Overview of Cloud Computing

Lin Qian1,a, Jun Yu1,b, Guangxin Zhu1,c, Lin Wang1,d, Haiyang Chen1,e, Hengmao Pang1,f, Feng Mei2,g, Wenda Lu2,h, Zhu Mei1,i

1State Grid Electric Power Research Institute Nanjing, China
2State Grid Zhejiang Electric Power Company Information & Telecommunication Branch Hangzhou, China

aqianlin@sgepri.sgcc.com.cn, byujun@sgepri.sgcc.com.cn,
czhuangxin@sgepri.sgcc.com.cn, dpanghengmao@sgepri.sgcc.com.cn,
nmeifeng@zj.sgcc.com.cn, lwenda@zj.sgcc.com.cn,
emeifeng@zj.sgcc.com.cn, hwanglin18@sgepri.sgcc.com.cn,
imeizhu2016@aliyun.com

Abstract. With the development and operation of the Internet, the amount of data generated every day growing bigger, and the demand for data processing capacity is increasing. Cloud computing has become a hot research topic. The core problem of cloud computing is cloud resource scheduling. Current scholars have studied cloud resource scheduling problems from various aspects. This article mainly describes and analyzes cloud resource scheduling based on QoS performance constraints, cloud resource scheduling based on load balancing, and forecasting based on cloud resource scheduling, which provides a useful reference for cloud computing research.

1. Introduction
With the rapid development of cloud computing and big data, cloud computing organizes and aggregates computing and communication resources in a networked manner, and provides computing resources that can be scaled down or scaled for major data processing platforms in a virtualized manner. Cloud computing is a further development of distributed computing, parallel processing and grid computing. It is an Internet-based computing system that provides hardware services, infrastructure services, platform services, software services, and storage services to various Internet applications [1]. Usually the cloud system is served by a mechanism owned by a third party, and the user only cares about the services provided by the cloud.

With the participation of a large number of scholars, many concepts have been proposed and its definition has been unified. The definition of cloud computing in China's cloud computing standardization white paper: Cloud computing is a model that provides scalable, resilient, and shared physical and virtual resource pools on an on-demand, self-service basis and provides network access. The cloud computing model consists of key features, cloud computing roles and activities, cloud capability types and cloud service classification, cloud deployment models, and cloud computing common concerns [2]. According to the service model, cloud computing can be divided into software as a service (SaaS) [3], platform as a service (PaaS) [4], and infrastructure as a service (IaaS) [5]. According to the deployment method, it can be divided into: private cloud, community cloud, public cloud, hybrid cloud [6].
The cloud service provider allocates resources for it through the dispatch center according to the task application submitted by the user. The solution space of cloud resource scheduling is very large, so it belongs to the NP-hard problem [7]. In the past research, experts and scholars have studied the cloud resource scheduling problem from various angles, such as load balancing, reducing platform cost, establishing energy perception model, and improving resource utilization. In the constraints, consider the CPU, memory, network bandwidth and other conditions of the server and virtual machine on the model. For the cloud resource scheduling problem, this paper conducts an in-depth investigation and analysis of its related research status, focusing on the analysis of cloud resource scheduling based on QoS performance constraints, cloud resource scheduling based on load balancing, and cloud resource scheduling based on prediction.

2. Current research on cloud resource scheduling

In recent years, cloud computing has been applied to various industries in society. The rapid development of cloud computing is inseparable from the rapid development of cloud computing technology. The core technology of cloud computing is the management of resources, that is, how to schedule and allocate resources. In recent years, scholars have found that the method based on Meta heuristics has been proved to obtain approximate optimal solutions in a reasonable time [8]. The current research directions of cloud resource scheduling include: cloud resource scheduling based on QoS performance constraints, based on Load-balanced cloud resource scheduling, prediction-based cloud resource scheduling, energy-aware-based cloud resource scheduling, and cloud resource scheduling based on maximum utilization.

2.1. Cloud Resource Scheduling Based on QoS Performance Constraints

When scheduling tasks in a cloud environment, the tasks involved in scheduling usually have various quality of service (QoS) target constraints. They want to obtain better QoS guarantees from the cloud computing system to ensure meet the respective QoS target constraints of the scheduling task.

D. W. Sun et al. [9] analyze the impact of five aspects of final completion time, request response time, cost, availability, and security on resource allocation. The final completion time of the task indicates whether the resource can complete the corresponding task. The latest completion time is, it can be completed in the interval, and the value is 0 if the time out of the interval. Response time is used to evaluate the user's experience. Set and are two time points. In the interval, the effect value is constant 1. In the interval, effect value increases linearly with time. Otherwise the utility value is 0. In terms of expenses, is the maximum budget. In the interval. The higher the cost, the smaller the effect value. In other cases, the effect value is 0. Availability indicates the likelihood that the service can process the result, if the value of availability greater or equal to 0, the bigger, the higher effect value. The higher the security, the more the user experience can be improved., represents two security values. In the interval, effect value is 0. In the interval, the effect value is 0.5.In the interval, the effect value is 1. By introducing effect values, the literature achieves constraints on QoS from five aspects.

