Journal of Physics: Conference Series

PAPER • **OPEN ACCESS**

Large-Scale Power Cloud Resource Scheduling Algorithm in an Edge Computing Environment

To cite this article: Xiang Huang *et al* 2022 *J. Phys.: Conf. Ser.* **2404** 012055

Large-Scale Power Cloud Resource Scheduling Algorithm in an Edge Computing Environment

Xiang Huang[1*], Zhihong Liang[1], Qiankun Zhang[1], Jianfeng Mo[1], Lei Zhao[1]

[1]CSG Digital Power Grid Research Institute Co., Ltd, Guangzhou, Guangdong, 510700, China

[*]Corresponding author: huangxiang@csg.cn

Abstract. Aiming at the problem that the current power cloud resource scheduling is inefficient and time-consuming, a large-scale power cloud resource scheduling algorithm in an edge computing environment is proposed. Firstly, the resource management method of large-scale power clouds is optimized. Considering the resource sharing and task scheduling security requirements between secondary nodes, and combined with edge computing technology, the resource distribution management and scheduling algorithm of the power cloud are optimized. Finally, the experiment proves that the resource scheduling time of a large-scale power cloud resource scheduling algorithm under an edge computing environment is reduced by more than 15 minutes in the actual application process, And the scheduling effect is also better than the traditional method.

1.Introduction

With the improvement of the intelligence level of the power grid in China and the construction of the energy Internet in the world, more and more intelligent devices are connected to the power grid. Therefore, how to efficiently dispatch these massive resources is an urgent problem to be solved. With the continuous development of power grid informatization and the development of cloud computing, power enterprises are more and more applied in the virtual environment [1]. With the continuous growth of power grid services, the characteristics of real-time, high parallelism and high capacity of power grid equipment make the load and calculation amount in the power grid environment gradually increase. Therefore, how to effectively improve the computing performance and utilization efficiency of the power grid, and how to establish a scalable virtual resource library are the topics that many people are discussing. At present, in a large number of existing research, the most commonly used resource planning based on job planning has a high self-adaptive ability. It is an effective large-scale computing method and has been widely used in virtual machine clusters [2]. However, this method has some defects, that is, it is an application-oriented, object-oriented, distributed and application-level sorting technology. The current resource allocation method can realize real-time resource allocation, but it can only make corresponding adjustments and responses when the system is short or excessive. Therefore, under this background, a large-scale power grid resource allocation algorithm will be given based on edge computing [3].

2.Large-scale power cloud resource scheduling algorithm

2.1.Large scale power cloud resource management scheme

With the rapid development of the Internet and the widespread application of cloud computing technology in the power field, a large number of intelligent terminals are connected to the cloud. To improve management efficiency and reduce maintenance cost, the traditional resource management scheme uploads a large amount of data collected from the terminal to the cloud [4]. However, with the rapid development of the smart grid technology, an increasing number of access devices and the increasingly frequent information exchange, this cloud resource management mode can't meet the needs of large-scale data processing. In the field of power business, there are many device nodes with computing capabilities at the edge of the network. If the computing capabilities of these edge devices can be utilized, cloud computing-based services can be well provided [5]. Therefore, this paper proposes a power cloud resource management scheme based on edge computing. In this scheme, the application platform sinks the computing task to the big data platform of the provincial company, shares the storage resources, computing resources and data resources of the big data platform for business algorithm calculation, and sends the calculation results back to the cloud. In the power cloud resource management scheme based on edge computing, data computing is divided into two levels: the first-level analysis and control node and the second-level calculation node of the headquarters, as shown in Figure 1.

