ISSN:2790-1661 Volume-14-(2025)

Research on Intelligent Factory-oriented Edge Cloud Collaborative Automated Scheduling and Resource Optimal Allocation Technology

Jiatong Zhu*

Shanghai Huachuang Automation Engineering Co., LTD, Shanghai, China

409746040@qq.com

Abstract. Smart factories are faced with the challenges of heterogeneous equipment, dynamic tasks and bandwidth and delay of massive data processing. Traditional centralized cloud computing scheduling has high delay, bandwidth pressure and single point failure risk, while single edge computing is limited by insufficient resources. This study focuses on the automatic scheduling and resource optimization allocation technology under the edge cloud collaborative architecture. This paper designs a three-tier edge cloud collaboration architecture of "hierarchical awareness-dynamic collaboration", and realizes refined resource management and collaboration through edge layer resource awareness agent, collaborative layer dynamic service bus and cloud layer digital twin engine. A hybrid scheduling algorithm named "Dynamic Divide and Reinforcement Scheduling" (DDRS) is proposed, which combines the dynamic divide and conquer of tasks with feedback reinforcement learning mechanism to achieve low delay and high robustness scheduling. A multi-dimensional game-cooperative equilibrium resource allocation model is constructed, and multi-resource joint optimization is realized based on Nash equilibrium and ADMM (Alternative Direction Method of Multipliers) algorithm, and an elastic migration mechanism is designed to deal with resource overload. The experimental results show that compared with pure edge and pure cloud scheduling, DDRS reduces the average delay to 63ms and the overtime task rate to 5.3%. The resource optimization model reduces the standard deviation of resource utilization from 0.41 to 0.18, reduces the number of elastic migration to 9 times, and can recover within 120ms when the node fails, which verifies the effectiveness of the technology in improving efficiency, stability and fault tolerance. This research provides theoretical and practical support for the intelligent upgrading of the whole chain of smart factories.

Keywords: resource optimal allocation; edge cloud collaborative; intelligent factory; automatic scheduling; dynamic divide and reinforcement scheduling.

1. Introduction

With the development of Industry 4.0 and intelligent manufacturing, the global manufacturing industry is undergoing a transformation to intelligence and flexibility. As the core of this transformation, smart factories realize device interconnection, data-driven and real-time decision-making through technologies such as Internet of Things, AI and digital twinning. However, the promotion of smart factories still faces three major challenges: first, the compatibility problem caused by the heterogeneity of equipment; second, the ability to adapt quickly under the dynamic requirements of tasks; and third, the bandwidth and delay problems when processing massive data. These challenges limit the large-scale implementation and application efficiency of smart factories.

Edge-cloud collaborative computing (ECCC) is a computing model that combines edge computing and cloud computing, which can effectively solve the problems of delay and bandwidth limitation faced by traditional cloud computing [1]. By processing data near the data source, ECCC can significantly reduce the network transmission delay and improve the data processing speed, thus better meeting the real-time and reliability requirements of smart factories. Resource allocation and task scheduling in smart factories is a complex optimization problem, which involves the geographical dispersion and heterogeneity of computing resources and the requirements for performance, energy consumption, cost and stability [2-3]. In order to solve these problems, researchers have proposed a variety of optimization algorithms and scheduling strategies. For

example, reference [4] proposed a heuristic algorithm based on second-order difference to solve the resource allocation problem of IO-intensive virtual machines under cloud-side collaborative computing architecture. In addition, literature [5] reviews the research on the application of virtualization, intelligent algorithm, digital twinning and zero-defect manufacturing in scheduling framework, and discusses how to optimize the production process by realizing zero interference and zero interruption SMS mode.

Traditional intelligent factory scheduling system mostly relies on centralized cloud computing, which has powerful computing power, but faces high delay, bandwidth pressure and single point of failure risk, and it is difficult to meet the real-time and reliability requirements [6-7]. With the development of edge computing, localization has become a new direction, but due to the lack of resources and heterogeneity, the ability to support complex tasks alone is limited. Therefore, edge cloud collaboration has become a development trend, aiming at combining the advantages of edge and cloud to build an intelligent scheduling system with low delay, high reliability and scalability [8].

