Digitalization and Management Innovation
A.J. Tallón-Ballesteros and P. Santana-Morales (Eds.)

doi:10.3233/FAIA230053

Virtualized Resources Scheduling in Multi-Tenant and Multi-Data Centers Based on Artificial Intelligence Algorithm: A Review

Cui Delong[a], Peng Zhiping[b,1], Qirui Li[a], Jieguang He[a],
Jiangtao Ou[c], Chengyuan Fan[c]
[a]*College of Electronic Information Engineer, Guangdong University of Petrochemical Technology, Maoming, China*
[b]*Jiangmen Polytechnic, China*
[c]*AI Sensing Technology, Chancheng District, Foshan, China*

Abstract. With the development and popularity of cloud computing, various artificial intelligence algorithms have been applied more and more widely in the field of cloud computing. While the complexity of artificial intelligence technology itself is also increasing, modern people are not satisfied with directly using artificial intelligence algorithm to solve complex practical problems, more and more researchers turn their eyes to explore the theoretical basis behind artificial intelligence algorithm, and try to explain artificial intelligence algorithm from various angles. This paper selects the latest application mode in the current field of cloud computing, and helps relevant researchers to clarify the development context of artificial intelligence technology in this typical mode by sorting out the latest artificial intelligence algorithm, so as to further explain the purpose of artificial intelligence technology.

Keywords. Artificial Intelligence; Virtualized Resources Scheduling; Multi-Tenant; Multi-Data Centers

1. Introduction

With the development and popularization of cloud computing, the complexity of cloud computing are also constantly improve. AI (Artificial Intelligence, AI) is the most commonly technology used to solve the problem of scheduling in the cloud computing environment. The present situation of AI not only technical ability put forward higher request, the interpretability of AI algorithm brings greater challenges.This paper selects the latest application modes in the field of cloud computing. Virtualized resources scheduling in Multi-Tenant and Multi-Data Centers using AI. The sope of this paper is helps relevant researchers to clarify the development context of AI technology in this typical pattern, and serves as a reference for further explanation of AI technology.

[1] Corresponding Author, Peng Zhiping, Jiangmen Polytechnic, China; Email: zhipingpeng@gdupt.edu.cn

Multi-tenant is one of the key features of cloud computing, which provides services for the users in an isolated manner by sharing the same cloud platform and its underlying infrastructure resources [1-4]. With the emergence, development, and growth of the novel virtualized container (Docker) technology in recent years, container virtualization technology and cloud platforms built on this technology are being widely adopted by the major cloud service providers due to its inherent advantages, including micro-servicing of applications, standardization of operation and maintenance, automation of integration/deployment, low testing and development costs. By June 2020, over 80% of internet enterprises (three times of that in 2018) are using container services in their production environments; 67% of the internet enterprises are using the hybrid cloud or the multiple public cloud services, with an average increase of 5% from 2019.

A common scenario for multi-tenant multi-data centers selection is shown in Figure 1, which has nine tenants and 10 data centers across four continents. Tenants select multi-data centers based on data center availability and preferences. Compared with one data center, using multiple data centers not only meets the requirements for comprehensive data analysis, but also ensures faster local data analysis and lower cost.

Although, the virtual cloud computing architecture with containers as the basic unit of operation offers new opportunities to address cost and efficiency issues in conventional virtual machine resource management, but it also poses following new challenges for resource management in container-based cloud platforms. There is an uncertain complex nonlinear dynamic relationship between computing resources and network resources, the two main virtual resources in the cloud platform for the large-scale multi-tenant and multi-data center, which is based on the docker as being the new virtual unit. These two resources are adaptively allocated in a collaborative mode to balance the maximum interests of both supply and demand of cloud services on the premise of meeting the SLA, (Service Level Agreements, SLA).

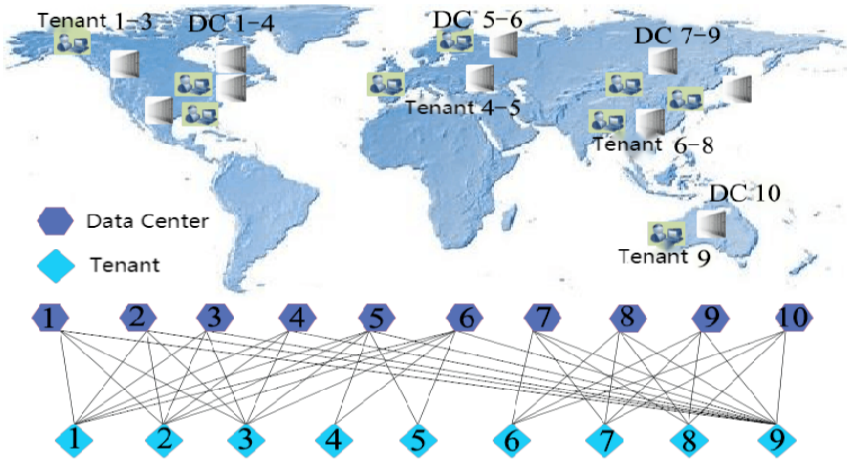

Figure 1. Data centers and tenants distributed across geographies

In order to satisfy the need of the multi-tenant services quality in this study,, including the selection of data centers, docker clusters and the server, and the adaptive scheduling mechanism of computing resources and networks resources in the docker cloud platform, the rest content is organization as follows. Firstly, the adaptive tasks scheduling theory in terms of the cloud computing and networks resources in the multi-

data centers with the virtual docker technology,; secondly, the co-adaptive scheduling algorithm for virtual network resources in the docker cloud platform for the multi-tenant and multi-data centers, including the high-performance co-adaptive scheduling mechanism and the knowledge transfer mechanism; thirdly, the shortcomings of the existing work are summarized and the possible research directions in the future are pointed out.