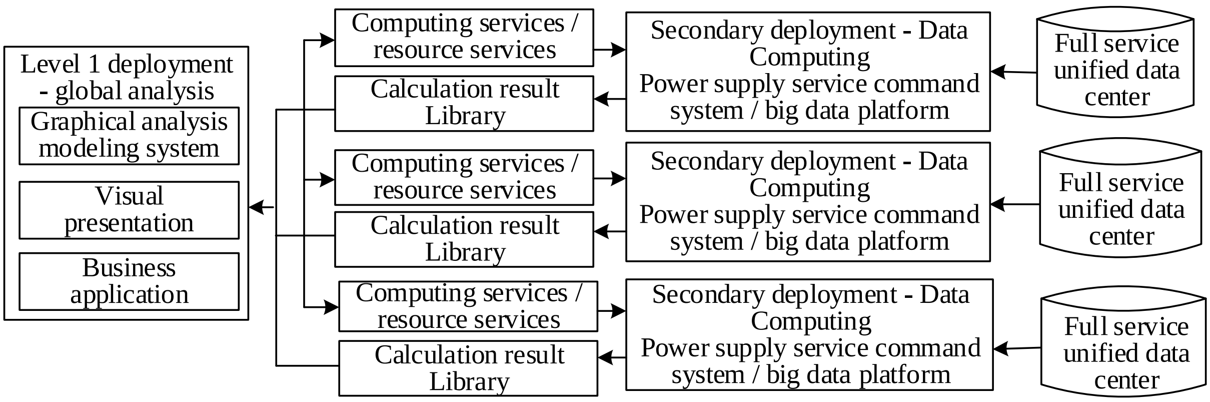

Figure 1.Power cloud resource management scheme based on the edge computing

The first level node mainly provides functions such as graphic analysis and modeling, algorithm distribution and management. The provincial level node is a sinking computing node, which performs unified scheduling of provincial computing service resources and connects with the power supply service command system and big data platform. The main application scenarios of this computing mode include holographic customer portrait construction and hot spot positioning. In the power cloud resource management scheme based on edge computing, the secondary node acts as the computing node, but the resources are limited, and sometimes large-scale computing tasks cannot be successfully performed locally [6]. Therefore, this paper considers that resource sharing and reasonable scheduling between secondary nodes can be realized by allocating the redundant sub-tasks that cannot be completed locally to other nodes with idle resources. Compared with traditional methods, edge computing has managed and allocated resources in a smaller granularity, improving the utilization and energy consumption of edge computing resources. However, if virtual resource management is fully implemented according to the user's application, a large number of active resources will be idle, and the resources and energy consumption are still wasted. Since the user is not sure about the resources, the applied resources often exceed the actual needs. To make more effective use of resources, we need to have a more accurate grasp of resource demand and make resource planning in advance [7]. The capacity demand for cloud computing increases gradually with the passage of time. The total demand does not increase monotonously, but fluctuates to a certain extent. To avoid the waste of idle resources as much as possible and ensure service quality, the resources are divided into two parts. The first part is called "periodic resources", which is mainly to meet most of the resource requirements of the cloud center and ensure the smooth operation of the resource center. The resource requirements can be found according to the historical data analysis to meet the relatively stable resource requirements of the data center; the second part of resources is called "peak period resources", which mainly meet the peak resource demand of each stage. Through resource division, resources can be more finely managed, periodic resources can be used to meet basic needs, and then peak period resources can be dynamically enabled or closed according to peak demand. Of course, the startup and shutdown of virtual resources take time, and the startup and shutdown of resources have different impacts on the application system. Therefore, in practical applications, the time granularity of resource management can be adjusted according to the actual situation of the user's application system [8]. The technical framework of resource planning and scheduling based on data mining is shown in Figure 2.