This study focuses on automatic scheduling and resource optimization under the edge cloud collaborative architecture, and is committed to solving three key problems: dynamic resource perception, low delay and high robustness scheduling algorithm design and joint optimal allocation of edge cloud resources. By overcoming these technical bottlenecks, we can provide an efficient, flexible and landing collaborative scheduling scheme for smart factories, and promote the manufacturing industry to realize the intelligent upgrade of the whole chain from perception to decision-making to execution. This research not only helps to improve the theoretical system of edge cloud collaboration in industrial scenes, but also is expected to improve production efficiency, reduce energy consumption and enhance system fault tolerance in practice.

2. Design of edge cloud collaborative architecture

With the rapid development of intelligent manufacturing, intelligent factory, as an important carrier to realize intelligent manufacturing, its internal automatic scheduling and resource optimization allocation technology is particularly important [9]. As a new computing paradigm, ECCC provides an efficient resource management and task scheduling solution for smart factories by combining the powerful processing power of cloud computing with the low latency of edge computing [10]. In order to adapt to the computing power and storage limitations of edge devices, it is necessary to compress and adapt the model to reduce the model size and improve the operation efficiency. Through effective resource management strategies, the reasonable allocation and utilization of computing resources can be realized and the resource utilization rate can be improved [11]. In the edge cloud collaborative environment, data security and privacy protection is an important issue. It is necessary to adopt corresponding encryption and anonymization technologies to ensure the security of data during transmission and processing [12]. ECCC can be used in many fields of smart factories, such as manufacturing execution system (MES), product quality control, equipment maintenance and so on.

The edge cloud collaboration architecture is designed with a three-tier architecture of "hierarchical awareness-dynamic collaboration", aiming at solving the problems of device heterogeneity and communication bottleneck. The architecture consists of edge layer, collaboration layer and cloud layer (as shown in Figure 1). At the edge layer, a lightweight Resource Agent (RA) is deployed to collect the state information of devices in real time, and model based on the device capability vector to quantify the computing power, memory capacity and real-time level of each device, thus realizing the fine management of resources [13]. The collaboration layer realizes protocol conversion and task routing through Dynamic Service Bus (DSB), and supports the conversion between various protocols such as OPC UA, Modbus and MQTT. At the same time, the service matching function is introduced to evaluate and select the optimal edge node to perform the task, and the transmission delay and resource satisfaction are considered comprehensively. In the cloud layer, the digital twin engine runs

on the long-term data analysis, trains a better scheduling model, and sends the update strategy to the edge layer to realize the continuous optimization and adaptive adjustment of system performance.