1.1. Challenges to computering resource management in container-based cloud computing

- Load balancing and elastic provisioning: The value of cloud computing lies in the construction of the user needs. In actual application environments, a considerable number of real-time online processing services often exist in addition to a large number of asynchronous processing services. Real-time online services have a short processing time with fluctuations in demand, while asynchronous services have a long processing time with huge data volume. The existing container resource management mechanism is incapable of adjusting the load of each server adaptively, resulting in poor resource utilization.
- Synergistic configuration of parameters: The container technology at the same time is also evolving rapidly. At present, the subcommands in a container have been up to 34, where core subcommands involve complex parameter configurations. For instance, the run command can be configured with up to 28 parameters. In terms of functionality and application scenarios, the commands contained in a container can be classified into several types, such as environmental information, system operation and maintenance, log information, and Hub service. The lack of coordination between front- and back-end commands and parameters can lead to the increased complexity of fault tolerance and significant degradation of system performance.
- Difficulty in measuring unexpected traffic and redundant resources: New tenants are not able to accurately predict the volume of the user requests. In the event of large fluctuations in the user needs, the conventional processing model for such unplanned requests is to quickly conduct scale-up and validation to go live, but this not only is time-consuming but also requires preparation to cope up the unexpected traffic, which is prone to cause significant resource redundancy.

1.2. Challenges to network resource management in a container-based cloud computing

- Large scale and high failure rate: With the number of servers interconnected in a public cloud data center exceeding the magnitude of 10^5 and the number of exchange nodes reaching the magnitude of 104, the increasingly large scale of the data center places new demands on the network architecture, transport protocols, and system management. Additionally, the network failure rate increases rapidly with the size of the system, and of the failures, the failure of network configuration (38%) and unexplained failures (e.g., switches suddenly stop forwarding traffic, accounting for 23%) show the most significant increase.
- Traffic complexity and high vertical scaling costs: Due to the Incast problem incurred by highly bursty and dynamic many-to-one communications, the growth of computing-intensive applications such as MapReduce and Hadoop, and the widespread use of virtualization technologies, the traffic behavior of networks gets more complex,

and severe transport loads are brought about. Meanwhile, as a result of the high occupancy of "east-west traffic" within a data center and the convergence rate issue in tree structures, the vertical scaling in a data center becomes extremely expensive and unsustainable.

- Low resource utilization and diverse forms: Traditional data exchange (e.g., Vlan) and communication identity (e.g., IP) technologies are effective in avoiding the interference between multiple applications deployed simultaneously in a data center, but they also limit the flexibility of the network resource reuse, resulting generally in low utilization of network resources. Additionally, driven by different performance requirements, networks in diverse forms co-exist, including the enhanced Ethernet, high-speed InfiniBand interconnect storage networks[5], and dedicated high-speed networks.

Although how data center resources are managed and leased by both the supply and demand sides of cloud services is gradually changing with the development of new technologies such as containers, resource utilization and SLA are still two fundamental concerns for both sides of cloud services. Meanwhile, cloud resources are still managed and used on a pay-as-you-go basis. Therefore, under the premise that SLA is ensured, cloud tenants are more concerned about how to reduce the occupancy of data center resources in terms of the lease to lower payment, while cloud service providers focus more on how to improve the resource utilization in terms of resource portfolio to reduce the operational costs. However, a cloud computing environment is an open and heterogeneous environment where loads, infrastructure, containers, and application deployments are varying rapidly. Here, we take a web application as an example. In this case, a single unexpected event can cause a spike in site traffic, which is often unpredictable. In cloud computing, a massive distributed system consisting of thousands of cheap PC servers, hardware failures are inevitable, and it is common for a physical node to be dynamically added or removed. The dynamic uncertainty is even more pronounced in a multi-tenant, multi-data center cloud service environment. Therefore, a multi-tenant, multi-data center cloud service environment objectively requires that the allocation of resources in a cloud platform, especially the two primary resources of computing and network resources, achieve adaptive scheduling with the dynamic change in the environment.

However, in a rapidly varying multi-tenant, multi-data center cloud environment, the adaptive scheduling of computing and network resources is still very difficult to achieve, even with the novel container virtualization technologies. For instance, Amazon has at least 11 data centers across four continents [6], each with hundreds of thousands of servers, and Google has at least 13 data centers across four continents [7], each with more than one million servers. After the physical resources in the data centers are virtualized, the number of servers is even larger, making it technically challenging and complex to manage. Also, connecting the data centers with private networks is costly. On the other hand, there is a complex non-linear and uncertain relationship between computing resources and network resources. It means that adjusting one aspect alone may not improve the resource utilization and application service performance, and dynamic and coordinated configuration is required [8, 9]. Therefore, under the rapidly changing multi-tenant, multi-data center cloud environment, it is objectively required not only to achieve the adaptive scheduling of computing resources and network resources but also to perform adaptive scheduling in a synergistic method. Nonetheless, little research has been conducted in this field.

In summary, this study is of great theoretical and practical significance to improve the resource utilization and achieve a balance of interests between both the supply and demand sides of cloud services in a multi-tenant, multi-data center cloud environment built on the novel container virtualization technology, under the premise that the tenant SLA is guaranteed.

2. Literature Review and Development Dynamics Analysis

The essence of the cloud computing is to provide pay-as-you-go services for the users by virtualizing the resources in the data centers. Therefore, the data centers play a key role in the era of cloud computing. Among various resources managed by the data centers, computing and management of network resources are the dominant ones. The research methods used in the studies can broadly be classified into two categories namely, static scheduling methods and dynamic scheduling methods.