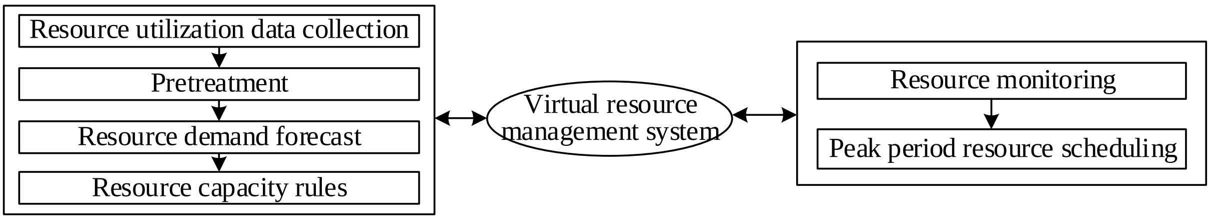

Figure 2.Cloud resource planning and scheduling technical framework

The development of information technology and communication technology has created favorable conditions for edge computing technology [9]. The basic principle of edge computing technology is to physically separate and logically integrate the processing and calculation of information by using distributed hardware systems and distributed software systems, and then summarize the results as required. In edge computing technology, the core is based on an edge computing process, which can combine huge system hardware resources to serve the same goal [10]. At the same time, the physical results of the processing can also be stored in many distributed computer systems according to a certain logical relationship, and provide on-demand access and calculation. In terms of physical entities, the technical architecture of edge computing is divided into four parts, which constitute the four-tier relationship of the edge computing technology platform [11]. Each layer provides physical facilities support for the upper layer and logical services for the lower layer.

2.2.Resource distribution algorithm based on edge computing

The load-balancing problem of edge computing is proposed because of the emergence of distributed systems. In distributed systems, load balancing is the core function of system scalability. A task is waiting for the service of a resource in the queue, and at the same time, another resource may be idle. This is an objective situation, which leads to a load of a server being too high and running slowly, while other servers are not running and consuming power in vain. The significance of load balancing is to prevent this situation as much as possible [12]. Load balancing divides the traffic to multiple servers through algorithms, makes full use of the server's computing resources and ensures that the server's workload is maintained in a balanced state to prevent its load from being too high [13]. By distributing customer requirements to various servers, the time delay of the server is reduced, and the response speed of the system is greatly improved. In the cloud environment, the servers can be distributed in the data centers of different companies and institutions all over the world. Even if a data center is damaged due to some irresistible factors, it can still coordinate the processing of customer requests and tasks through servers distributed in other regional data centers. The infrastructure construction of the private cloud is relatively complete. Load balancing can use a group of dedicated servers to serve specific enterprise clients [14]. For parallel programs, load balancing tries to distribute the workload to multiple processors as much as possible to improve the performance of the system. Generally speaking, a load-balancing mechanism includes information collection, policy selection and data migration. In the information collection stage, the load balancer will collect information about the workload and the current state of the computer to determine whether the current system load is balanced. The strategy selection stage focuses on calculating the most ideal data distribution mode, while the data migration stage transfers the excessive load from the overloaded server to other servers. The power cloud is a multi-task parallel mode, and it needs to be able to predict the system performance as accurately as possible before assigning tasks. If a task is assigned to a processing node s in the power cloud, it is necessary to predict the completion time of the task [15]. Completing a user task mainly includes two operations: transferring the task to the computing node and executing the task on the computing node. Therefore, it is necessary to predict the following two points: how long it will take to transmit this task to the processing node s, that is, to predict the expected transmission time of each line transmission unit task from the master node to the node $ETT\left(e_{ij}\right)$. When the task reaches i, how long does j need to complete it. Under the current processing capacity of the node, the estimated execution time of the processing unit task $ETT\left(e_{ij}\right)$. Suppose a certain moment t, for a transmission line: v, we need to predict the number of $k+1$ unit tasks in a. The time required for the transmission of the task package is M. Can be obtained by edge computing t timeline v Available bandwidth on Bandwidth, then t time can be obtained by the formula: $k+1$ unit tasks on the line v estimated transfer time on $ETT\left(e_{ij}\right)$:

$$ETT\left(e_{ij}\right) = M(i,j)/\text{ Bandwidth } [a - tv + (k+1)] \tag{1}$$

Hypothetical node d and node r between f transmission line is connected with: $f_0, f_1, \dots \dots, f_{n-l}$ then the first $k+1$ unit tasks from d node transfer to r estimated transmission time of the node:

$$W_{d\to r} = k + 1\prod ETT\left(\mathrm{e}_{ij}\right) + f_{n-l}(r-d) \tag{2}$$

