Resource Allocation in Cloud Computing Environment based on NSGA-II

Hesheng Gong
College of Economics and Management
Nanjing University of Aeronautics and Astronautics
Nanjing, China
15150591001m@sina.cn

2021 2nd International Conference on Computing and Data Science (CDS) | 978-1-6654-0428-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/CDS52072.2021.00014

***Abstract*—In the cloud computing environment, physical resources are abstracted into the same virtual resource, but the allocation of virtual resources also needs to be scheduled to make the entire allocation system more reasonable, rapid, efficient and minimize the total cost. How to schedule virtual resources to physical resources is a basic and complex problem in cloud computing. The scheduling of virtual resources is modeled and proved difficult to solve. The solution of the model is transformed into a multi-objective optimization problem with system load balancing as the optimization objective, and an improved genetic algorithm based on non-dominated sorting (NSGA-II) is proposed. To solve this problem, genetic algorithm is used to finally map out the best scheduling scheme through the termination population condition. Compared with scheduling algorithms for specific environments, abstract models are more representative of virtual resource scheduling problems in typical cloud computing environments. First, the proposed model was simulated, and the quota of different resources of each virtual machine was restricted when the initial population was generated; the load balancing idea was considered when designing the fitness function, and based on this, the formula for the measurement standard was carried out. The experimental results show the effectiveness of the model and the feasibility of the NSGA-II algorithm to solve the problem.**

Keywords—Cloud computing, multi-objective optimization, non-dominated sorting genetic algorithm, scheduling

I. INTRODUCTION

Cloud computing is a business service model and computing model that distributes computing tasks in different data centers composed of a large number of computers, enabling various application systems to obtain storage space, information services, and computing capabilities as needed. The service usage model of cloud computing enables computing power to be circulated as a commodity, just like currency, water, and electricity, with convenient access and low cost [1-5]. The biggest difference is that it is transmitted through the Internet. How to effectively schedule and allocate virtual resources of cloud computing data centers in a dynamic environment according to user needs is a key technology. Advanced dynamic resource scheduling is of great significance for improving the utilization efficiency of computing resources of various educational, administrative and research institutions and enterprises, saving resources, improving resource sharing and reducing operating costs, and is worthy of in-depth research.

At present, on the one hand, the dynamic allocation and scheduling of physical and virtual resources in the data center is facing many new challenges: real-time dynamic changes in user requirements are difficult to accurately predict; physical and virtual resources are widely distributed and diverse; system performance and cost need to be considered, these considerations make the problem very complicated. Therefore, there is an urgent need for efficient resource scheduling algorithms to adapt to different business needs and business goals. On the other hand, when considering the allocation and migration of dynamically adjustable virtual machines and the overall performance of physical machines, how to solve the problem of inconsistent user needs when considering resource factors such as CPU, storage, and network instead of a single factor. Taking into account a series of characteristics of cloud computing and the ability to efficiently allocate appropriate computing resources for user operations in this environment, the cloud environment service cluster must provide a better resource recommendation for each resource demander. According to the characteristics of cloud computing environment, this paper analyzes the impact of factors such as resource utilization, overhead cost and response time on scheduling, proposes a multi-objective comprehensive evaluation model for virtual machine resource scheduling, and applies typical multi-objective evolutionary algorithms to this problem.

Traditional scheduling algorithms use Min-Min, Max-Min, polling algorithms, and simulated annealing algorithms. The advantage of using the Min-Min algorithm lies in its rapid execution. However, these two algorithms have shortcomings: the former will cause node overload or almost no workload, which does not meet the optimal state of the entire system network load balance, and the algorithm does not have adaptability and scalability, and the scope of application Narrow; the latter is better than the former, but changes in external conditions have a greater impact on the advantages of the algorithm, and its advantages are not stable. The advantage of the polling algorithm is that it is not complicated to implement, but when the server software and hardware performance is different, the algorithm is likely to cause unbalanced load when the server time difference is large. The advantage of the simulated annealing algorithm is that it avoids the algorithm from falling into the local minimum dilemma and can eventually tend to the global best result, but there will be a problem of longer execution time.