2.1. Current Research on Computing Resource Management under a Cloud Environment

2.1.1 Current Research on Computing Resource Management in a Data Center

Static scheduling method: The current research on static scheduling of computing resources in data centers mainly focuses on the placement, migration, and synergistic and adaptive configuration of virtualized resources with changes in the user needs and application system parameters, in the hope of improving computing resource utilization while reducing the resource fragmentation [10]. To measure the overall data transfer rate of a tenant, Li *et al*. introduce a new concept, that is the metric progress, which is defined as the minimum demand-normalized bandwidth allocation on all network links. The progress is an essential metric to indicate how fast a tenant can complete data transfer. By maximizing the tenant's progress,application performance such as execution time can be optimized[11].Yan *et al.* [12] advocated a data center selection algorithm based on steady state probability (SPP). This algorithm greatly reduces the probability of cloud services being blocked as it predicts the probability of being selected based on the state of the data center.

Dynamic scheduling method: Li *et al*. [13] aiming at the problem of the existing runtime prediction methods for workflow activities fail to effectively extract categorical and numerical features, propose a multi-dimensional feature fusion-based runtime prediction approach for workflow tasks. Guo *et al.* [14] put forward a “shadow router” based model for real-time adaptive virtual machine provisioning in large-scale data centers which allows for setting diverse targets and/or constraints and automatically adapts to changes in virtual need rates and system parameters. Based on this model, a combined selection algorithm of virtual machine-router and virtual machine-physical machine was designed to solve the min-max-DC-load issue. In face of the problem that the virtual machines of conventional cloud service providers are unable to build a completed virtual data center, Shi *et al.* [15] decomposed the problem into three sub-problems namely, virtual data center clustering, virtual machine allocation and virtual chain allocation, and used a virtualized data center allocation algorithm to find the optimal solution to the problem.

Limitations of computing resource management in a data center are as follows:

● The computing resource scheduling by most of the data centers is usually based on the assumption that the state of data centers and network connections are fully or partially known and that data centers have sufficient servers which can be switched at will. However, the state of data center resources is rapidly changing and becomes highly unpredictable in the real cloud computing environments. Besides, the situation when data centers are overloaded is not considered.

● Although, the dynamic scheduling algorithms of most data centers can accept the variable amount and variable types of resource constraints, the computational feasibility of the algorithms lacks verification.

2.1.2 Current Research on Server Computing Resource Management

Currently, the virtual machine placement policy under a cloud environment is a hot research topic in cloud computing and has received extensive attention from the experts and scholars at home and abroad. Most studies focus on a single or comprehensive performance indicators such as lowering user payment, reducing job execution time, improving resource utilization, decreasing system energy consumption through resource allocation prediction, load balancing, and other measures [16, 17].

Static scheduling method: Existing studies on server resource management mainly focused on the server consolidation problem, and most of the studies modeled this problem as a bin-packing problem. Built on the bin-packing idea, Song *et al.* [18] introduced a data center virtualization resource allocation algorithm based on the user application needs. This algorithm effectively improved the resource availability both intra- and inter-physical machines and increased the competition ratio of the algorithm to 3/2. To further reduce the number of bins in the bin-packing problem, Song *et al.* [19] suggested an improved algorithm known as Harmonic Mix, which supposes that at the most 10 items can be moved per run. This algorithm not only reduced the maximum size of the items from 1/6 to 1/8 but also increased the competition ratio to 4/3. Cao *et al.* [20] recommended an on-demand resource allocation algorithm based on grey waveform prediction. This algorithm used grey waveforms to predict the load of virtual machines in the next resource allocation cycle and designed a utility function for virtual machine services that accounted both the resource needs and the service priority to maximize the overall service utility value of each virtual machine in the physical machine. Wang *et al.* [21] introduced a predictive resource management method to monitor and allocate the memory resources, which not only enabled the load balancing of virtual machines among multiple physical hosts but also effectively improved the memory resource utilization in the data center.

Dynamic scheduling method: Sun *et al.* [22] came up with a dynamic resource scheduling algorithm based on the virtual machine auction. Based on the two-way combinatorial auction protocol, and combined with neural network and swarm search optimization techniques, this algorithm achieved efficient resource allocation under cloud environments. For the problem of combinatorial auction for heterogeneous virtual machines, Zhang *et al.* [23] introduced a dynamic virtual machine resource scheduling model based on this technique. On the other hand, Xu *et al.* [24] propounded a joint placement algorithm for virtual machines based on the loading peak features. This algorithm computes the loading peak features by using a loading peak similarity formula to model the load of virtual machines and then achieves resource complementation by combining the virtual machines with the load peaks occurring at different points of time thus, improving the resource utilization. Based on the K-Means

clustering algorithm, Zhang *et al.* [25] divided the workflows with similar resource needs and performance needs into the same set of tasks for unified debugging and dynamically adjusted the number of virtual machines with the principle of minimizing energy consumption, which improved the throughput of the platform. To address the problem of severe blocking of service requests that are about to expire, Le *et al.* [26] divided resource management into two phases that is, resource provisioning and resource scheduling. In the scheduling phase, a mixture of several mechanisms, including the first-in-first-out (FIFO) algorithm, short task priority, and urgent task priority, were used to schedule tasks according to the different task requirements to ensure the timely completion of the tasks. Pawar *et al.* [27] designed a priority-based computing resource management system. The SLA parameters involved include timely task completion indicators, as well as CPU processing time, memory size, network bandwidth, and service priority. In this way, the execution time of the tasks can be reduced.

Limitations of the server computing resource management are as follows:

- Most studies only consider the single allocation of CPU or storage resources during resource allocation and not on how to integrate virtual machines by taking advantage of the difference in needs between the virtual machines and the resources.
- Although, most studies can effectively improve resource utilization but affect the implementation of the SLA and fail to achieve the synergy between the two aspects.