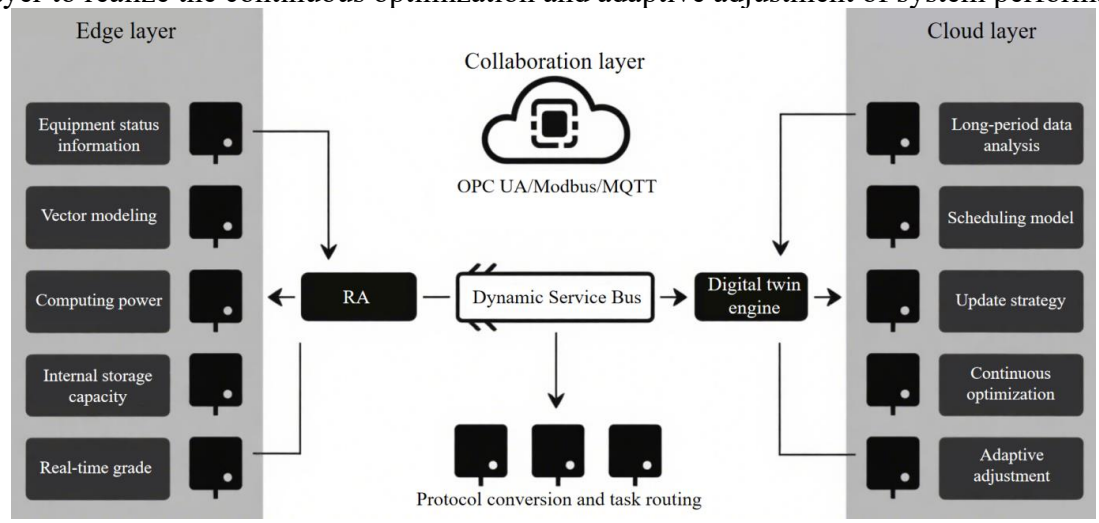

Figure 1 Edge cloud collaborative architecture

3. Automated scheduling strategy

A hybrid scheduling algorithm named "Dynamic Divide and Reinforcement Scheduling" (DDRS) is proposed to achieve low delay and high robustness of scheduling tasks in the edge cloud collaborative environment. The algorithm integrates task decomposition and dynamic decision-making mechanism, and combines reinforcement learning (RL) method for real-time optimization to adapt to complex and changeable industrial scenarios [14].

The core of DDRS lies in the organic combination of "task dynamic divide and conquer" and "feedback RL". In the task partition stage, the system decomposes each task i into multiple subtasks, and determines whether it should be deployed in the edge node or executed in the cloud according to the set decomposition decision function $\Psi(i)$.

$$\Psi(i)=\begin{cases} Edge & if\ \dfrac{Data_i}{BW_{edge}}+\dfrac{Comp_i}{Cap_{edge}} \le T_{critical} \\ Cloud & otherwise \end{cases} \tag{1}$$

The function comprehensively considers the data volume $Data_i$, the network bandwidth BW_{edge}, the computational complexity $Comp_i$ and the computing capacity Cap_{edge} of the edge device, and compares it with the preset critical delay threshold $T_{critical}$, thus making the optimal deployment decision. If the real-time requirement is met, the task is assigned to edge processing; Otherwise, it will be done by the cloud.

On the level of collaborative scheduling, DDRS introduces the feedback RL mechanism and constructs a scheduling model based on the optimization objective function:

$$\min\left(\alpha Delay_{total}+\beta Energy+\gamma Cost_{fail}\right) \tag{2}$$

Among them, α,β,γ is the weight coefficient of different optimization objectives, which represents the degree of attention to total delay, energy consumption and fault punishment respectively. In order to improve the adaptive ability of the scheduling strategy, the system adopts the improved TD3 algorithm and integrates the equipment state prediction module, which can adjust the scheduling strategy in real time according to the running state changes of the production line and enhance the fault tolerance and response speed of the system [15].

4. Optimal allocation method of resources

Construct a multi-dimensional game-collaborative equilibrium resource allocation model to realize the joint optimal allocation of computing, storage and bandwidth resources under the edge cloud collaborative architecture [16]. The model takes into account the dynamic characteristics of edge nodes and cloud resources, and improves the overall resource utilization efficiency and task scheduling performance of the system through the combination of game theory and optimization algorithm.

In the resource modeling stage, the system abstracts the available resources of each edge node into a supply matrix $S_j = \left[CPU_j, Mem_j, BW_j\right]^T$, which respectively represents its CPU computing power, memory capacity and network bandwidth. At the same time, the resource requirements of tasks are described as demand matrix $D_i = \left[CPU_i^{req}, Mem_i^{req}, BW_i^{req}\right]^T$, which is used to quantify the specific consumption of various resources by tasks. Through the matching of supply and demand matrix, the system can realize the fine mapping between tasks and resources.