In this paper, genetic algorithm is used to analyze the problem, considering that genetic algorithm is more convenient to write and the logical structure is concise; in addition, parallel searching is based on the group, and comparison among multiple entities is convenient to find the best individual; in a probabilistic environment Iteratively, it has enough randomness; it has excellent global search ability; finally, genetic algorithm has good scalability, making it easy to combine with other algorithms to solve more complex problems.

978-1-6654-0428-0/21/$31.00 ©2021 IEEE
DOI 10.1109/CDS52072.2021.00014

II. Related Work

Nowadays, the demand to the storage, operation and analysis of data are continuously increasing under the environment that Big Data is getting more and more popular among various fields. Cloud computing and Cloud manufacturing then appeared in order to fit the urgent needs.

A. *Cloud Computing*

Since Cloud computing was created by Google for the first time, the concept has been applied and improved by a huge amount of companies combining with their own situation. The Amazon cloud computing platform architecture based on EC2 elastic cloud computing and many other storage services of the Amazon company have been using virtual computing resource. In addition, he also introduced the core virtual computing resource scheduling and its algorithm based on Hadoop MapReduce framework established by IBM. Chrysa Papagianni et.al described cloud computing as a service, such as SaaS (software as a service), IaaS (infrastructure as a service) and PaaS (platform as a service), to offer technical support needed in various subjects. However, there exists the need to minimize the cost within the process of allocating the resources [6]. The OCRP (an optimal cloud resource provisioning) algorithm which was proposed by Sivadon Chaisiri and his team can optimize the model efficiently and, to some extent, solve the problem [7].

B. *Cloud Manufacturing (CM)*

Under the trend of the returning of manufacturing industry throughout the world, the low-cost manufacturing in China which used to be an advantage is suffering from the stress from the various subjective and objective problems. Traditional manufacturing model does not fit the industry condition as well as it did before, while the cloud manufacturing model based on the Internet which also combines many emerging technologies such as the Internet of things, cloud computing and information manufacturing suits the trend better. Except for the statement above, Guo et.al summarized Cloud manufacturing architecture which consists of logistics resource layer, virtual resource layer, core service layer, application interface layer and application layer and concluded the six main characteristics of CM proposed earlier [8].

The concept of CM was firstly formed in China in [9] which described the definition, architecture and key technologies in detail. The research on CM abroad was started from [10], which created a new cloud manufacturing prototype system—ICMS (Cloud-based manufacturing system). After that, Wu et.al built various models matching different situations through research and innovation, and several different VMs models were used to analyze the allocation methods among various Actors from the perspective of SaaS model [11].

III. Mathematical Formulation

Based on the above information, the model can be established in the following steps:

- The model defines the utilization rate as the percentage corresponding to the ratio of the task's demand for a certain resource at the node to the amount of the resource provided by the resource node to the virtual machine. That is, the following definition calculation of the utilization rate.

Virtual machine CPU utilization

$$L_{CPU} = \frac{c_i}{c_j} \cdot 100\% \tag{1}$$

Virtual machine memory utilization

$$L_{\mathrm{mem}} = \frac{m_i}{M_j} \cdot 100\% \tag{2}$$

Virtual machine storage space utilization

$$L_{DS} = \frac{d_i}{D_j} \cdot 100\% \tag{3}$$

- According to the utilization and total amount of different resources, the effective influence proportion Z of parameter a can be set.

$$Z_a = L_a \cdot \frac{J_a}{J_{total}} \cdot 100\% \tag{4}$$

which are,

$$Z_{CPU} = L_{CPU} \cdot \frac{C_j}{C_{total}} \cdot 100\% \tag{5}$$

$$Z_{\mathrm{mem}} = L_{mem} \cdot \frac{M_j}{M_{total}} \cdot 100\% \tag{6}$$

$$Z_{DS} = L_{DS} \cdot \frac{D_j}{D_{total}} \cdot 100\% \tag{7}$$

- For the entire virtual machine cluster, there is another definition of the influence of virtual machine i as Y.

$$Y_{\mathrm{i}} = k_1 \times Z_{CPU} + k_2 \times Z_{mem} + k_3 \times Z_{DS} \tag{8}$$

- Total node resource utilization of each virtual machine.

$$V = (k_1 \times \frac{c_i}{c_j} + k_2 \times \frac{m_i}{M_j} + k_3 \times \frac{d_i}{D_j}) \times 100\% \tag{9}$$

In the third and fourth parts, k_1、 k_2、 k_3 show the weight coefficients of CPU, memory and storage space.

