A Review of Virtual Resource Management Research in Cloud Data Centers

Zhijie Han
School of Software, Henan University
KaiFeng, China
https://orcid.org/0000-0002-7362-7520

Guang Yang
School of Software, Henan University
KaiFeng, China
240416127@qq.com

Fang Zuo
Henan International Joint Laboratory of Intelligent Network and Key Technology, Henan University
Software Engineering Intelligent Information Processing Innovation Base-Subject Innovation Base of Henan Higher Universities
KaiFeng, China
https://orcid.org/0000-0001-5673-8870

***Abstract*—Cloud computing data centers represent the direc-tion of data center architecture and development. Virtual ma-chine (VM) management is one of the primary research directions for resource management in cloud data centers. Optimization of VM scheduling often has a significant impact on resource man-agement. This article aims to comprehensively investigate and analyze the current development of VM scheduling. Starting with VM scheduling-related technologies, the article further analyzes VM deployment, optimization methods for VM scheduling, and goals for VM scheduling optimization from three perspectives. Finally, the article provides an analysis and summary of the research trends and directions in VM scheduling.**

Keywords— Cloud data center, Virtualization, Virtual Ma-chine, Scheduling optimization

I. INTRODUCTION

In today's digital age, cloud computing, a revolutionary computing model, is increasingly becoming the preferred way for businesses and individuals to process and store massive amounts of data. Cloud computing greatly improves the efficiency of data processing and storage by centralizing computing resources in data centers and providing them as a service to users. And cloud data center VM management, as one of the core technologies of cloud computing, is committed to the efficient use of resources and elastic allocation, further promoting the development of cloud computing. However, with the continuous evolution of cloud computing application scenarios and the continuous progress of technology, how to reasonably schedule and manage VM under the premise of ensuring performance and reliability has become a core issue that needs to be urgently solved in the field of cloud data center VM management. Cloud data center VM management still needs more comprehensive and in-depth research. Therefore, the purpose of this paper is to provide an overview of the current research status of cloud data center VM management, summarize the existing results and problems, and look forward to the future development direction. Through in-depth research on cloud data center VM management, we can further optimize resource utilization, improve system performance, and provide users with better computing services and experiences.

II. RELATED TECHNOLOGIES

A. Cloud Data Center

Cloud computing is a novel computing model based on virtualization technology, on-demand payment as a business model, and characterized by elasticity, dynamicallocation, and resource sharing. The services provided by cloud computing can be divided into three categories: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). Companies such as Amazon, Google, and Salesforce have established massive data centers worldwide to accom- plish extensive computing tasks and store vast amounts of information, thereby successfully providing cloud computing services in commercial applications. The rapid development of cloud computing technology has driven the transformation of traditional data centers, giving rise to a new generation of data centers — cloud data centers.

Cloud data center is the most important infrastructure for cloud computing, which generally consists of a large number of physical machines (PM). In order to provide diversified services, virtualization technology has been applied in a large number of applications. Therefore, arranging VM on PM to provide various types of services for users has become a new way. Cloud data centers build server resources (CPU, memory.

B. Resource Virtualization

Due to differences in hardware equipment in data centers, compatibility issues arise, posing a challenge to unified resource management. To solve this problem, resource abstraction and virtualization serve as the basis for establishing virtual computing environments. Reference [1] conducts indepth research on resource virtualization models and methods, proposing an autonomous element resource abstraction model featuring environment-dynamic awareness and autonomous behavioral decision-making capabilities to address inconsistent access due to resources' diversity and autonomy. In addition, reference [2] achieves a virtual cluster system through virtualization, eliminating coupling between system software and hardware, thereby enabling rapid deployment and switching of clusters and achieving significant performance advantages. Cloud data centers use virtualization technology to construct a virtual resource pool, enabling effective and unified management of large-scale basic resources. Then, computing tasks are distributed across a large number of dynamic and scalable virtual resource pools, allowing users to obtain computing power, storage space, and information services according to their needs.

