Resource Allocation Algorithm of Cloud Computing Infrastructure Services based on Continuous Optimization Algorithm

Wenyi Lu*
Jiangxi Vocational and Technical College of Communication, Nanchang, China

*Corresponding author e-mail: 81695311@qq.com

Abstract. The cloud environment is extremely complex, dynamic and changeable. How to use cloud services efficiently to create economic benefits for enterprises and even society is a matter that most scholars and manufacturers pay close attention to at present. Therefore, the resource allocation of cloud computing becomes the focus of research. The purpose of this paper is to further study the resource allocation algorithm of cloud computing infrastructure services. Firstly, the IAAAs mode in three main application modes of cloud computing is introduced in detail. Based on the understanding of optimization theory, the resource allocation algorithm of cloud computing infrastructure services with continuous optimization algorithm is established. The experimental results show that the number of iterations of the three algorithms is generally on the rise with the increase of the number of task nodes. In general, the number of iterations of ant colony algorithm is less than that of genetic algorithm, and the performance of continuous optimization algorithm is better than that of both, which shows that it can promote the overall efficiency of the algorithm.

Keywords: Continuous optimization algorithm, Cloud computing system, Infrastructure services, Resource allocation

1. Introduction

Cloud computing system is developed on the basis of traditional computer system, which is the inheritance and development of traditional PC system architecture. Therefore, before discussing the system architecture of cloud computing, we should first introduce the traditional PC system architecture. The traditional PC system has a clear hierarchical architecture [1-2]. The operating system layer refers to the operating system software used to manage the software and hardware. The software layer includes system software layer and application software layer [3]. The most critical is the operating system layer, whose task is to implement hardware resource management, software task management and file system management [4]. The system architecture of cloud computing can be divided into cloud basic resource layer, cloud operating system layer and cloud software layer, and the cloud software layer can be divided into cloud system software and cloud application software. Like the traditional PC system architecture, it takes the hardware resource layer as the support of the system,
which is located at the bottom of the system [5]. Infrastructure is service (IAAs). The service mode is based on virtualization technology. Cloud operators provide users with a large number of computing resources for their practical use through the Internet, including storage space, computing power, database, etc. [6]. In IAAs mode, the key technology of infrastructure as a service is the virtualization of server and the management and scheduling of basic resources. The purpose of virtualization is to change the organization of resources, and then improve resource utilization [7-9]. The role of resource management is to effectively allocate the massive resources in the cloud, effectively meet the needs of users, and provide support for the security and reliability of services [10].

Zijian Chao proposes a new cost-oriented optimization model, which is based on the information and communication technology infrastructure of cloud. Through numerical simulation, the effectiveness of the model to reduce the operation cost of cloud platform in demand side management is verified [11]. Jianbo Du solves the problem of computing offload in hybrid fog / cloud system by jointly optimizing the decision of offload and the allocation of computing resources, transmitting power and radio bandwidth, while ensuring the fairness and maximum tolerable delay of users. The optimization problem is formulated to minimize the delay between all UEs and the maximum weighted cost of energy consumption (EC). They propose a low complexity suboptimal algorithm to solve the problem, in which the unloading decision is obtained by semi definite relaxation and randomization, and the resource allocation is obtained by fractional programming theory and Lagrange dual decomposition. Simulation results verify the convergence performance and fairness between UEs of the proposed algorithm, as well as the performance gains in terms of delay, EC and effective UE number of existing algorithms [12].

This paper starts from the real problem of resource allocation of cloud computing infrastructure services, expounds the concept of resource allocation of infrastructure services, designs a continuously optimized resource allocation algorithm based on this, and carries out simulation experiments by setting parameters, and verifies the effectiveness and efficiency of the algorithm with specific examples and comparative analysis experiments. By comparing the data, we can see that with the increase of the network size, the number of iterations of the three algorithms generally shows an upward trend. In general, the performance of ant colony algorithm is better than that of genetic algorithm, the number of iterations is less, and the number of iterations of continuous optimization algorithm is better than that of both, which shows that it can promote the overall efficiency of the algorithm.