2.1.3 Resource Selection with Containers as the Scheduling Unit

Currently, virtualized container technology is the most prevalent way of resource provisioning in cloud computing. Unlike the conventional virtual machine technology, the novel technology does not require a full OS instance in the containers, which can greatly reduce the consumption of the server CPU, memory, and other resources [28]. Li *et al.* [29] used global and local resource managers to establish an elastic resource management framework for data centers with containers as the basic unit of virtualized resource management. To cope with the mixed deployment of the tasks with different priorities and resource needs on nodes, they designed a QoS-guaranteed task-resource matching algorithm. The experimental results showed that the algorithm not only reduced the occurrence of the node resource shortage significantly, but also improved the application performance under the same resource needs. Xu *et al.* [30] investigated a cloud computing resource scheduling algorithm using containers instead of virtual machines. The results demonstrated that using the containers as the basic scheduling unit could significantly reduce the response time of the cloud users and improve the utilization of the cloud provider resources.

2.1.4 The Study Conducted by the Research Group on Dynamic Resource Provisioning

The research group proposed a reinforcement learning-based resource scheduling algorithm [31] by abstracting resource scheduling under cloud environments into a coherent decision problem and designed a novel payoff function by introducing two performance indicators, which are segmented SLA and utilization of costs per unit time. For the virtualization placement problem, a multi-objective comprehensive evaluation model for virtual machines was designed [32], and a multi-objective particle swarm optimization algorithm was proposed for the dynamic placement of virtual machine resources.

Limitations are as follows:

● Most of the studies only consider the relationship between resource provisioning methods and resource types as well as between service types but fail to take the relationship between dynamic resource provisioning methods and data distribution strategies into account.

● Although, majority of the studies focus on how to improve resource utilization, but not enough research is done on load balancing.

2.2. Current Research on Network Resource Management under a Cloud Environment

2.2.1 Network Resource Management in a Cloud Platform

Off late, in pursuit of providing services to the users using multi-clouds and hybrid clouds [33], cloud brokers are recommended as a basic cloud service model [34, 35] to minimize the costs and maximize the profits by renting cloud service provider instances for cloud network selection and reusing relatively small tenant needs. Figure 2 shows the tree-like topology structure of data center represented by Fat-tree, which consists of access layer, aggregation layer and core layer. The number of communication paths across container clusters is determined by the number of switches at the core layer, while the number of communication paths within container clusters is determined by the number of switches at the aggregation layer in the cluster.

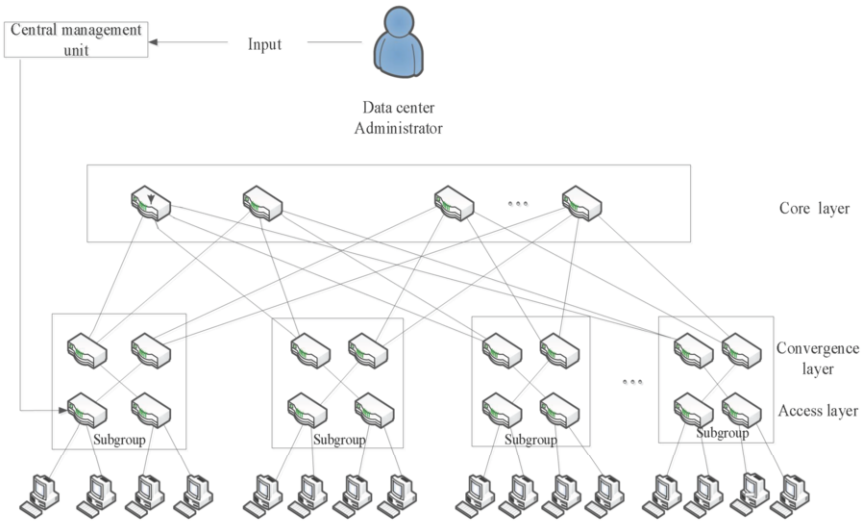

Figure 2. A typical fat-tree topology

Static scheduling method: Choi *et al.* [36] proposed a graph clustering-based cloud network selection algorithm to minimize the costs of the cloud brokers. In this algorithm, the clustered objects include data centers (nodes), inter-cloud networks, and intra-cloud networks. Truong *et al.* [37] maximized the profits for multi-cloud service providers based on a synergistic mechanism of cloud brokers. Apart from the costs and profits, service quality is the most primary concern of cloud service providers. In order to meet the needs of the users in terms of service quality, Amato *et al.* [38] scored various service capabilities of cloud service providers based on a multi-objective strategy and selected cloud service providers for cloud brokers based on their scores.

To achieve two performance indicators, that is, response time minimization and profit maximization, Kessaci*et al.* [39] formalized the cloud broker scheduling problem as a multi-objective planning problem using Pareto optimality theory and used a multi-objective genetic algorithm to search the optimal scheduling policy. To deploy latency-sensitive applications across multiple cloud service providers, Diaz *et al.* [40] designed a mixed integer programming algorithm subject to the constraints of resource capacity, load balancing, and latency, and devised two strategies to cope with the failure of the cloud service providers. Lin *et al*. [41] proposed a two-stage job scheduling and resource allocation framework that adopts multiple intelligent schedulers to solve the cooperative scheduling problem between job scheduling and resource allocation. A heterogeneous distributed deep learning model is used in the job scheduling stage to schedule multiple jobs to multiple cloud data centers.

Dynamic scheduling method: Dynamic resource reservation [42, 43] provides a new way for cloud brokers to reduce costs. Cloud brokers can adopt dynamic resource reservation strategies to lower the costs [44], with further explorations shown in the study [45]. Wan *et al.* [46] designed a balanced evaluation system for nodes, paths, and flows based on the computing capacity of centralized and scalable controllers, which effectively improved the operational efficiency of data (especially large volumes of data) during the forwarding process.