- Since the CPU specifications of the virtual machine directly affect the calculation speed of the virtual machine, that is, the speed of completing the task, set α as the correlation coefficient, and the calculation speed can be expressed as αC_{j}, and the time T

$$T(j)=\frac{\mathrm{m}_i}{\alpha C_{\mathrm{j}}} \tag{10}$$

Because the environment is set to turn on and off all virtual machines at the same time, the running time of the physical machine is the longest time required as $T_{\max}$.

Calculate the consumption of physical machines and virtual machines. According to the VRSA-EO model, the energy consumption of the physical machine can be obtained

$$P_u = k_0 \cdot P_{max} + (1-k_0) \cdot P_{max} \cdot L_{CPU}(t) \tag{11}$$

Among them $P_{\max}$ is the maximum energy consumption of the physical machine during operation, and the percentage of energy consumption of the physical host during no-load operation is k_0 while $\mathrm{k}_0 \approx 0.7$.

The energy consumption of the physical machine at t can be shown as

$$E=\int_t P_{\mathrm{u}}(t)dt = \int_t P_{\max} \times (k_0 + (1-k_0) \times L_{CPU}(t))dt \tag{12}$$

The cost of migration can be defined by

$$U_{\mathrm{di}} = 0.1 \cdot \int_{t_0}^{t_0+T_{mj}} L_{CPU}(t)_j \, dt \tag{13}$$

While,

$$L_{CPU}(t)_j = \sum_{i=1} \frac{CPU_i(t)}{CPU_{total}} \tag{14}$$

$$T_{\mathrm{mj}} = \frac{M_j}{B_j} \tag{15}$$

Effective cost calculation. After integrated calculation, the formula is obtained. The virtual machine consumption is equal to the sum of the virtual machine power consumption and the virtual machine migration consumption, and the physical machine consumption is the boot cost. The calculation formula for the virtual machine and the physical machine consumption can be obtained as

$$VF=\mathrm{W} \times .T + X. \times U_{di} \tag{16}$$

$$PF = PS + E + \mathrm{T}_{\max} \times W. \times Amount \tag{17}$$

And then get the primary total cost calculation method

$$F = VF + PF \tag{18}$$

Define the effective cost set FE as the partial cost of all virtual machines under their respective influence proportions and comprehensive utilization. The accumulation of all items in the set is the total effective cost of the virtual machine cluster

$$FE = F \times V \times Y \tag{19}$$

IV. Resource Scheduling Solution Based on NSGA-II

The scheduling problem in the cloud environment has the characteristics of large data volume and high real-time requirements. Traditional algorithms are difficult to apply in the cloud environment and cannot meet the needs of all parties. The use of intelligent algorithms can process large amounts of data more reasonably and smoothly, and improve computing efficiency and speed. Especially after the company's business expansion in the later period, when data is growing exponentially, intelligent algorithms can better reflect its advantages.

The following figure shows the basic algorithm flow of genetic algorithm. The main parts include: initial population generation operation, selection operation, crossover operation, mutation operation, and population iteration. In order to facilitate each operation, the chromosome can be compiled in binary or decimal, that is, the operation of encoding and decoding. The complete operation process is shown in Fig. 1.