2. Proposed Method

2.1 IAAs Model Introduction

The application of PAAS service mode is still in the exploration stage. IAAs is the most important mode in cloud computing applications. In fact, both SAAS and PAAS can be developed on the basis of IAAs, so IAAs has a broader application prospect. The virtual machine allocation optimization algorithm in this paper is based on IAAs mode. In the following content, we will introduce the technical details of IAAs.

(1) IAAs application mode description

As mentioned before, IAAs is an application model that takes basic resources as services, and IAAs is recognized as the most potential service model. In short, in the implementation process of IAAs service, there is not much difference between the user's operation and access to ordinary Internet sites, and even the resources can be accessed through the browser. However, the difference with visiting ordinary websites is that what users get here is not simple information, but various basic resources including servers, storage, network bandwidth and computing power. Users use these resources online by paying for them.
In IAAs service mode, cloud providers provide users with virtual resources, which is also called “virtual machine”. Specifically, what users get is only a user name and corresponding login password, through the network login authentication.

IAAs mode has flexible resource supply capability. In the process of service, users pay according to the actual amount of resources used. After the end of occupancy, the resources they occupy will be released automatically and continue to invest in other services.

Under IAAs mode, the function of management platform is to integrate resources effectively, form a collection of resources that is convenient for unified management and control, and flexibly allocate and schedule these resources according to the needs of users, and at the same time carry out billing measurement of services.

2.2 Optimization Theory

(1) Optimization method

It is widely used in many fields such as natural science, social management, engineering application, etc. In short, optimization is to select the appropriate value of the variable to be determined on the basis of the variable value range and constraints, so as to obtain the best solution of the objective function expression. According to whether there are variable values and relationship constraints, it can be divided into two types: constrained programming and unconstrained programming.

2.3 Resource Allocation Algorithm

(1) Objective function

Cloud computing resource allocation has the characteristics of large-scale, dynamic adaptive and resource constraints. Based on these characteristics, we add several constraints to make the algorithm more close to the actual problem. These constraints are task processing time constraint, network bandwidth constraint and network delay constraint.

(2) Fitness function design

In this paper, we introduce the concept of fitness. Genetic algorithm requires that the fitness function must be the maximum value, and the minimum value of the total cost under the constraints of cloud computing resource allocation should be converted into the maximum value function according to the non-negative principle of the fitness function. In order to meet the principle of “survival of the fittest” of genetic algorithm and select excellent chromosomes, we define the fitness function as:

$$f(i) = \left( \frac{\sum_{k=1}^{M} (\text{totaltime}(k) - Mn + 1)}{M \times (\text{totaltime}(i) - Mn + 1)} \right)^2, i = 1, 2, ..., M$$  \hspace{1cm} (1)

(3) Genetic code

Genetic algorithm is used to solve the resource allocation problem. The decision variables of the problem are transformed into genes, and then the genes are encoded. The chromosomes formed by the encoded genes are generated into new chromosomes through crossover and mutation operations, so that better chromosomes can be selected under certain conditions. Finally, the chromosomes are decoded and the most original decision variables are returned to explain the resource allocation The situation. Common coding methods include direct coding and indirect coding. Due to the large scale of resources and complex constraints in the cloud environment, the selection of coding methods will directly affect the performance and efficiency. Follows the coding principles of completeness, soundness and non-redundancy. Compared with other coding techniques, real coding directly codes the solution space of the problem variable, which is convenient to introduce the specific field information, and greatly shortens the coding length, facilitates the breadth search, improves the solution accuracy, and guarantees the algorithm efficiency. Therefore, this paper selects real coding technology for coding.

(4) Generate initial population
There are two steps to generate the initial population: the first step is to generate the left string of one-dimensional string, that is, to randomly assign the subtask to a resource; the second step is to generate the right string of one-dimensional string.

(5) Crossover and mutation operation

In this paper, the consistent crossover strategy is adopted, that is, every gene on the chromosome is randomly crossed according to the same probability.