At the level of resource allocation, the game mechanism based on Nash equilibrium is introduced to construct utility function:

$$U = \sum_{j=1}^{n} \left(\log\left(1 + \eta_j\right) - \lambda Load_j^{std} \right) \tag{3}$$

Among them, η_j represents the resource utilization rate of node j, $Load_j^{std}$ is the standard deviation of its load fluctuation, and λ is the balance factor, which is used to adjust the trade-off between resource utilization efficiency and load stability. The optimal resource allocation matrix $X_{n\times m}$ is solved by the distributed ADMM (Alternating Direction Method of Multipliers) algorithm to ensure that the following constraints are met:

$$\sum_i X_{ji} \circ D_i \le S_j, \quad \forall j \tag{4}$$

Where "$\circ$" stands for Hadamard product, that is, allocation on demand strategy, to ensure that the resource usage of each edge node does not exceed its upper limit.

In order to cope with sudden resource overload, the system also designs an elastic migration mechanism. When it is detected that the ratio of the task queue length of an edge node to its processing capacity exceeds the preset threshold value $\theta_{\max}$, the migration decision is triggered:

$$if \ \frac{Task_{queue}}{Cap_{edg}} > \theta_{\max} \Rightarrow Migrate\_to\_Cloud(\ ) \tag{5}$$

This mechanism can dynamically migrate some non-critical tasks to the cloud for execution when marginal resources are tight, thus ensuring the stable operation of the system and improving the task completion rate. In the process of migration, the task priority and data dependency are combined to make intelligent judgment, which ensures the minimum migration cost and enhances the fault tolerance and adaptability of the system.

5. Experiment and simulation

5.1 Simulation environment configuration

The simulation environment is built on six industrial gateways as edge nodes (4-core CPU/8GB memory/100Mbps bandwidth) and AWS EC2 c5.4xlarge cloud servers (16 core vCPU/32GB memory), using an automatically generated dataset of 1000 manufacturing tasks (covering real-time quality inspection, equipment control, and other scenarios). The baseline includes pure edge scheduling and pure cloud scheduling, with core parameters set as critical delay threshold

$T_{critical} = 150ms$, resource overload threshold $\theta_{\max} = 80\%$, and optimized weight ratio $\alpha : \beta : \gamma$, to comprehensively evaluate the performance of edge cloud collaborative scheduling strategies.

5.2 Experimental results and analysis

The experimental results in Table 1 show that compared with pure edge scheduling and pure cloud scheduling strategy, the edge cloud collaborative scheduling method DDRS proposed in this paper shows significant advantages in task scheduling efficiency. The average delay is reduced from 92ms in pure edge scheduling and 210ms in pure cloud scheduling to 63ms, and the overtime task rate is also greatly reduced, from 18.2% and 41.5% to only 5.3%, which fully verifies the effectiveness of this method in reducing system delay and improving the reliability of task completion.

Table 1 Comparison of task scheduling efficiency

Scheduling strategy	Average delay (ms)	Overtime task rate
Pure edge scheduling	92	18.2%
Pure cloud scheduling	210	41.5%
DDRS	**63**	**5.3%**

Figure 2 The experimental results of resource utilization optimization show that compared with the scheduling mode without resource balancing mechanism, the ADMM optimization model proposed in this study has obvious advantages in reducing the standard deviation of resource utilization, and the standard deviation is reduced from 0.41 to 0.18, indicating that the load of each edge node on CPU, memory and bandwidth is more balanced; At the same time, the number of elastic migration triggers has been reduced from 27 to 9, which effectively reduces the migration overhead caused by uneven resources and improves the stability of the system and the efficiency of resource utilization.

The high robustness verification experiment evaluates the resilience of the system and the adaptability of the scheduling strategy through the fault injection test of sudden downtime at the edge node 2. The test results in Figure 3 show that the system can complete the task redistribution and execution within an average of 120ms after the node failure, which is 3.2 times higher than the baseline method and shows excellent fault tolerance. After detecting the routing failure, DSB in the collaboration layer quickly fed back the fault information to RL module, generated the fault penalty item and updated the scheduling strategy, and finally moved the task to the available node 1 for execution. Although an extra delay of 85ms was introduced, it still achieved rapid recovery, which verified the efficient decision-making ability of TD3 algorithm under dynamic disturbance.