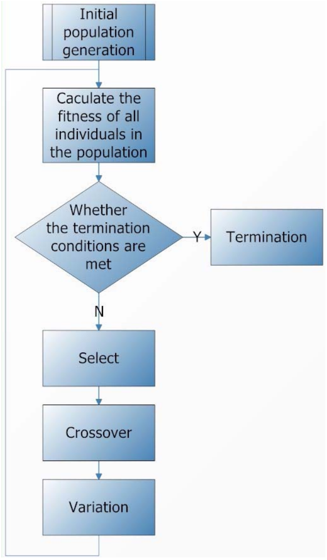
Initial population generation
Caculate the fitness of all individuals in the population
Whether the termination conditions are met
Y
Termination
N
Select
Crossover
Variation

Fig. 1. Genetic algorithm operation process.

The selection operation is to select the chromosomes or genes that meet the required conditions in the population, and use them as the parent population before the crossover. The selection method is to combine the different resource utilization rates of each individual with a certain weight ratio (the ratio in this article is 7:2:1) to synthesize the comprehensive utilization rate of an individual. The array formed is the population fitness set, and each individual The ratio of the comprehensive utilization rate to the total comprehensive utilization rate of the population is sorted by size, and the area of the fan-shaped area is set according to the ratio of each adaptive degree to the total fitness according to the roulette method, as shown in Figure 2. Generate an equal amount of random numbers. As the area of the sector gradually increases, the greater the probability that the random numbers are included, it plays a role in selection.

Crossover is to randomly generate crossover positions, and perform crossover operations on different individuals at specific gene positions to generate new individuals and populations. The new population gold set is used as the parental individuals and the crossover positions of the two parental chromosomes are randomly generated, and the two genes are swapped for crossover operation, as shown in Fig. 3.

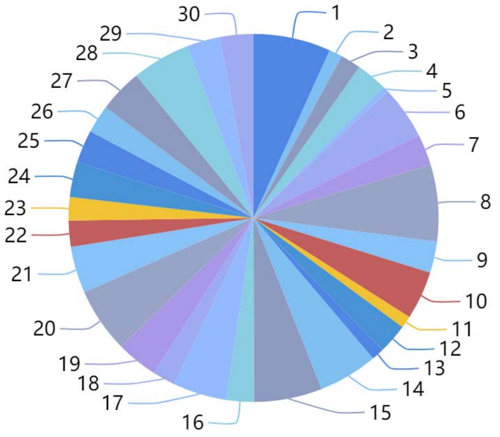

Fig. 2. Roulette example

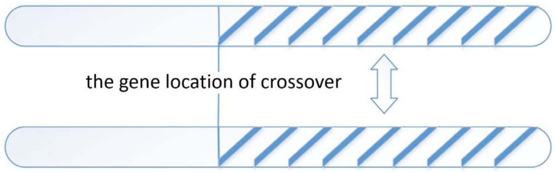

Fig. 3. Schematic diagram of cross operation.

The mutation operation is to calculate the number of individual variant loci, and randomly select the corresponding number of loci, randomly change the original genes, and generate new individuals and populations. Randomly generate the position of the mutant gene in the three chromosomes, and then randomly generate an integer power of 2 within the value range to replace the gene at the locus of the original chromosome to complete the mutation operation, see Fig. 4.

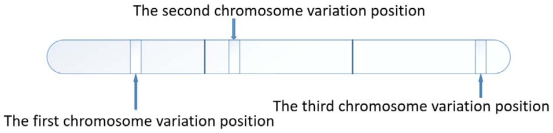
The second chromosome variation position
The first chromosome variation position
The third chromosome variation position

Fig. 4. Schematic diagram of mutation operation.

The new population generated by the selection operation, crossover operation and mutation operation is used as the initial population of the next iteration process, and iterates repeatedly until the termination condition of the set number of iterations is reached. At this time, the population is the terminated population finally formed after multiple generations of selection, crossover and mutation operations. Calculate the comprehensive utilization rate of the population at this time, that is, the final fitness matrix.