J. X. Ren et al. [10] consider the case of assigning tasks to virtual machines, mainly considering the three-dimensional QoS requirements of users, namely QoS time requirements, QoS performance requirements and QoS cost requirements. In terms of QoS time requirements, only the task calculation time is considered regardless of the communication time of the task. The length of the task and the performance ratio of the virtual machine are used to represent the time. In terms of QoS performance, the performance of the server is measured in terms of computing power(Yuan/million instructions). In terms of QoS fee requirements, (yuan/million instructions) is used to indicate the price the customer needs to pay. By comparing the three parameters to customer-defined requirements, if it meets the customer's requirements the value is 1 and if it exceeds the customer's request the value is 0. Establish user time, performance, and cost satisfaction matrix respectively. Finally, the three matrices are added to obtain the user satisfaction matrix.
2.2. Cloud Resource Scheduling Based on Load Balancing

The resource balance scheduling problem under cloud computing is currently a hot research field. Due to the diversity of user requirements, it is difficult for cloud data centers to achieve load balancing when allocating resources, which leads to service performance imbalance [11]. Therefore, cloud data centers resource scheduling and optimization has become a hot topic in recent years, and virtual machine placement is a means of maximizing cloud resource scheduling, and has become an important research topic of resource scheduling. Virtual machine placement is aimed at solving virtual and physical machines. The mapping relationship between the two is based on the set placement strategy, looking for an optimized virtual machine placement scheme to achieve the goal of maximizing resource load balancing [12].

K. Pradeep et al. [13] explain load balancing to evenly share the load between two or more different servers to get the job done quickly, use resources efficiently, and save energy. Load balancing can be achieved by scheduling virtual machines. Virtual machine scheduling is divided into static scheduling and dynamic scheduling. Static scheduling occurs during the initial allocation phase of the virtual machine. Dynamic scheduling, also known as dynamic load balancing, occurs during the redistribution phase. Dynamic scheduling will increase the running cost of the server.

C. Y. Qi et al. [14] point out that the load balancing algorithm can allocate tasks according to the remaining situation of each node resource, and achieve the balance of load between nodes. G. Ritchie et al. [15] proposes to calculate the average completion time of the task, and then assign the task to each computing node according to the time required by the task, so that each computing node is in a busy state; B. Chen et al. [16] proposes a computing cluster load balancing coefficient. The formula, and considers the strategy of task maximum priority and least server priority. Priority is given to tasks with high demand, and servers with low resource utilization are preferred.

J. Z. Zhu et al. [17] consider the virtual machine migration strategy. The load in the first K time period of the peak CPU load exceeds the threshold or the physical host with a particularly large change needs virtual machine migration. The physical host that needs to be migrated A virtual machine with a small memory footprint takes precedence over migration; Z.G.Hu et al. [18] point that when considering the virtual machine migration strategy, the timing of virtual machine migration is predicted, and the load of the t+1th time point is predicted by the pre-t time period. For selecting the virtual machine to be migrated, consider the CPU and memory usage of the virtual machine. If the memory usage of the physical machine is too high, migrate the virtual machine with high memory and CPU usage. Otherwise, migrate the physics with low memory usage. A virtual machine with too high CPU and memory usage in the machine.

2.3. Cloud resource scheduling based on prediction

In the cloud resource scheduling, there is a problem of virtual machine dynamic migration which may waste energy. By migrating virtual machines to shut down unnecessary servers, the effect of reducing energy consumption is achieved. By predicting the load of the server or virtual machine, the load of the server or virtual machine at the next moment is predicted, which provides a basis for selecting which virtual machines to migrate to which servers.

B. Wang et al. [19] propose a virtual machine scheduling method for load sensing and prediction. The model use the second exponential smoothing method to predict the resource load of the server in the prediction module. In the virtual machine migration module, the combination of MMT and MM is used to monitor the virtual machine's CPU, memory, and network bandwidth to filter out the virtual machines that need to be migrated, and then use the resource optimal adaptation strategy to select the target server, finally placing the virtual machine in the target server. This predictive model is more accurate than the auto regressive predictive model to predict server load conditions.

K. K. Qi [20] propose a cloud computing resource scheduling model based on dynamic trend prediction ant colony algorithm. The model is divided into load forecasting module, virtual machine scheduling module, virtual machine management module and host management module. In the load prediction module, the current load is compared with the previous load, if it is high, it is an upward trend,
and vice versa. The current increased load value or the reduced load value is compared with the historical average load value, and the load value is increased or decreased by a formula. If the predicted result is an upward trend, the added value is added to predict the next load value; if the predicted result is a downward trend, the reduced value is subtracted to predict the next load value.