Virtualization technology abstracts and isolates physical resources to create a virtual computing environment. It enables flexible management and utilization of resources by dividing a physical server into multiple independent VM instances and running different operating systems and applications in each VM instance,as shown in figure 1,VMware is a leader in cloud computing and virtualization technology.

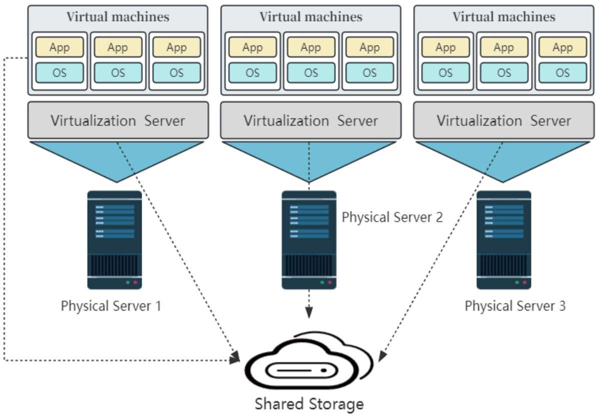

Fig. 1. . Cloud data center resource virtualization

Their products, including vSphere and vCloud Suite, enable virtualization of server, storage, and network resources. By abstracting physical resources into virtual machines and virtual networks, users can flexibly configure and manage resources on a unified management platform and realize resource sharing and automation.AWS is one of the world's largest cloud service providers. Their Elastic Compute Cloud (EC2) service uses virtualization technology to divide physical servers into multiple VM instances, which are made available to users through the cloud platform. Users have on-demand access to computing resources as needed and can configure and monitor resources through the AWS management console. As another leading cloud service provider, Azure offers a variety of cloud computing services with virtualization of resources.Azure Virtual Machines allows users to run applications on VM, while Azure Virtual Networks provides management and configuration capabilities for virtual networks. All of these services are based on virtualization technology and allow users to easily access and manage resources in the cloud

III. VIRTUAL MACHINE DEPLOYMENT

Rapid resource deployment is a critical functional requirement in data centers. After adopting virtualization technology, cloud computing data centers need to build virtual resource pools and deploy VM on different physical hosts to effectively and uniformly manage large-scale infrastructure resources,as shown in figure 2. Cloud computing environments have high demands for VM deployment, aiming to achieve efficient, fast, energy-saving, low-consumption, and load balancing deployment, maximizing the utilization of computing, storage, and network resources. VM deployment is a complex problem. Firstly, in cloud environments, the range of resource and application changes is wide and highly dynamic, and userrequired services primarily adopt on-demand deployment. Secondly, different cloud data centers and different levels of cloud computing environments have varied deployment patterns for services, with diverse supported software system forms and varying system architectures, resulting in diverse deployment strategies.

The low-cost and low-carbon VM deployment strategy. High energy consumption and carbon emissions are two important issues in distributed cloud computing. Due to the geographical distribution of data centers (DCs), distributed cloud computing needs to comprehensively consider various resources, energy prices, and carbon emission rates.

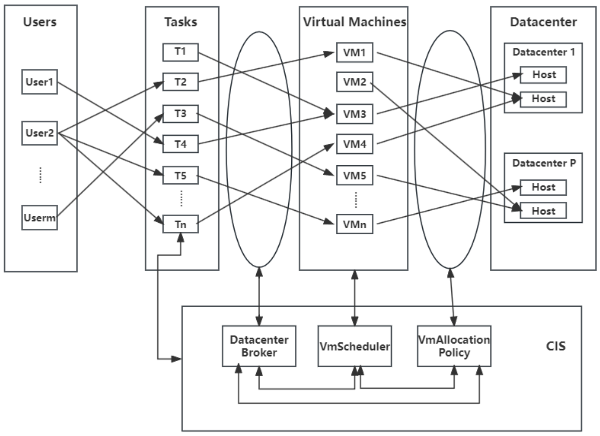

Fig.2. Virtual machine deployment process

Therefore, in VM deployment, the comparison of cost and carbon efficiency is particularly important, especially when compared with centralized cloud computing. Reference [3] proposes a low-cost and low-carbon distributed cloud VM deployment strategy that combines prediction-based A* algorithm with fuzzy set technology to optimize supplier's cost and carbon emissions decision-making, considering different energy prices and carbon emission rates geographically, and simultaneously optimizing network and server resources.