Limitations are as follows:

- Since most of the studies do not consider the price difference between reserved resources and real-time resources, they are unable to reflect the dynamic features of the tenant needs in a timely manner.
- Majority of the studies tend to assume that cloud service providers have only one reserved instance cycle, which is not in line with the actual business operation situation in cloud computing.
- Further, these studies tend to choose among multiple cloud platforms owned by the same cloud service provider, making it difficult to implement across multiple cloud service providers.

2.2.2 Network Resource Management in a Data Center

The performance of a cloud data center is essentially determined by the performance of the network connecting all the servers inside the data center. Therefore, network bandwidth allocation in data centers has been the hot topic of research in cloud computing.

Static scheduling method: In static allocation method the data center provides bandwidth to the leasers in the data center in a static resource reservation way [47]. Alicherry *et al.* [48] investigated the selection of network resources in data centers. By modeling the system using fully vertex-weighted graphs and proving the problem to be NP-hard, they designed the Find-Min-Star algorithm to discover the data center clusters and finally select the data center cluster with the smallest diameter as the optimized solution to the problem. Currently, key results in this field have been applied to major cloud service providers. For example, by means of dedicated switches, Google's B4 network [49] achieves over 95% network link utilization; The SWAN controller [50] designed by Microsoft exhibits high link utilization and can automatically solve the congestion update problem.

Dynamic scheduling method: With an emphasis on the time-dependent feature of data volume, Zhang *et al.* [51] chose a single data center to store user data and designed

an online algorithm to optimize the costs of bandwidth, storage, and data movement. Femminella *et al.* [52] proposed a similar scheduling algorithm for genetic big data. Guo *et al.* [53] designed an OFPF-based SOTE algorithm for hybrid node networks, which reduced local congestion paths by dynamic weight assignment. In smart synergistic network architecture, Miao *et al.* [54] designed a multi-parameter multipath routing algorithm for smart synergistic networks, developed a multi-parameter multipath routing protocol for smart synergistic networks and weighted the network performance parameters, such as CPU occupancy, round-trip delay and bandwidth to obtain the path weights.

Limitations are as follows:

- These studies assume that the capacity of a data center is measured by the number of virtual machines it can accommodate. As a result, only the virtual machine homogeneity issues can be solved.
- The studies assume that there always exists one data center that can be used to store all the user data, which often is not true.

2.2.3 Network Resource Management in a Server

Server selection studies mainly focused on the server consolidation problem [55, 56]. Especially in computing resource management, most of the studies also model this problem as a bin-packing problem.

Static scheduling method: Wang *et al.* [57] introduced the bandwidth-aware server consolidation problem and provided an approximation algorithm with an approximation ratio of $(1+\varepsilon)(\sqrt{2}+1)$, which was increased to 2 in the subsequent study [58]. Meng *et al.* [59] thoroughly investigated the components of the bandwidth costs and used a graph partitioning method for optimal scheduling.

Dynamic scheduling method: There are relatively few studies conducted in this area. Yue *al.* [60] proposed a software-defined hybrid routing algorithm for data center networks with tree structures to address the problems of uneven traffic distribution and different transmission performance needs of the data center networks. Through statistical analysis, the algorithm divides the data flow into two categories, that is, large flow and small flow. To meet their different transmission performance needs, adaptive routing algorithms are used for large flows while non-traffic-aware routing algorithms are used for small flows[61].

Limitations are as follows:

- These studies model the server network resource management problem as a bin-packing problem, but the bin-packing problem allows items to overlap in the same resource dimension, while the resources in a server generally can be used by only one virtual machine at a given time.
- Since the cost optimization algorithms designed in these studies are not integrated with the server networking topology, the networking information cannot be utilized fully to find a more optimal solution.

2.2.4 Work of the Research Group on the Data Center Networks

To ensure the service quality of security situation information under the constraints of delay limitations and network resources [62], the research group designed a solution to schedule security situation information according to the urgency of data packet delay

[63], to control the data flow into the system according to the capacity of the link, and to design a fast queueing method for data packets according to the queue leader [64]. For the data packet timeout and energy consumption problems caused by the dynamic nature of the wireless environment, the research group demonstrated the feasibility of actively dropping time-out data packets from both theoretical analysis and simulation validation [65].

2.3. Current Research on Synergistic Management of Computing Resources and Network Resources in a Cloud Environment

Considering both data center and server layer resources, Yao *et al.* [66] designed a two-time-scale Lyapunov optimization algorithm for the selection of data centers and servers to reduce energy consumption. This algorithm was originally intended to be applied to latency-insensitive tasks like MapReduce but did not involve specific Map and Reduce phases. Wang *et al.* [67] studied MapReduce across data centers. They used a stochastic mechanism to deploy Reducer in a hierarchical architecture. Zhang *et al.* [68] studied the scheduling strategy for remote sensing data across the data centers. They reduced data transmission between data centers by applying a combination of hypergraphs and task trees and selected critical workflow paths to optimize the task completion time.

3. Conclusion and Prospect

3.1 Conclusion

This paper selects the latest application mode in the current field of cloud computing, and helps relevant researchers to clarify the development context of artificial intelligence technology in this typical mode by sorting out the latest artificial intelligence algorithm, so as to further explain the purpose of artificial intelligence technology.

3.2 Prospect

At present, most of the relevant studies in China and abroad focused on the unilateral adaptive scheduling of computing resources or network resources in data centers, ignoring the inherent nonlinear dynamic uncertainty relationship between the two, which makes it difficult to achieve a balance of interests between both the supply and the demand sides of the cloud services while guaranteeing SLA. Besides, studies on synergistic adaptive scheduling between the two are very few, and the depth of the research and effective solutions are still lacking. In particular, there is scant research on synergistic adaptive scheduling in large-scale multi-tenant and multi-data center cloud platforms using the novel container virtualization technology. The adaptive scheduling of computing resources and network resources in a synergistic manner will be one of the core problems to be solved in the deployment of multi-tenant and multi-data center services with containers as the core virtualization technology.