The comprehensive benefits of edge computing task scheduling are divided into performance benefits and economic benefits. Performance benefit mainly refers to task execution time and system load, while economic benefit mainly refers to user task execution cost. The evaluation function of low-performance performance benefit should be avoided due to excessive user expense and excessive pursuit of performance or low cost:

$$TimePer\ (i,j) = N - \frac{W_{d\to r}}{\sum_{j=1}^{N}\sum_{i=1}^{N} qt_{ij}\times sl_{ij}} \tag{3}$$

Where t_{ij}, l_{ij} are the expected completion time matrix of the task q and resource task allocation matrix sl elements corresponding to positions, N is the reference coefficient. The performance benefit evaluation function shows that the task execution time is different with different task allocation methods. The migration quality is that after the virtual machine is transferred from a hot spot, this problem can be well solved. From the above, it can be seen that physical nodes use CPU utilization to determine the heat source. Therefore, this paper proposes a method that combines CPU utilization with the volume of the virtual machine to ensure the quality of the migration when selecting the migration virtual machine.

$$S = U_V - R_{\text{ram}}/\text{TimePer } (i,j) \tag{4}$$

Where, U_V is the utilization rate of the virtual machine CPU, $and\, R_{\text{ram}}$ is the memory of the virtual machine. When the CPU utilization is high and the memory is small, the s value is large. At this time, the CPU utilization of virtual machines has an important impact on SLA. By transplanting it, the CPU utilization of physical nodes can be greatly reduced. As this system has a very small storage space and the amount of storage required in the migration process is also very low, the bandwidth and CPU clock cycle required for migration can be reduced, thus greatly reducing the running time in the migration process [16]. The virtual machine migration program requires copying the CPU field, memory data and other related parameters. If the migration is not performed in the same data environment, more data will be generated. In this paper, a space volume measurement formula based on a virtual machine is proposed. By selecting a smaller virtual machine, the virtual machine can be transplanted quickly.

Virtual machine volume calculation function (such as formula):

$$\mu = S - \frac{1}{1-U_v} \times \frac{1}{1-R_{ram}} \tag{5}$$

Among them, the higher the utilization rate of μ, the larger the virtual machine volume Z, and the more data γ_s^i needs to be copied during migration, so the more the migration time and the physical node η_s^i occupied. Suppose that the estimated execution time $T_s^n(k+1)$ for a certain node k to process a certain unit task can be obtained by the following formula:

$$T_s^n(k+1) = \mu \prod_{i=2}^{n} \left[(1-Z)\eta_s^i(k) + Z\gamma_s^i(k)\right] \times S \tag{6}$$

Based on the research of the power cloud computing system architecture and combined with the user task execution process, the power cloud computing task scheduling system based on edge computing is designed. The overall architecture of the system is shown in Figure 3 of the power cloud task scheduling system architecture below, which is mainly composed of a task request queue, task segmentation component, task scheduler, resource scheduler, resource pool, virtualization layer and physical resource layer, et.