A fault-tolerant VM deployment strategy. With the introduction of virtualization technology, the failure of host server systems has become a critical concern [4]. Since VM are built upon physical devices and virtualization platforms of host servers, any failures occurring on the host servers will affect all running VM, necessitating the development of corresponding response strategies. In reference [5], a method of deploying redundant configured VM is proposed. This method evaluates the minimum number of VM required based on the performance needs of the applications and determines optimized VM deployment decisions to minimize the impact caused by the failure of any k host servers.

A network-based VM deployment strategy. VM on different PM in cloud data centers require frequent data communication, thus considering factors such as network bandwidth and latency becomes essential when evaluating application performance. However, traditional VM deployment strategies primarily focus on efficiency and utilization of computational resources, neglecting network-related factors. In reference [6], a network-based VM deployment strategy is proposed with the aim of minimizing data transmission time between VM to optimize overall application performance.

IV. . VIRTUAL MACHINE SCHEDULING METHODS

A. . Optimization Methods Based on Operations Research

Operations research is frequently employed to tackle complex problems in real-life scenarios, particularly to improve or optimize the efficiency of existing systems. In the context of VM scheduling, mathematical modeling is commonly used to determine the optimal mapping between virtual machines and hosts, followed by the design of efficient algorithms to solve this problem. Within the realm of current operations research approaches, techniques such as linear programming, dynamic programming, and stochastic programming have been extensively applied to optimize VM scheduling problems.

Based on linear programming, Charitopoulos,Papageorgiou and Dua [7] proposed an algorithm for analyzing and solving the multiple-parameter mixed-integer linear programming (mp-MILP) problem under global uncertainty (i.e., RHS, OFC, and LHS). The algorithm computes exact explicit solutions and corresponding regions in the parameter space. Chen and Liu [8] introduced a two-tier VM placement framework based on linear programming. Their LP-based oblivious mutation VM placement algorithm generates VM placements with minimum energy consumption. Additionally, the FeasibilityDriven Stochastic Virtual Machine Placement (FDSP) algorithm seamlessly collaborates with LP-based approaches to achieve desirable feasible placements. Lopez,Kushik and Zeghlache [9] demonstrated how to formulate and leverage integer linear programming problems to derive test suites and optimal solutions, providing judgments on the quality of VM placement implementations.

Dynamic programming is an approach that decomposes a problem into overlapping subproblems to obtain the optimal solution for the original problem. It is commonly applicable to problems that exhibit both overlapping subproblems and optimal substructure, and VM scheduling optimization is no exception. Zhang,Wu and Chen [10] proposed a VM selection algorithm based on greedy algorithms and dynamic programming. Experimental results have demonstrated that this algorithm effectively reduces energy consumption while satisfying SLA constraints.

Stochastic programming is an optimization method that considers uncertainties. It probabilistically models decision variables and stochastic parameters, and utilizes relevant algorithms to solve for optimal strategies. It supports decisionmakers in making better decisions in uncertain environments and is commonly used to address VM scheduling optimization problems. In the case of fluctuating VM workloads, Nandi,Banerjee and Ghosh [11] proposed a stochastic model for optimizing data center consolidation by formulating the problem as a stochastic integer programming problem. They also introduced intelligent decision-making for statistical VM reuse on PM to ensure optimized hardware resource utilization while providing service guarantees. Chaisiri,Lee and Niyato [12] presented an Optimal Virtual Machine Placement (OVMP) algorithm that minimizes cost for hosting VM in multiple cloud provider environments, considering future demand and price uncertainties. The OVMP algorithm makes decisions based on the optimal solutions of Stochastic Integer Programming (SIP) and leases resources from cloud providers.