Acknowledgments

The work presented in this paper was supported by: National Natural Science Foundation of China (62273109, 61772145, 61672174); Key Realm R&D Program of Guangdong Province(2021B0707010003). Guangdong Basic and Applied Basic Research Foundation (2021A1515012252, 2020A1515010727, 2022A1515012022). Key Field Special Project of Department of Education of Guangdong Province (2020ZDZX3053). Key Realm R&D Program of Guangdong Province(2021B0707010003).Guangdong Province Ordinary Universities Characteristic Innovation Project (2019KTSCX108) .Maoming Science and Technology Project (210429094551175, mmkj2020008, mmkj2020033).

Reference

[1]. Bhaskar Prasad Rimal, Martin Maier. Workflow scheduling in multi-tenant cloud computing environments [J]. IEEE Transactions on Parallel and Distributed Systems, 2017, 1(28): 290-304.
[2]. Delong Cui, Zhiping Peng, Jianbin Xiong, et al. A Reinforcement Learning-Based Mixed Job Scheduler Scheme for Grid or IaaS Cloud[J]. IEEE Transactions on Cloud Computing, 2020, 8(4):1030-1039.
[3]. Zhiping Peng, Jianpeng Lin, Delong Cui, et al. A multi-objective trade-off framework for cloud resource scheduling based on the Deep Q-network algorithm[J]. Cluster Computing, 2020, (23):2753-2767.
[4]. Kaibin Li, Zhiping Peng, Delong Cui, et al. SLA-DQTS SLA Constrained Adaptive Online Task Scheduling Based on DDQN in Cloud Computing[J]. Applied Sciences, 2021, 11(20):9360-9360.
[5]. http://aws.amazon.com/about-aws/global-infrastructure/?nc2=h_ls2
[6]. http://www.google.com/about/datacenters/inside/locations/index.html
[7]. Jia Rao, Xiangping Bu , ChengZhong Xu. A distributed self-learning approach for elastic provisioning of virtualized cloud resources[C]. Proceedings of the 2011 IEEE 19th Annual International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, Washington , DC, USA, 2011: 45-54..
[8]. Jia Rao, Xiangping Bu, Kun Wang, ChengZhong Xu. Self-adaptive provisioning of virtualized resources in cloud computing[C]. Proceedings of ACM Special Interest Group on Measurement and Evaluation, New York, USA, 2011: 129-130.
[9]. Shuyi Yan, Xue Wang, Miguel Razo, Marco Tacca, Andrea Fumagalli. Data center selection: A probability based approach[C]. Proceedings of International Conference on Transparent Optical Networks, Graz, Austria , 2014: 1-5.
[10]. Wanjun Huang, Marco Tacca, Ning So. Cooperative data center selection for optimal service performance: An ILP formulation[C]. Proceedings of 10th IEEE International Symposium on Parallel and Distributed Processing with Applications, Madrid, Spain, 2012: 523-526.
[11]. Yan Li, DeKe Guo, Xiaofeng Cao, Honghui Chen. Application-aware network sharing for multi-tenant data center. Chinese Journal of computers, 2021,44(07),1363-1377
[12]. Zhen Xiao, Weijia Song, Qi Chen. Dynamic resource a llocation using virtual machines for cloud computing environment [J]. IEEE Transactions on Parallel and Distributed Systems, 2013, 24(6): 1107-1117.
[13]. Yang Guo, Alexander Stolyar. hadow-routing based dynamic algorithms for virtual machine placement in a network cloud [J]. IEEE Transactions on Cloud Computing, 2015, (99): 1-10.
[14]. Li Shi , Katramatos, Dimitrios, Dantong Yu. Virtual data center allocation with dynamic clustering in clouds[C]. Proceedings of 2014 IEEE 33rd International Performance Computing and Communications Conference, Austin, U SA, 2014: 1-10.
[15]. Sukhpal Singh, Inderveer Chana. Cloud based development issues: A methodical analysis[J]. International Journal of Cloud Computing and Services Science, 2013, 2(1): 73-84.
[16]. Sukhpal Singh, Inderveer Chana. Cloud resource provisioning: survey, status and future research directions[J]. Knowledge and Information Systems, 2016, 49(3): 1005-1069.
[17]. Wangzhang Cao, Bin Li, Bing Qi, Yi Sun, Songsong Chen , Kun Shi. Two Stage Robust Optimization Model of Colocation Data Centers Cluster Under Computing Resource Sharing Mode. Power System Technology, 2022, 46(10): 4102-4115.