W. Zhou et al. [21] use the Kalman predictor to predict the remaining virtual resource. The Kalman filter can predict the next resource according to the existing resources, and then compare it with the current situation and the forecast. Changing the feedback parameters, and iteratively optimize the prediction model to make the model better. The predicted resource situation is provided to the cloud resource scheduling model to improve scheduling efficiency.

3. Conclusion
Cloud computing has become the infrastructure of the next generation IT industry, but cloud resource scheduling is still an important issue to be solved by cloud computing. This paper focuses on the research and analysis of cloud resource scheduling based on QoS performance constraints, cloud resource scheduling based on load balancing, and cloud resource scheduling problem based on prediction. In the future, we can consider solving cloud resource scheduling problems from the following perspectives. For example, improving the accuracy of cloud resource scheduling from the perspective of multi-objective issues, saving operating costs of the platform from the perspective of reducing cloud resource scheduling costs, and considering cloud resource scheduling issues from the perspective of improving server resource utilization et al.

Acknowledgments
This work was financially supported by the science and technology projects of headquarters of State Grid Corporation (Research on key technologies of secure network coding and data protection for secure renewable storage System).

References
[1] J. X. Zhang, Z. M. Gu, C. Zheng. Summary of research progress in cloud computing. Computer Application Research, 2010.
[2] http://www.cac.gov.cn/2014-09/29/c_1112671503.htm.
[3] C. Michael. Cloud computing and SaaS as new computing platforms. Communications of the ACM, 2010.
[4] C. Pahl. Containerization and the PaaS Cloud. IEEE, 2015.
[5] H. Madni, A. L. M Shafie, C. Yahaya, et al. Resource Scheduling for Infrastructure as a Service (IaaS) in Cloud Computing: Challenges and Opportunities. Journal of Network and Computer Applications, 2016.
[6] W. W. Lin, D. J. Qi. A Survey of Cloud Computing Resource Scheduling. Computer Science, 2012.
[7] L. Z. Du, H. D. Fu, H. Zhang, Y. L. Huang. Research on P and NP. Computer Technology and Development, 2013.
[8] M. Kalra, S. Singh. A review of metaheuristic scheduling techniques in cloud computing. Egyptian Informatics Journal, 2015.
[9] D. W. Sun, G. R. Chang, F. Y. Li, C. Wang, X. W. Wang. A Preference Multidimensional QoS Cloud Resource Scheduling Optimization Algorithm Based on Immune Cloning. Chinese Journal of Electronics, 2011.
[10] J. X. Ren, X. K. Zhong. Multidimensional QoS Constraint Cloud Task Scheduling. Microelectronics and Computer, 2018.
[11] A. G. Greenberg, J. R. Hamilton, D. A. Maltz, et al. The cost of a cloud: research problems in data center networks. ACM SIGCOMM Computer Communication Review, 2008.
[12] Y. B. Lu, J. Y. Wang, J. M. Wu. Research on cloud computing resource load balancing scheduling optimization algorithm, 2017.
[13] K. Pradeep and T. P. Jacob, Comparative analysis of scheduling and load balancing algorithms
in cloud environment, 2016 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), Kumaracoil, 2016.

[14] C. Y. Qi, W. S. Wen, Y. Xiao, Y. X. Liu, Y. Jia. A load balancing strategy for adaptive file system metadata services. Journal of Software, 2017.

[15] G. Ritchie, J. Levine. A fast, effective local search for scheduling independent jobs in heterogeneous computing environments. Journal of Computer Applications, 2003.

[16] B. Chen, M. L. Gan, J. Li. A Cloud Resource Scheduling Method Based on Load Balancing. Computer Technology and Development, 2017.

[17] J. Z. Zhu, D. Xiao, F. Wang. Multi-dimensional QoS constraint task scheduling mechanism for load balancing under cloud computing. Computer Engineering and Applications, 2013.

[18] Z. G. Hu, C. OuYang, C. K. Yan. Resource load balancing method for reducing energy consumption under cloud environment. Computer Engineering, 2012.

[19] B. Wang, Q. W. Wang, K. Dong, J. F. Sheng. Research on virtual machine scheduling method based on quadratic exponential smoothing prediction. Computer Application Research, 2017.

[20] K. K. Qi. Research on cloud computing resource scheduling optimization based on dynamic trend prediction ant colony algorithm. Science and Technology Bulletin, 2016.

[21] W. Zhou, Z. Z. Wang. Research on Cloud Computing Resource Scheduling Based on Kalman Predictor. Software, 2015.