B. *Optimization Method Based on Heuristic Algorithm*

VM scheduling is often regarded as a type of classical packing problem (CPP) in computer science literature. The goal is to allocate a large number of candidate VM to a set of physical nodes. Various greedy heuristic algorithms are commonly employed to address VM scheduling problems. Notable approaches include the First Fit Decreasing (FFD) algorithm, the Best Fit Decreasing (BFD) algorithm, and the Worst Fit Decreasing (WFD) algorithm.

The sorting method of the FFD algorithm is to arrange the VM in descending order according to their demands, placing the larger VM at the beginning. The purpose of doing this is to prioritize placing VM that occupy more resources into physical servers, in order to better utilize the remaining resources to accommodate subsequent VM. Building upon this idea, Verma,Ahuja and Neogi [13] proposed a pMapper architecture that utilizes multiple methods to capture cost-aware application placement problems. These methods can be applied in various environments and adopt the FFD algorithm for VM placement, thereby reducing server energy consumption. Tang,Mo and Li [14] introduced a scheduling algorithm called Virtual Machine Dynamic Forecasting and Scheduling (VM-DFS), which aims to deploy VM in cloud computing environments. In this algorithm, by analyzing the historical memory consumption, the most suitable physical machine for placing VM can be predicted based on future consumption. This paper formulates the VM placement problem as a bin packing problem and solves it using the FFD approach. By adopting this method, for applications with specific VM requirements, the number of PM can be minimized to the greatest extent. The advantages of the FFD algorithm lie in its simplicity, ease of implementation, and ability to achieve good packing results in many cases. However, this algorithm may result in uneven utilization of server resources, and in certain problem instances, the FFD algorithm may not find the optimal solution.

The basic idea of the BFD algorithm is to first sort the VM to be allocated according to a specific rule, and then allocate them to the physical server with the smallest amount of remaining resources. By sorting the VM demands in descending order, placing the larger VM at the beginning, it is possible to better utilize the remaining resource space of the server with subsequent smaller VM. Based on this, Beloglazov,Abawajy and Buyya [15] proposed an energyaware BFD algorithm that provides data center resources to customer applications in a way that improves data center energy efficiency while ensuring Quality of Service (QoS). Abdullah,Lu and Wieder [16] proposed an improved BFD algorithm for intelligently allocating VM to hosts, as well as a Dynamic Utilization Rate (DUR) algorithm for space utilization and VM migration, reducing migration frequency and improving energy consumption and SLA violation rates. The advantages of the BFD algorithm include its ability to reduce waste of server's remaining resources to a certain extent and improve resource utilization efficiency. However, this algorithm may result in a more complex allocation process and may not find the optimal solution in certain situations. Therefore, in practical applications, the suitability of using the BFD algorithm needs to be balanced based on the specific requirements and constraints of the problem.

WFD is one of the heuristic algorithms commonly used to solve multidimensional boxing problems. The main goal of this algorithm is to select the physical server that occupies the most resources when allocating VM in order to minimize the remaining resources so as to maximize the resource utilization. Based on this idea, Yan,Wang and Li [17] designed and implemented C4, a low-cost consolidation service for scale computing in AliCloud.They analyzed the user model, resource utilization and migration cost of AliCloud and migrated the instances using the WFD algorithm to achieve a cost-effective, load-balanced, and non-oscillating consolidation service.The advantage of the WFD algorithm is that it is able to reduce the residual resources of the servers to a certain extent and improve the resource utilization efficiency. However, the algorithm may lead to an unbalanced distribution of server resources and may not be able to find an optimal solution in some cases. Therefore, in practical applications, the characteristics and constraints of the problem need to be considered comprehensively.