[18]. Kamali, Shahin Efficient b in packing algorithms for resource provisioning in the cloud[C]. Proceedings of International Workshop on Algorithmic Aspects of Cloud Computing, Patras, Greece, 2016: 84-98.
[19]. Cao Jie, Zeng Guosun, Kuang Guijuan, Zhang Jianwei, Ma Haiying, Hu Kekun. Cloud virtual machine on-demand physical resource allocation method that supports random service requests [J].Software Journal, 2017,28 (2): 457-472.
[20]. Wang Zhigang, Wang Xiaolin, Jin Xinxin, Wang Zhenlin, Luo Yingwei.Mbalancer: Virtual machine memory resource dynamic forecasting and provisioning[J]. Software Journal, 2014, 25(10): 2206-2219.
[21]. Sun Jiajia, Wang Xingwei, Gao Cheng Xi, Huang Min. Resource allocation mechanism optimized based on neural network and group search in the cloud environment[J]. Software Journal, 2014,25 (8): 1858-1873.
[22]. Linquan Zhang, Zongpeng Li, Chuan Wu. Dynamic resource provisioning in cloud computing: A randomized auction approach[C]. Proceedings of IEEE 33rd Conference on Computer Communications, Toronto, Canada , 2014: 433-441.
[23]. Xu Siyao, Lin Weiwei, Wang Zijun.Virtual Machine Placement Algorithm based on load peak feature [J]. Software Journal, 2016,27 (7): 1876-1887.
[24]. Qi Zhang, Mohamed Faten Zhani, Raouf Boutaba. HARMONY: dynamic heterogeneityaware resource provisioning in the Cloud[C]. Proceedings of IEEE 33rd international conference on distributed computing systems, Philadelphia, USA, 2013: 510-519.
[25]. Guan Le, Ke Xu, Junde Song. Dynamic resource provisioning and scheduling with deadline constraint in elastic Cloud[C]. Proceedings of International Conference on Service Science , Shenzhen, China, 2013: 113-117.
[26]. AF Dhaya, R., Kanthavel, R., IoE based private multi-data center cloud architecture framework. Computers & Electrical Engineering, 2022, 100: 107933
[27]. Krishan Kumar, Manish Kurhekar. Economically efficient virtualization over cloud using Docker containers[C]. Proceedings of IEEE International Conference on Cloud Computing in Emerging Markets, Bangalore, India, 2016: 95-100.
[28]. Li Qing, Li Yong, Tu Bi, Bo Mengdan. QoS guarantee[J]. Journal of Computer Science, 2014,37 (12): 2396-2407.
[29]. Xin Xu, Huiqun Yu, Xin Pei. A novel resource scheduling approach in container based clouds[C]. Proceedings of IEEE 17th International Conference on Computational Science and Engineering, Chengdu, China, 2014: 257-264.
[30]. Delong Cui, Zhiping Peng, Qirui Li, et al. A Cloud Workflow Scheduling Algorithm for Multi-object Optimization Using Reinforcement Learning[J]. Journal of Nonlinear and Convex Analysis, 2020, 8(4):1677-1687.
[31]. Zhiping Peng, Jianpeng Lin, Delong Cui, et al. A multi-objective trade-off framework for cloud resource scheduling based on the Deep Q-network algorithm[J]. Cluster Computing, 2020, (23): 2753-2767.
[32]. Rodrigo N. Calheiros, Adel Nadjaran Toosi, Christian Vecchiola. A coordinator for scaling elastic applications across multiple clouds[J]. Future Generation Computer Systems, 2012, 28(8): 1350-1362.
[33]. Alba Amato, Beniamino Di Martino, Salvatore Venticinque. Cloud brokering as a service [C]. Proceedings of 8th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, Compiegne, France, 2013: 9-16.
[34]. Mechtri Marouen , Zeghlache Djamal, Zekri Elyes, Marshall IainJames. Inter and Intra cloud networking gateway as a service[C]. Proceedings of IEEE 2nd International Conference on Cloud Networking, San Francisco, USA, 2013: 156-163.
[35]. Lizi Zheng, Delong Cui. Collaborative adaptive scheduling scheme for multi-source big data tasks in the cloud[C]. 2019 International Conference on Image and Video Processing, 2020, (2020):1-6.
[36]. Truonghuu Tran, Chen khong Tham.A novel model for competition and cooperation among cloud providers [J]. IEEE Transactions on Cloud Computing, 2014(3): 251-265.
[37]. Alba Amato, Salvatore Venticinque. Multi-objective decision support for brokering of cloud SLA[C]. Proceedings of 27th International Conference on Advanced Information Networking and Applications Workshops , Barcelona, Spain, 2013: 1241-1246.
[38]. Yacine Kessaci, Nouredine Melab, Elghazali Talbi. A pareto-based genetic algorithm for optimized assignment of VM requests on a cloud brokering environment [C]. Proceedings of IEEE Congress on Evolutionary Computation, Cancun, Mexico, 2013: 2496-2503.
[39]. Díazsánchez Felipe, Al Zahr Sawsan. An exact placement approach for optimizing cost and recovery time under faulty multi-cloud environments[C]. Proceedings of 5th International Conference on Cloud Computing Technology and Science, Bristol, UK, 2013, 2: 138-143.