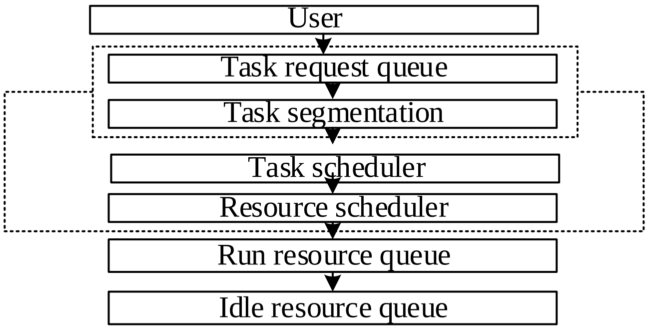

Figure 3.Power cloud task scheduling system architecture

The core components of the architecture comprise work demand queuing and task demand queuing. Task demand queuing refers to the work proposed by the user, and the user's work content includes: first, the nature of the task itself, such as whether the work has a high computing capacity or a high storage capacity, and the task and other related contents; the second is the limitation of the task, such as the service quality requirements and the cost of task execution. The third is the minimum requirements for system configuration, CPU and memory capacity, and power grid bandwidth [17]. Task decomposition element: the task decomposition element divides the tasks and programs proposed by the user into several smaller operations. The main basis of task decomposition is to identify the parallel operations required by each operation, the operations required by the operation in multiple operations, and the data and data required by the task. Task scheduler: make use of the results of task division, correlate the working states of software and hardware resources in the resource allocation program, establish the corresponding task allocation scheme, and map it to the resource allocation program [18]. According to the existing power grid environment, power grid bandwidth and power grid availability, energy can design an efficient dispatching scheme. This method makes a good correspondence between the user's task requirements and the resources of the power cloud so that the task requirements can be met to a certain extent and the security of the power grid can be ensured. On this basis, the optimal scheduling scheme is determined by several evaluation criteria, and each resource scheduling system assigns each task to the corresponding resource node according to the mapping relationship. Resource Planning: to realize the operation planning in the power grid of the power grid, it is necessary to understand the use of various resources in the power grid in advance, including both utilized and available resources. A special monitoring node is set in the power grid to monitor the status of each node in real-time, and can know the current use status, availability and load balance of each node at any time. In resource scheduling, the resources in the resource pool are configured by selecting the resources in the resource pool, to realize the mapping of resources and match their relationship with resources. According to the existing power grid environment, power grid bandwidth and power grid availability, energy can design an efficient dispatching scheme. This method makes a good correspondence between the user's task requirements and the resources of the power cloud, so that the task requirements can be met to a certain extent and the security of the power grid can be ensured [19]. On this basis, the optimal scheduling scheme is determined by several evaluation criteria, and each resource scheduling system assigns each task to the corresponding resource node according to the mapping relationship. Resource Planning: to realize the operation planning in the power grid of the power grid, it is necessary to understand the use of various resources in the power grid in advance, including both utilized and available resources. A special monitoring node is set in the power grid to monitor the status of each node in real-time and can know the current use status, availability and load balance of each node at any time. On this basis, the job plan generated by the job plan is used to select the appropriate resource nodes from the resource pool and allocate the jobs, to realize the real correspondence between user jobs and resource nodes.

2.3.Realization of power resource dispatching

Through the analysis of the business system, the resource scheduling related to the system is summarized. According to the different characteristics of resource abstraction levels and objectives targeted by various scheduling technologies, the resource scheduling technologies under the virtualization environment can be divided into physical layer resource scheduling, virtual layer resource scheduling and application layer resource scheduling from bottom to top. The research objective is to ensure the stable operation of the application system, Combined with the functions of current mainstream virtualization products (therefore, the dynamic allocation of internal resources of virtual machines is not within the scope of this study), energy consumption is saved and user service quality is improved [20]. Therefore, the model of this study is the research of resource scheduling adjustment strategy based on virtual machine migration technology. To dynamically deploy cloud resource capacity, it should be considered when to deploy, how much to deploy, and how to deploy. This paper will provide an effective capacity evaluation method to analyze and evaluate cloud resource capacity based on the cloud resource application status evaluation model, as shown in Figure 4.