C. Optimization Methods Based on Meta-heuristics

Metaheuristic algorithms are a class of advanced heuristic algorithms that are typically designed based on certain behaviors or processes found in nature or biology, such as simulated annealing, genetic algorithms, particle swarm optimization, and so on. These algorithms imitate biological evolution, natural selection, collective behavior, and other mechanisms to search for optimal or near-optimal solutions within the search space. Metaheuristic algorithms exhibit the following characteristics:

- Randomness: Metaheuristic algorithms typically include random components to increase the diversity of the search and avoid getting stuck in local optima.

- Global Search Capability: Metaheuristic algorithms are dedicated to conducting global search in the solution space, rather than just local search. They often possess strong exploration capabilities, enabling them to escape from local optima and continue searching for other potential solutions.

- Robustness: Metaheuristic algorithms are insensitive to variations in problem input conditions and can be applied to different types of optimization problems. They exhibit a high degree of versatility and adaptability.

- Scalability: Metaheuristic algorithms have the ability to handle large-scale and complex problems. They can be adapted to different situations by adjusting parameters or introducing problem-specific constraints.

Currently, in the field of VM scheduling, widely adopted swarm intelligence algorithms include Genetic Algorithm (GA), Ant Colony Optimization (ACO), Particle Swarm Optimization (PSO), and Artificial Bee Colony Algorithm (ABC), among others.

GA is an optimization algorithm that simulates the evolutionary process in nature to solve problems. It optimizes and evolves individuals in a population generation by generation by simulating the genetic mechanism in biological evolution in order to solve optimal or near-optimal solutions. Based on this idea, Sarker and Tang [18] proposed a penalty-based genetic algorithm considering the migration cost from the current VM placement to the new optimal VM placement, and the gain of optimizing VM placement may be less than the loss of migration cost from the current VM placement to the new VM placement, the algorithm also considers the energy consumption of the new VM placement and the new VM placement in addition to the migration cost. the total interVM traffic in the new VM placement. In order to reduce the computation time and the number of VM migrations due to server consolidation, Sonklin,Tang and Tian [19] proposed a diminishing and conquering genetic algorithm, which employs a ”diminishing-conquering” strategy to reduce the problem size and the number of VM migrations without significantly affecting the quality of the solution.

ACO simulates the behavior of ants when searching for food and marking paths, through the information exchange and pheromone updating among ants, the better paths will be gradually selected and increase the pheromone concentration by the ants, while the bad paths will gradually decrease the pheromone concentration. In this way, the ants will be more inclined to choose the path with higher pheromone, thus gradually searching for the optimal or near-optimal solution of the problem.Alharbi,Tian and Tang [20] proposed an ACO for solving the VM placement problem, which effectively updates the pheromone to obtain the optimal solution of the problem through the introduction of a new energy-efficiencyaware heuristic factor.Liu et al [21] proposed an ant colony optimization-based approach to solve the virtual machine placement problem, which effectively utilizes the physical resources and reduces the number of running physical servers.

PSO simulates the cooperative behavior between individuals in a flock of birds or a school of fish, where particles move and update according to the individual best position and the global best position. Through the information exchange between particles and the guidance of the global best position, excellent solutions will be gradually gathered and searched by the particle swarm to find the optimal or near-optimal solution of the problem.Ramezani,Naderpour and Lu [22] proposed a fuzzy logic PSO algorithm to solve the multi-objective optimization problem of VM scheduling, which improves the efficiency of the traditional PSO through the use of a fuzzy logic system, reduces data center power consumption, improve resource utilization, and shorten VM migration time.Wang,Liu and Zheng [23] improved the parameters and operators of PSO, adopted an energy-aware local fitness prioritization strategy, and designed a novel coding scheme to effectively reduce the energy consumption of data centers.