[40]. Alexei A. Gaivoronski, Alexei A. Gaivoronski, Per J. Nesse, Xiaomeng Su. Modeling and economic analysis of the cloud brokering platform under uncertainty: Choosing a risk/profit trade-off[J]. Service Science, 2013, 5(2): 137-162.
[41]. Jianpeng Lin, Delong Cui, Zhiping Peng, et al. A Two-Stage Framework for the Multi-User Multi-Data Center Job Scheduling and Resource Allocation[J]. IEEE Access, 2020(8): 197863-197874.
[42]. Qirui Li, Zhiping Peng, Denglong Cui, et al. Data Center Selection Based on Reinforcement Learning[C]. 2019 4th International Conference on Cloud Computing and Internet of Things (CCIOT), 2019: 14-19.
[43]. Nirnay Ghosh, Soumya K. Ghosh, Sajal K. Das. SelCSP: A framework to facilitate selection of cloud service providers[J]. IEEE Transactions on Cloud Computing, 2015, 3(1): 66-79.
[44]. Wei Wang, Di Niu, Baochun Li. Dynamic cloud resource reservation via cloud brokerage [C]. Proceedings of IEEE 33rd International Conference on Distributed Computing Systems, Philadelphia, USA, 2013: 400-409.
[45]. Wei Wang, Di Niu, Baochun Li. Dynamic Cloud resource reservation via IaaS cloud brokerage[J]. IEEE Transactions on Parallel and Distributed Systems, 2014, 7973(6): 1580-1593.
[46]. Kaiyang Liu, Jun Peng, Weirong Liu, Pingping Yao. Dynamic resource reservation via broker federationin cloud service: A fine-grained heuristic-based approach[C]. Proceedings of IEEE Global Communications Conference, Austin, USA, 2014: 2338-2343.
[47]. Jihai Zhong, Delong Cui, Zhiping Peng, et al. Multi Workflow Fair Scheduling Scheme Research Based on Reinforcement Learning[C]. 9th Annual International Conference of Information and Communication Technology, 2019, (154): 117-123.
[48]. Li Chen, Baochun Li, Bo Li. Allocating Bandwidth in Datacenter Networks: A Survey [J]. Journal of Computer Science and Technology, 2014, 29(5): 910-917.
[49]. Mansoor Alicherry, T.V. Lakshman. Network aware resource allocation in distributed clouds [C]. Proceedings of IEEE Conference on Computer Communications, Orlando, USA, 2012: 963-971.
[50]. Jain Sushant, Kumar Alok , Mandal Subhasree , Poutievski Leon .B4: Experience with a globally-deployed software defined WAN[C].Proceedings of Annual Conference of the ACM Special Interest Group on Data Communication , Hong Kong, China , 2013, 43(4): 3-14.
[51]. Chiyao Hong, Srikanth Kandula, Ratul Mahajan, Ming Zhang. Achieving high utilization with software-driven WAN[C]. Proceedings of Annual Conference of the ACM Special Interest Group on Data Communication , Hong Kong, China, 2013, 43(4): 15-26.
[52]. Linquan Zhang, Chuan Wu, Zongpeng Li. Moving big data to the cloud: An online cost-minimizing approach[J]. IEEE Journal on Selected Areas in Communications, 2013, 31(12): 2710-2721.
[53]. Femminella Mauro, Reali Gianluca, Valocchi Dario, Nunzi Emilia. The ARES project: network architecture for delivering and processing genomics data[C]. Proceedings of IEEE Symposium on Network Cloud Computing and Applications, Rome, Italy, 2014: 23-30.
[54]. Yingya Guo, Zhiliang Wang, Xia Yin, Xingang Shi, Jianping Wu. Traffic engineering in SDN/OSPF hybrid network[C]. Proceedings of IEEE 22nd International Conference on Network Protocols, North Carolina, USA, 2014: 563-568.
[55]. Miao Chuntang, Su Wei, Zhang Hongke, Zhou Huachun. A multi-path routing algorithm for intelligent collaborative network [J]. Electronic Journal, 2015,43 (10): 1881-1887.
[56]. Tang Lun Wu Ting Zhou Xinlong Chen Qianbin. A Virtual Network Function Migration Algorithm Based on Federated Learning Prediction of Resource Requirements[J]. Journal of Electronics & Information Technology, 2022,44(10): 3532-3540.
[57]. Xin Sun, Sen Su, Fangchun Yang. Adaptive Virtual Machine Replacement for Multi-dimensional a ware server consolidation in data centers [J]. Journal of Information and Computational Science, 2013, 10(3): 633-643.
[58]. Maheshwari Nitesh, Nanduri Radheshyam, Varma Vasudeva Dynamic energy efficient data placement and cluster configuration algorithm for map-reduce framework[J]. Future Generation Computer Systems, 2012, 28(1): 119-127.
[59]. Xiaoqiao Meng, Vasileios Pappas , Li Zhang Improving the scalability of data center networks with traffic-aware virtual machine placement [C].Proceedings of INFOCOM 2010, San Diego, USA, 2010: 1-9.
[60]. Even Guy, Naor Joseph, Rao Satish, Schieber Baruch. Fast a pproximate graph partitioning algorithms[J]. SIAM Journal on Computing, 1999, 28(6): 2187-2214.
[61]. Shao Sujie, Wu Lei, Zhong Cheng, Guo Shaoyong, Bu Xiande. Container Based Microservice Selection for Multi-workflow in Edge Computing Paradigm, Journal of Electronics & Information Technology, 2022, 44(11): 3748-3756.
[62]. CAI Yueping, Wang Changping.Software-Defined Data Center Network Hybrid Routing Mechanism[J]. Journal of Communications, 2016,37 (4): 44-52.

[63]. Mian Guo , Shengming Jiang, Quansheng Guan. QoS Provisioning performance of intserv, diffserv and DQS with multi-classself-similar traffic[J]. Transactions on Emerging Telecommunications Technologies, 2013, 24(6): 600-614.
[64]. Qirui Li, Zhiping Peng, Delong Cui, et al. A Two-stage Approach for Virtual Resources Adaptive Scheduling in Container Cloud[C]. 2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), 2020, :90-95.
[65]. Mian Guo, Quansheng Guan , Shengming Jiang. A differentiated queueing service based admission control policy for wireless multimedia [C].Proceedings of IEEE 1st International Conference on Communications, Sydney, Australia, 2014: 1308-1313.
[66]. Yuan Yao, LongboHuan, Abhishek B. Sharma. Power cost reduction in distributed data centers: A two-Time-scale approach for delay olerant workloads[J].IEEE Transactions on Parallel and Distributed Systems, 2014, 25(1): 200-211.
[67]. Lizhe Wang, Jie Tao, Rajiv Ranjan, Holger Marten, Achim Streit. G-Hadoop: Map reduce across distributed data centers for data-intensive computing [J]. Future Generation Computer Systems, 2013, 29(3): 739-750.
[68]. Wanfeng Zhang, Lizhe Wang, Yan Ma, Dingsheng Liu.Design and implementation of task scheduling strategies for massive remote sensing data processing across multiple data centers[J].Software Practice & Experience, 2014, 44(7): 873-886.