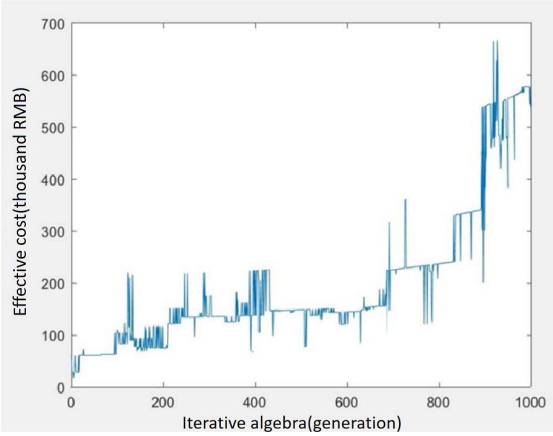

Fig. 5. Generate the optimal fitness result distribution map.

V. Experimental Results

First, input the task demand matrix R and related basic parameters as the basic data of the algorithm. Set the initial population popsize to 30 and the number of iterations gene to 1000. Set selection probability ps, crossover probability pc, and mutation probability pm to help complete crossover and mutation operations.

In the research process, a scheduling method based on traditional scheduling theory, task distribution theory and load balancing theory was selected. Compared with the scheduling method and distribution plan used by Z company, the genetic algorithm can get more optimized satisfaction. Solution, compare the scheduling plan corresponding to the obtained satisfactory solution with the original resource scheduling plan of Company Z. The total usage of CPU, memory and storage and the effective cost of the two plans are shown in Table I.

After comparison, it can be concluded that the optimized solution is not only more economical in terms of resource usage, but its effective cost is also higher than that of the original solution, so the optimized solution has better performance than the original solution in many aspects.

TABLE I. PROGRAM EVALUATION TARGET COMPARISON TABLE

	Original Plan			**Optimized Plan**		
Resource	***CPU***	***Memory***	***Storage***	***CPU***	***Memory***	***Storage***
Total use	1156	3018	11545	1013	2822	11175
Effective Cost (ten thousand RMB）	1172.648			1224.349		

REFERENCES

[1] Varghese, Blesson, and Rajkumar Buyya. "Next generation cloud computing: New trends and research directions." Future Generation Computer Systems 79 (2018): 849-861.

[2] Deng, Ruilong, et al. "Optimal workload allocation in fog-cloud computing toward balanced delay and power consumption." IEEE internet of things journal 3.6 (2016): 1171-1181.

[3] Ooi, Keng-Boon, et al. "Cloud computing in manufacturing: The next industrial revolution in Malaysia ?." Expert Systems with Applications 93 (2018): 376-394.

[4] Samimi, Parnia, Youness Teimouri, and Muriati Mukhtar. "A combinatorial double auction resource allocation model in cloud computing." Information Sciences 357 (2016): 201-216.

[5] Zhou L, Zhang L, Zhao C, Laili Y, Xu L. Diverse task scheduling for individualized requirements in cloud manufacturing. Enterprise Information Systems. 2018 Mar 16;12(3):300-18.

[6] Merlino, Giovanni, et al. "Mobile crowdsensing as a service: a platform for applications on top of sensing clouds." Future Generation Computer Systems 56 (2016): 623-639.

[7] Chase, Jonathan, et al. "A scalable approach to joint cyber insurance and security-as-a-service provisioning in cloud computing." IEEE Transactions on Dependable and Secure Computing (2017).

[8] Guo, Liang, and Jingxiong Qiu. "Optimization technology in cloud manufacturing." The International Journal of Advanced Manufacturing Technology 97.1-4 (2018): 1181-1193.

[9] Li BH, Zhang L, Wang SL, Tao F, Cao JW, Jiang XD, Song X, Chai XD. Cloud manufacturing: a new service-oriented networked manufacturing model. Computer integrated manufacturing systems. 2010 Jan 16;16(1):1-7.

[10] Xu X. From cloud computing to cloud manufacturing. Robotics and computer-integrated manufacturing. 2012 Feb 1;28(1):75-86.

[11] Wu, Linlin, et al. "SLA-based resource provisioning for hosted software-as-a-service applications in cloud computing environments." IEEE Transactions on services computing 7.3 (2013): 465-485.