ABC is an optimization method proposed based on the imitation of bee behaviors, which is a specific application of swarm intelligence. Its main characteristic is that it does not require understanding of specific information about the problem but relies on comparing the quality of solutions. Through the local search behavior of individual artificial bees, it ultimately brings forth the global optimal value within the colony, thus achieving fast convergence. Based on this, Jiang,Feng and Zhao [24] introduced a fast and energy-efficient real-time VM consolidation strategy named DataABC, which is based on a data-intensive energy model. DataABC adopts the concept of the artificial bee colony algorithm to achieve fast and globally optimized decisions for VM consolidation while reducing data center energy consumption without compromising the QoS.

V. . VIRTUAL MACHINE SCHEDULING METHODS

A. . Optimization Methods Based on Operations Research

With the increasing scale of cloud data centers, the energy consumption of devices is also increasing. Therefore, the optimization of energy consumption in VM scheduling has become a hot research issue in the field of cloud computing. By optimizing VM scheduling and controlling the energy consumption of data centers, we can significantly reduce energy costs and save enterprise operating expenses.

To optimize energy consumption in data centers, various optimization approaches based on VM scheduling have been proposed by many researchers. Hasan and Huh [25] introduced a VM selection and allocation method based on heuristic resource allocation. This approach aims to maximize the reduction of data center energy consumption and operational costs while ensuring user SLA requirements. Alboaneen,Pranggono and Tianfield [26] proposed a novel approach for classifying host loads within an energy-efficient

VM consolidation framework. By applying an underloaded detection algorithm, the hosts with underloaded loads are further categorized into three states: underloaded, normal, and critical. They also designed overload detection and virtual machine selection strategies to simultaneously reduce data center energy consumption while meeting QoS requirements.

B. . Load Balancing

In cloud data centers, user requests and tasks can enter the system in a constantly changing manner with different demands and priorities. The goal of load balancing is to ensure that each server or virtual machine is in a balanced state by dynamically adjusting resource allocation and avoiding situations where some resources are overloaded while others are idle. Load balancing can be achieved through various strategies and algorithms, such as round-robin, weighted load balancing, least connection algorithm, etc. These algorithms dynamically assign requests to relatively idle servers based on factors such as workload, server performance metrics, and network conditions to achieve load balancing. Lin and Wu [27] proposed a dynamic round-robin algorithm for energy-aware VM scheduling and consolidation to achieve load balancing in data centers. Panwar and Mallick [28] proposed a dynamic load management algorithm for effectively allocating all received requests among VM. This algorithm evenly distributes the workload among servers by effectively utilizing resources.

VI. CONCLUJION AND REMARKS

With the development of cloud computing, cloud data centers have gained widespread applications and have become the mainstream of computing services. By partitioning physical servers into multiple VM instances, cloud data centers achieve efficient resource utilization and flexible allocation, allowing users to quickly obtain computational power and storage space according to their needs. However, the current management of VM in cloud data centers still faces several challenges. Firstly, as the scale of cloud data centers continues to expand, the cost of management and energy consumption increases, necessitating more efficient resource utilization and energy management strategies. Secondly, with the diversification of cloud computing scenarios and the constantly changing business requirements, VM management systems need to possess better adaptability and intelligence to meet the demands of different scenarios and workloads.

This paper provides an overview of resource management in cloud data centers, including cloud data center overview, VM deployment, VM scheduling methods, and optimization objectives. In terms of virtual resource management, future research directions include: 1) designing high-performance and highly reliable cloud data center network topology to meet the needs of different application scenarios and workloads. Reasonable network topology design can improve the overall performance and resource utilization of the data center. Additionally, combining VM management with network topology for optimized management of network traffic, load balancing, and bandwidth allocation. Furthermore, addressing faults and security issues in data center network topology to ensure the reliability and confidentiality of user data is also an important direction for future research. 2) Incorporating machine learning and artificial intelligence technologies to combine VM management systems with automated scheduling, intelligent resource allocation, and other aspects to enhance system adaptability. Additionally, exploring research areas such as crossdata center VM migration, multi-tenant isolation, and security protection. 3) With the continuous expansion of cloud computing scale, the VM management system needs to meet the management needs of large-scale data centers, such as efficient migration algorithms, concurrency management strategies, and resource optimization technologies. Secondly, with the development of emerging technologies such as containerization and serverless architecture, how to combine them with traditional VM management to achieve diversified resource management and adapt to a variety of application scenarios is also a future research direction.