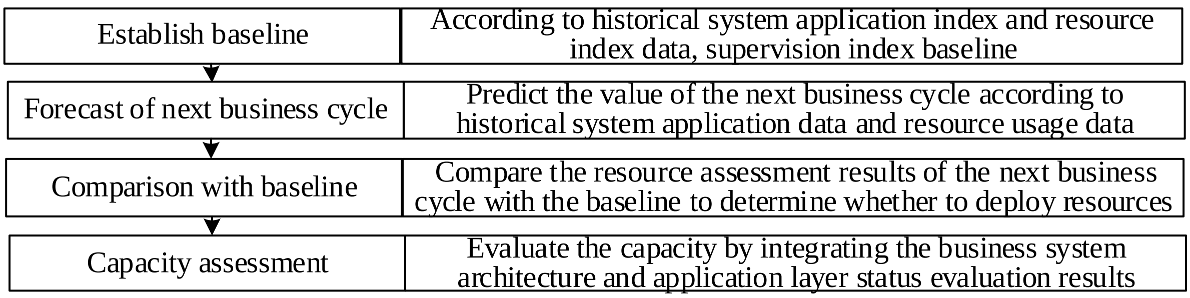

Figure 4.Overall idea of resource capacity assessment model construction

Each online business system operates on the cloud resource platform and generates a large amount of historical data. For example, the historical data of system operation indicators includes cumulative visits, the number of online users, the number of concurrent users, the average response time of the system, TPS and other indicators. Many units have corresponding network management systems or cloud resource management platforms to monitor business system resources and ensure the reliability of resources, as shown in Table 1:

Table 1.Weight of dispatching evaluation factors

Level I indicator	First level weight	Secondary indicators	Secondary weight
Number of the system online	0.357	Number of online users	0.722
		Concurrent users	0.278
System response time	0.303	Average response time	0.658
		Maximum response time	0.345
Resource utilization	0.343	CPU utilization	0.496
		Memory usage	0.506

The power equipment condition assessment center is an important part of the power data center. Its main task is to realize the management and maintenance of data, regularly carry out equipment assessment, formulate maintenance strategies and plans, and realize the professional management of equipment condition maintenance. The power equipment resource scheduling model is shown in Figure 5.

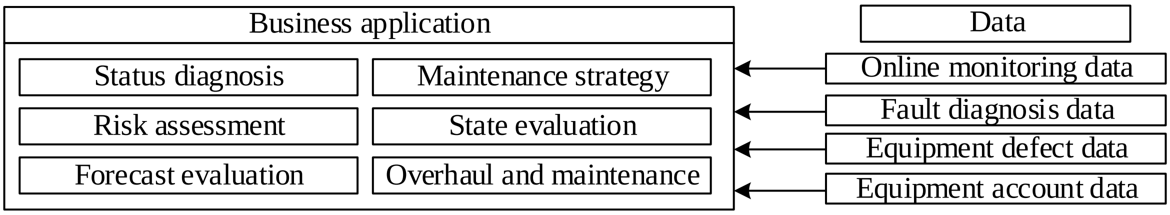

Figure 5.Power equipment resource scheduling model

It can be seen from the figure that the power equipment resource dispatching model center has the functions of state diagnosis, prediction and evaluation, state evaluation, risk assessment, overhaul strategy, overhaul and maintenance to improve the effective dispatching management of massive power resources and ensure the safety of resource dispatching.

3.Analysis of experimental results

To verify the advantages of the edge algorithm, this section carries out experimental verification of the algorithm. The experiment is divided into two parts for verification. Firstly, the model is verified. In the verification process of the proposed resource scheduling model, lingoll is used to calculate and solve the model because it has a powerful computing function and data processing ability, and is more powerful than MATLAB. The power cloud platform is composed of four computers with the same configuration. The specific configuration process is as follows: first, the following software is installed on each machine. All computers are installed with Ubuntu 604 system, and the JDK version is jdk-8u5 - linux-x64argz. The Hadoop version is Hadoop 0.20.0. The hardware configuration standard of the four machines in the cluster is: 24g memory, 320G hard disk, 2.6g main frequency. The configuration of each node in the Hadoop cloud platform is shown in Table 2:

Table 2.Configuration hostname of each node in cloud resources

host	IP address	function
Master	192.168.199.86	NameNode(JobTracker)
Slaver1	192.168.199.87	DataNode(JobTracker)
Slaver2	192.168.199.88	DataNode(JobTracker)
Slaver3	192.168.199.89	DataNode(JobTracker)

Cloudsim is used to build a simulation platform for power transportation resource scheduling. The specific experimental environment configuration is shown in Table 3:

Table 3.The experimental environment

Experimental environment	Equipment parameters
Hardware environment	Intel(R)Core(TM) i5-2430M 2.4GHz
Software environment	Windows XP
Development platform	My Eclipse 8.5 IDE, Matlab 10

Finally, the optimal solution of the scheduling algorithm is obtained by analyzing the scheduling effect value, completion time and load balancing degree; among them, efficiency is the main evaluation method, and the results are shown in Figure 6:

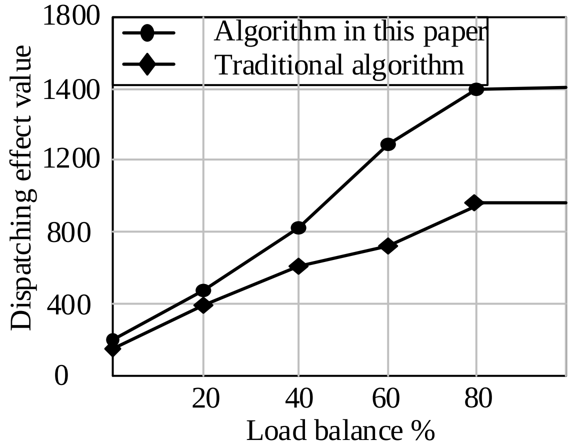

Figure 6.Comparison of scheduling effect values under different task numbers

It can be seen from the figure that with an increase in the number of tasks, the overall scheduling effect values of the two methods show an upward trend, but the scheduling effect under the guidance of this method is better. The completion time of the scheduling task is one of the indicators to measure the scheduling effect, and is also an important indicator considered by most scheduling algorithms. Based on this, the resource scheduling completion time of the two methods is further compared. The specific results are shown in Figure. 7:

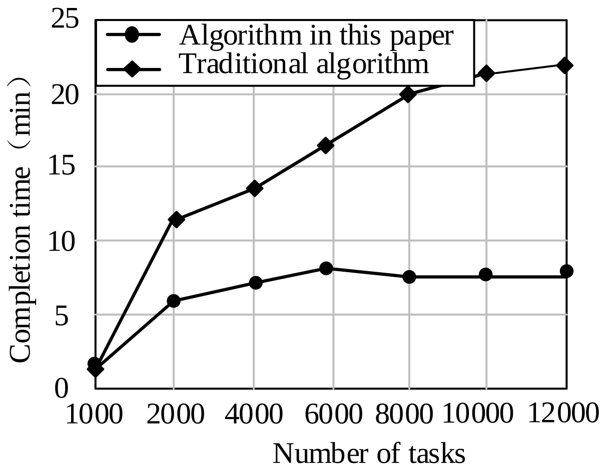

Figure 7.Comparison of completion time under different task numbers

Based on the above detection results, compared with the traditional methods, the scheduling effect and time consumption of this method are significantly better than those of the traditional methods, achieving the scheduling objectives of high efficiency, real-time and safety and meeting the research requirements.

4.Conclusion

In this paper, the technologies of distributed environment construction, data noise reduction, data integration and edge operation are used to build a benchmark that can be used for resource adjustment. And the method of boundary operation is used to understand the demand of enterprises for resources through boundary operation, and the reasoning ability is configured according to different service demands, to achieve the best utilization efficiency, improve the stability of the system, and ensure the efficient operation of various businesses in the power market.