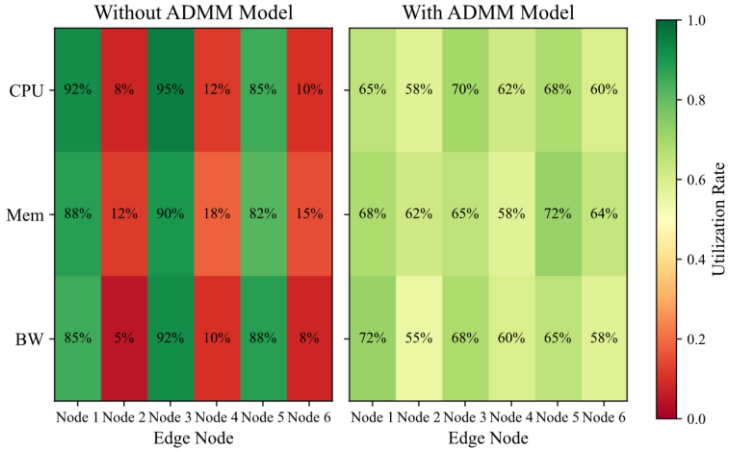

Figure 2 Optimization effect of resource utilization rate

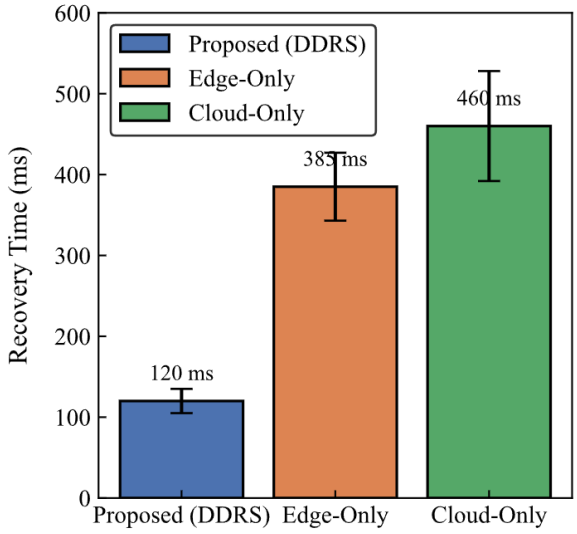

Figure 3 High robustness verification

As a new computing paradigm, ECCC shows great potential in automatic scheduling and optimal resource allocation of smart factories. But there are still many challenges to be overcome. Future research needs to focus on algorithm optimization, system integration, standardization and practical deployment, so as to promote the wide application of ECCC in smart factories.

6. Conclusion

In this study, the technology of collaborative automatic scheduling and optimal resource allocation of edge clouds for smart factories is deeply discussed, and a resource allocation model called DDRS and a multidimensional game-collaborative equilibrium model are proposed. Experiments and simulations show that the scheme has obvious advantages in task scheduling efficiency and resource utilization efficiency. Specifically, compared with pure edge scheduling and pure cloud scheduling strategies, DDRS reduces the average delay from 92ms and 210ms to 63ms, and optimizes the overtime task rate from 18.2% and 41.5% to 5.3%. The ADMM optimization model is also significantly superior to the scheduling mode without resource balancing mechanism in terms of standard deviation of resource utilization, with the standard deviation reduced from 0.41 to 0.18, and the number of flexible migration triggers reduced from 27 to 9, effectively reducing the migration overhead caused by uneven resources. The high robustness verification experiment shows that the system can complete the task redistribution and execution within 120ms even when the edge node suddenly goes down, showing excellent fault-tolerant performance. These results show that the automatic scheduling and resource optimization allocation technology under the edge cloud collaborative architecture can effectively solve the challenges faced by smart factories, such as equipment heterogeneity, task dynamics and massive data processing, and provide an efficient, flexible and landing solution for intelligent upgrading of manufacturing industry. Future research needs to pay more attention to algorithm optimization, system integration, standardization and practical deployment, so as to promote the wide application of edge cloud collaborative computing in smart factories.