(6) Decoding and obtaining the optimal solution

The optimal solution obtained by genetic algorithm is a chromosome string. By decoding the string, the optimal scheme of resource allocation is obtained. Decoding is the inverse process of encoding. Assuming that the optimal solution is: (111222 123456), the allocation scheme is:

\[ Y_1: X_1 \rightarrow X_2 \rightarrow X_3 \]  
\[ Y_2: X_4 \rightarrow X_5 \rightarrow X_6 \]  

(2)  
(3)

3. Experiments

3.1 Experimental Evaluation

In order to evaluate the resource allocation algorithm of cloud computing infrastructure services in the continuous optimization algorithm environment proposed in this chapter, two simulation experimental environments (deterministic demand simulation and fixed reservation virtual machine volume simulation) are designed in this section.

Table 1. Experimental basic configuration

| Hardware       | DELL OPTIPLEX330 machine with 2 Intel E2180 2.00GHz processors, and 1GB RAM |
|----------------|---------------------------------------------------------------------------------|
| Software       | Linux, kmem 2.6.38                                                              |
| Operating system | Matlab 7.10                                                                     |

The hardware environment of this experiment is Dell optiplex330 machine with 2 Intel E2180 2.00GHz processors, and 1GB ram. The operating system is Linux, and the software is matlab 7.10. See Table 1 for the details of experimental basic configuration. Chinese style

3.2 Algorithm Simulation Analysis

To create a virtual cloud computing network, there are 9 tasks to be processed, and 4 processors can process tasks at the same time, with a maximum depth of 9. The specific depth value is determined according to the generated relationship between the past and the future.

3.3 Parameter Setting of Resource Allocation Algorithm for Cloud Computing Infrastructure Services

This paper sets the basic parameter values of the resource allocation algorithm for cloud computing infrastructure services, as shown in Table 2, referring to the model with the most resource allocation applications at present. The left side of the table is the parameter value setting of genetic algorithm, and the right side is the parameter value setting of slave group algorithm.

Table 2. Basic parameter values of resource allocation algorithm

| (Parameter) symbol | Genetic algorithm | Ant algorithm |
|--------------------|-------------------|--------------|
| Population size M  | 30                | Ant number   | 8            |
| Crossing probability PC | 0.6              | Iterations maxinter2 | 500          |
| Probability of variation PM | 0.05       | Pheromone heuristics a | 3           |

4. Discussion
4.1 Algorithm Comparison and Analysis

In order to test the performance of continuous optimization algorithm, in the simulated cloud computing experiment, multiple tasks and resource points are used to demonstrate the hybrid optimization algorithm. Table and figure show the experimental data of continuous optimization algorithm. It should be noted that the parameter settings of individual genetic algorithm is consistent with those of continuous optimization algorithm.

(1) Comparison of iterations

When the number of task nodes is between 20-100. When the number of tasks is in the range of [20,40], the genetic algorithm has a faster convergence speed, but with the increase of the number of iterations, the convergence speed is slower and slower; the characteristics of ant colony algorithm is just the opposite, because the early lack of initial pheromone, the convergence speed of the algorithm is slower, but with the accumulation of later pheromone, the task processing speed is faster and faster. The task allocation time of continuous optimization algorithm is much better than the former two algorithms.

Table 3. Task allocation time comparison of three algorithms

| Number of task nodes | Ant colony algorithm | Genetic algorithm | Continuous optimization algorithm |
|----------------------|----------------------|-------------------|-----------------------------------|
| 20       | 123                  | 68                | 12                                |
| 30       | 368                  | 290               | 210                               |
| 40       | 998                  | 876               | 305                               |
| 50       | 1869                 | 2001              | 1120                              |
| 60       | 2600                 | 4000              | 1580                              |
| 70       | 3768                 | 6986              | 1809                              |
| 80       | 6898                 | 10287             | 3890                              |
| 90       | 9786                 | 19084             | 4587                              |
| 100      | 15879                | 38890             | 5800                              |

With the increase of the number of task nodes, the number of iterations of the three algorithms tends to increase. In general, the performance of ant colony algorithm is better than that of genetic algorithm, and the number of iterations of continuous optimization algorithm is also better than that of genetic algorithm, which shows that it can improve.

