6] C.F. Lin, R. Sheu, Y.S. Chang, et al. A relaxable service selection algorithm for QoS-based web service composition [J], Information and Software Technology, 2011, 53(12), pp. 1370-1381.

[7] Y. Yang, H.X. Yao, J.M. Ye, et al. Leveraging ontology-aided AI planning for automatic composition of semantic web services [C], Proceedings of ICIII 2010, pp. 110-115.

[8] Y. Yang, S. Wang, R. Li, An efficient approach to optimization of service composition in cloud environment [C], Proceedings of ICCSPA 2017.

[9] Y. Liu, Research on service composition and selection in cloud computing [D], Doctoral dissertation of Beijing University of Posts and Telecommunications, 2013.

[10] Y. Yang, S.Q. Tang, Y.W. Xu et al. An approach to QoS-aware service selection in dynamic web service composition [C], Proceedings of ICNS 2007.