Reference

[1] Jing, K. L., Si-yu, F., & Ya-fu, Z. (2020). Model predictive control of the fuel cell cathode system based on state quantity estimation. Computer Simulation, 37(7), 119-122.
[2] Li, B., Chen, P., Liu, H., Guo, W., Cao, X., Du, J., ... & Zhang, J. (2021). Random sketch learning for deep neural networks in edge computing. Nature Computational Science, 1(3), 221-228.
[3] Sheng, S., Chen, P., Chen, Z., Wu, L., & Yao, Y. (2021). Deep reinforcement learning-based task scheduling in iot edge computing. Sensors, 21(5), 1666.
[4] Liu, G., Chen, X., Zhou, R., Xu, S., Chen, Y. C., & Chen, G. (2021). Social learning discrete Particle Swarm Optimization based two-stage X-routing for IC design under Intelligent Edge Computing architecture. Applied Soft Computing, 104, 107215.
[5] Tritschler, N., Dugenske, A., & Kurfess, T. (2021). An automated edge computing-based condition health monitoring system: with an application on rolling element bearings. Journal of Manufacturing Science and Engineering, 143(7).
[6] Nie, L., Wu, Y., Wang, X., Guo, L., Wang, G., Gao, X., & Li, S. (2021). Intrusion detection for secure social internet of things based on collaborative edge computing: a generative adversarial network-based approach. IEEE Transactions on Computational Social Systems, 9(1), 134-145.
[7] Koo, S., & Lim, Y. (2021). A multi-objective computation offloading algorithm for dependent tasks based on a mobile edge computing environment. KIISE Transactions on Computing Practices, 27(2), 122-127.
[8] Zhou, Z., Jia, Z., Liao, H., Lu, W., Mumtaz, S., Guizani, M., & Tariq, M. (2021). Secure and Latency-Aware Digital Twin Assisted Resource Scheduling for 5G Edge Computing-Empowered Distribution Grids. IEEE Transactions on Industrial Informatics, 18(7), 4933-4943.
[9] Zhang, Q., Gui, L., Zhu, S., & Lang, X. (2021). Task offloading and resource scheduling in hybrid edge-cloud networks. IEEE Access, 9, 85350-85366.
[10] Lee, S., Kim, H., Shin, H., Kim, T. H., & Kim, W. (2018). A study on the estimation of optimal ESS capacity considering REC weighting scheme. The Transactions of the Korean Institute of electrical engineers, 67(8), 1009-1018.
[11] Wang, S., Pi, A., & Zhou, X. (2021). Elastic Parameter Server: Accelerating ML Training with Scalable Resource Scheduling. IEEE Transactions on Parallel and Distributed Systems, 33(5), 1128-1143.
[12] Dewangan, B. K., Venkatadri, M., Agarwal, A., Pasricha, A., & Choudhury, T. (2021). An Automated Self-Healing Cloud Computing Framework for Resource Scheduling. International Journal of Grid and High-Performance Computing (IJGHPC), 13(1), 47-64.
[13] Liang, B., Obaidat, M. S., Liu, X., Zhou, H., & Dong, M. (2021). Resource scheduling based on priority ladders for multiple performance requirements in wireless body area networks. IEEE Transactions on Vehicular Technology, 70(7), 7027-7036.

[14] Meyer, V., Kirchoff, D. F., Da Silva, M. L., & De Rose, C. A. (2021). ML-driven classification scheme for dynamic interference-aware resource scheduling in cloud infrastructures. Journal of Systems Architecture, 116, 102064.
[15] Duan, L., Ma, S., Li, Y., & Yao, X. (2021). Unified management method for multi-dimensional basic power resources. Procedia Computer Science, 183, 827-832.
[16] Kamruzzaman, M., Duan, J., Shi, D., & Benidris, M. (2021). A deep reinforcement learning-based multi-agent framework to enhance power system resilience using shunt resources. IEEE Transactions on Power Systems, 36(6), 5525-5536.
[17] Han, S. (2021, February). Research on interactive response strategy of power demand-side resources participating in the power grid. In IOP Conference Series: Earth and Environmental Science (Vol. 657, No. 1, p. 012109). IOP Publishing.
[18] Xiong, J., Hou, B., & Du, X. (2021). Research on the optimal allocation of protective resources of power network under malicious attacks. In IOP Conference Series: Earth and Environmental Science (Vol. 645, No. 1, p. 012043). IOP Publishing.
[19] Zhu, W., Cheng, X., Sun, L., Cong, L., Wang, B., Tian, Z., & Wang, N. (2021, February). Optimal utilization of load side power and heat resources based on aggregator mode. In Journal of Physics: Conference Series (Vol. 1754, No. 1, p. 012028). IOP Publishing.
[20] Xue, S., Wu, Z., Zhang, H., Zhang, G., & Cong, P. (2021, July). Evaluation model of key driving factors for different types of demand side distributed power resources to participate in market transactions. In IOP Conference Series: Earth and Environmental Science (Vol. 829, No. 1, p. 012009). IOP Publishing.