ACKNOWLEDGMENT

This work was supported in part by Special project for Key R&D and promotion of science,Technology Department of Henan Province(202102210327, 222102210272, 222102210052, 222102210007, 22210, 2210062, 212102210094, 232102211009, 222102210055), Postgraduate Education Reform and Quality Improvement Project of Henan Province under Grant (YJS2022JD26), Major Science and Technology Projects of Henan Province,and Major public welfare projects of Henan Province (201300210400), Higher EducationTeaching Reform Research and Practice Project(Graduate Education) of Henan Province, China(Grant No.2019SJGLX080Y), Postgraduate Education Reform and Quality Improvement Project of Henan Province, China (Grant No.YJS2022JD26), Graduate Education Innovation and Quality Improvement Project of Henan University, China (Grant No.SYLJD2022008,SYLKC2022028),2023 Graduate Education and Teaching Reform Research and Practice Project of Henan University, China (No.YJSJG2023XJ021), Henan Province Excellent Foreign Expert Introduction Project, China (Grant No.HNGD2023005), Key R&D and Promotion Special Project of Henan Province, China(Grant No.222102210055).

REFERENCES

[1] Y.M.Zhang, D.S.Li, X.C.Lu.Scalable distributed resource information service for internet-based virtual computing environment[J]. Journal of Software, 2007, 18(8): 1933-1942.

[2] J.Xin,K.Chen,W.M.Zheng.Studies on virtualization of cluster resource management technology[J].Journal of Frontiers of Computer Science and Technology,2010(4):324-329.

[3] E.Ahvar, S.Ahvar, Z.A.Mann, N.Crespi, J.Garcia-Alfaro, and R.Glitho, ”CACEV: A cost and carbon emission-effificient virtual machine placement method for green distributed clouds,” in Proc. IEEE Int. Conf. Services Comput. (SCC), San Francisco, CA, USA, Jun. 2016, pp. 275282.

[4] F.MACHIDA,D.S.KIM,J.S.PARK,et al.Toward optimal virtual machine placement and rejuvenation scheduling in a virtualized data center[C]//Proc of IEEE International Conference on Software Reliability Engineering Workshops.2008:1-3.

[5] F.MACHIDA,M.KAWATO,Y.MAENO.Redundant virtual machine placement for fault-tolerant consolidated server clusters[C]//Proc of Network Operations and Management Symposium.2010:32-39.

[6] S.Liu,G.Quan,S.P.Ren.On-line preemptive scheduling of real-time services with profit and penalty[C]//Proc of IEEE Southeast

[7] Conference.2011:287-292.

[8] V.M.Charitopoulos, L.G.Papageorgiou,V.Dua.Multi-parametric mixed integer linear programming under global uncertainty

[9] [J].Computers&Chemical Engineering,2018,116:279-295.

[10] Y.Chen,X.Chen,W.Liu,et al.Stochastic scheduling for variation-aware virtual machine placement in a cloud computing CPS[J].Future Generation Computer Systems,2020,105:779-788.

[11] J.Lopez,N.Kushik,D.Zeghlache.Virtual machine placement quality es- ´ timation in cloud infrastructures using integer linear programming[J]Software Quality Journal,2019,27(2):731-755.

[12] K.Zhang,T.Wu,S.Chen,et al.A new energy efficient VM scheduling algorithm for cloud computing based on dynamic programming[C]//2017 IEEE 4th International Conference on Cyber Security and Cloud Computing(CSCloud).Piscataway:IEEE,2017:249-254.