(2) Load data comparison

When the number of tasks is shown in Figure 1. From the resource load level of ant colony algorithm, genetic algorithm and continuous optimization algorithm. In order to clearly see the load balance level of each resource point, we draw the resource load balance comparison diagram of three algorithms. It can be seen from the figure that the load of ant colony algorithm and genetic algorithm is large at the resource point, and small at the resource point, while the load is relatively balanced, which shows that the overall load balancing performance.
5. Conclusions

The function of resource management is to effectively allocate the massive resources in the cloud, effectively meet the needs of users, and provide support for the security and reliability of services. This paper first describes the pseudo code implementation of the resource allocation algorithm, and uses programming language to simulate the algorithm, and under certain parameter assumptions, designs a specific example, and carries out simulation experiments through the set parameters, and verifies the effectiveness and efficiency of the algorithm by specific examples and comparative analysis experiments.

References

[1] Jyotiska Nath Khasnabish, Mohammad Firoj Mithani, Shrisha Rao. Tier-Centric Resource Allocation in Multi-Tier Cloud Systems[J]. IEEE Transactions on Cloud Computing, 2017, 5(3):576--589.
[2] Li-Der Chou, Hui-Fan Chen, Fan-Hsun Tseng. DPRA: Dynamic Power-Saving Resource Allocation for Cloud Data Center Using Particle Swarm Optimization[J]. IEEE Systems Journal, 2018, 12(2):1554-1565.
[3] YOU, C, Huang, K, Chae, H. Energy-Efficient Resource Allocation for Mobile-Edge Computation Offloading[J]. IEEE Transactions on Wireless Communications, 2017, 16(3):1397-1411.
[4] Wenyun Dai, Longfei Qiu, Ana Wu. Cloud Infrastructure Resource Allocation for Big Data Applications[J]. IEEE Transactions on Big Data, 2018, 4(3):313-324.
[5] J. Zhang, N. Xie, W. Li. Truthful Multi Requirements Auction Mechanism for Virtual Resource Allocation of Cloud Computing[J]. Journal of Electronics & Information Technology, 2018, 40(1):25-34.
[6] Bo Li, Songtao Guo, Yan Wu. Construction and Resource Allocation of Cost-Efficient Clustered Virtual Network in Software Defined Networks[J]. Journal of Grid Computing, 2017, 15(4):1-17.
[7] Jalal Khamse-Ashari, Ioannis Lambadaris, George Kesidis. An Efficient and Fair Multi-Resource Allocation Mechanism for Heterogeneous Servers[J]. IEEE Transactions on Parallel and Distributed Systems, 2018, 29(12):2686-2699.
[8] Lei Wei, Jianfei Cai, Chuan Heng Foh. QoS-Aware Resource Allocation for Video Transcoding in Clouds[J]. IEEE Transactions on Circuits & Systems for Video Technology, 2017, 27(1):49-61.
[9] Chase J., Niyato D. Joint Optimization of Resource Provisioning in Cloud Computing[J]. IEEE Transactions on Services Computing, 2017, 10(3):396-409.
[10] Wei L., Foh C H., He B., et al. Towards Efficient Resource Allocation for Heterogeneous Workloads in IaaS Clouds[J]. IEEE Transactions on Cloud Computing, 2018, 6(1):264-275.
[11] Zijian Cao, Jin Lin, Can Wan. Optimal Cloud Computing Resource Allocation for Demand Side Management in Smart Grid[J]. IEEE Transactions on Smart Grid, 2017, 8(4):1943-1955.
[12] Jianbo Du, Liqiang Zhao, Jie Feng. Computation Offloading and Resource Allocation in Mixed Fog/Cloud Computing Systems with Min-Max Fairness Guarantee[J]. IEEE Transactions on Communications, 2017, 66(4):1-1.