References

[1] Jafar Aminu,Rohaya Latip,Zurina Mohd Hanafi,Shafinah Kamarudin & Danlami Gabi.(2025).Efficient Task Allocation for Energy and Execution Time Trade-Off in Edge Computing Using Multi-Objective IPSO.Computers, Materials & Continua,84(2),2989-3011.

[2] XiaopingXiong & GengYang.(2024).A node deployment and resource optimization method for CPDS based on cloud-fog-edge collaboration.IET Generation, Transmission & Distribution,18(21),3524-3537.

[3] Supratik Banerjee & Sanjay Kumar Biswash.(2024).Performance-Driven Resource Allocation Strategy in NDN-Based Mobile Edge Computing (MEC) Networks.Arabian Journal for Science and Engineering,50(10),1-24.

[4] Jianhua Liu,Jincheng Wei,Rongxin Luo,Guilin Yuan,Jiajia Liu & Xiaoguang Tu.(2024).Computation Offloading in Edge Computing for Internet of Vehicles via Game Theory.Computers, Materials & Continua,81(1),1337-1361.

[5] Sampa Sahoo,Kshira Sagar Sahoo,Bibhudatta Sahoo & Amir H. Gandomi.(2024).A learning automata based edge resource allocation approach for IoT-enabled smart cities.Digital Communications and Networks,10(5),1258-1266.

[6] Ruti Gafni,Itzhak Aviv & Dror Haim.(2024).Multi-Party Secured Collaboration Architecture from Cloud to Edge.The Journal of Computer Information Systems,64(5),698-709.

[7] Neelakantan Puligundla,Gangappa Malige,Rajasekar Mummalaneni,Sunil Kumar Talluri & Suresh Reddy Gali.(2024).Resource allocation for content distribution in IoT edge cloud computing environments using deep reinforcement learning.Journal of High Speed Networks,30(3),409-426.

[8] LinZhu,LongTan,BingxianLi & HuiziTian.(2024).An optimization scheme for vehicular edge computing based on Lyapunov function and deep reinforcement learning.IET Communications,18(15),908-924.

[9] Duan Ying & Jiang Chunmao.(2024).Binary task offloading strategy for cloud robots using improved game theory in cloud-edge collaboration.The Journal of Supercomputing,80(10),14752-14772.

[10] Wang Yu,Zhang Zhiyi,Tang Peng & Bian Shiyao.(2024).A Model for Predicting Physical Health of College Students Based on Semantic Web and Deep Learning Under Cloud Edge Collaborative Architecture.International Journal on Semantic Web and Information Systems (IJSWIS),20(1),1-19.

[11] Wang Wei,Wang Xiaotian,Ma Xiaotian,Zhao Ruifeng & Yang Heng.(2024).Residential Electricity Consumption Prediction Method Based on Deep Learning and Federated Learning Under Cloud Edge Collaboration Architecture.International Journal of Gaming and Computer-Mediated Simulations (IJGCMS),16(1),1-19.

[12] Guo Wei,Sun Shengbo,Tao Peng,Li Fei,Ding Jianyong & Li Hongbo.(2024).A Deep Learning-Based Microgrid Energy Management Method Under the Internet of Things Architecture.International Journal of Gaming and Computer-Mediated Simulations (IJGCMS),16(1),1-19.

[13] Liu Xi & Liu Jun.(2023).A truthful mechanism for multi-access multi-server multi-task resource allocation in mobile edge computing.Peer-to-Peer Networking and Applications,17(1),532-548.

[14] RenbinFang,PengLin,YizeLiu & YanLiu.(2023).Task offloading and resource allocation for blockchain-enabled mobile edge computing.IET Communications,18(20),1889-1899.

[15] Liu Shuang,Tian Jie,Zhai Chao & Li Tiantian.(2023).Joint computation offloading and resource allocation in vehicular edge computing networks.Digital Communications and Networks,9(6),1399-1410.

[16] Mani A.,Kavya G. & Bapu B. R. Tapas.(2023).A proficient resource allocation using hybrid optimization algorithm for massive internet of health things devices contemplating privacy fortification in cloud edge computing environment.Wireless Networks,30(3),1187-1199.