ISSN: 2710-0588

Research on Worker Allocation Optimization Based on Real-time Data in Cloud Computing

Jingtian Zhang

Georgia Institute of Technology, Atlanta 30332, Georgia, USA

Abstract

The core of cloud computing task scheduling optimization for real-time data is to monitor the workload, task requirements, and resource utilization status of the real-time monitoring system, and then flexibly adjust the task allocation plan to improve resource utilization efficiency and accelerate response time. This article deeply analyzes the application of multi-objective optimization technology, optimization of data transmission and processing flow, changes in scheduling strategies based on real-time data, and the improvement of intelligent and adaptive task assignment capabilities, and constructs a comprehensive optimization architecture. By adopting these methods, the performance of cloud computing platforms can be significantly enhanced, task processing latency can be reduced, and scientific allocation of resources can be achieved.

Keywords

Cloud computing; Worker allocation; Real time data; Multi objective optimization.

1. Introduction

The traditional cloud computing task allocation model is mostly based on static scheduling, which often ignores the use of real-time data, resulting in uneven resource allocation and slow task processing. The optimization of worker allocation strategies based on real-time data can dynamically track system load, task requirements, and resource status. Combined with multi-objective optimization techniques, real-time adjustments can be made to task allocation strategies to improve system operational efficiency and effective resource utilization. The focus of this study is to explore how to use real-time data-driven dynamic scheduling mechanisms, optimize data transmission and processing processes, enhance the intelligence and adaptability of task scheduling, and other strategies to achieve efficient resource allocation optimization for cloud computing workers, providing theoretical guidance and practical technical support for the efficient operation of cloud computing platforms.

2. Overview of Cloud Computing Worker Allocation Based on Real-time Data

2.1. Definition of Cloud Computing Worker Allocation

Cloud computing worker allocation refers to the process of carefully arranging and optimizing the allocation of computing resources in a cloud computing environment, ensuring that tasks can run smoothly on appropriate computing units, thereby improving task processing efficiency. The so-called worker refers to a node or virtual machine that serves as a resource for computing, storage, networking, etc. in a cloud environment. In the multi tenant environment of cloud computing, the workload of tasks must be flexibly adjusted and optimized based on variables such as worker processing capacity, resource utilization status, and networking status[1]. In this environment, the allocation purpose of workers is to fully tap into the potential of computing resources, ensure real-time and accurate task execution, and meet

ISSN: 2710-0588

constantly changing business demands and system load pressures. A reasonable worker allocation mechanism can improve the overall performance of cloud computing systems, reduce resource consumption, and enhance user experience. Especially in large-scale distributed computing scenarios, effective scheduling of workers plays a decisive role in improving the operational efficiency of cloud systems.

2.2. The Importance of Cloud Computing Worker Allocation Based on Real time Data

In conventional cloud computing environments, resource allocation strategies often rely on fixed resource configurations or past data records, but often overlook the role of real-time dynamic factors. However, in the specific operation process, the system's load, network conditions, fault recovery, and other situations may change at any time. If only a fixed mode is used to schedule resources, it often leads to uneven task distribution, slow response, or excessive resource consumption. The task allocation mechanism based on real-time information can track multiple dynamic indicators in the cloud computing environment in real time, such as resource utilization status, task progress, and fluctuations in system load[2]. By relying on real-time data, task allocation can quickly adapt to changes in load, thus avoiding the limitations of relying solely on historical models or static settings. Faced with massive data processing demands, this real-time data based task allocation strategy can reduce the complexity of resource scheduling, accelerate system response efficiency, and ensure the quality and reliability of cloud computing services. This strategy not only optimizes the utilization efficiency of computing resources, but also provides more flexible processing strategies in the changing cloud computing scenarios, meeting the high requirements for resource scheduling accuracy and real-time performance.

3. Current Status of Cloud Computing Worker Allocation

3.1. Unbalanced task load

In cloud computing environments, the phenomenon of uneven workload distribution refers to significant inconsistencies in task allocation among numerous processing nodes, causing some nodes to bear greater pressure while the resources of other nodes are not fully utilized. This imbalance usually stems from inappropriate task allocation strategies or the system's failure to accurately assess the computational requirements of each task when allocating resources. Uneven workload distribution is mainly reflected in some nodes operating under overload, resulting in low efficiency of computing resource utilization, and in extreme cases, may cause node failures or response delays[3]. At the same time, some nodes may be in a low load or even idle state, and their resources may not be effectively utilized. Long term uneven workload may prolong the execution time of tasks, and even lead to long-term backlog of some tasks, thereby affecting the response efficiency and operational stability of the entire system.

3.2. Real time data latency

In cloud computing architecture, the so-called real-time data delay refers to the time delay that occurs during the transmission and processing of data, which makes task allocation and resource allocation unable to keep up with the real-time situation of the system. This type of delay is commonly present in multiple stages of data collection, transmission, processing, and feedback, especially in large distributed systems, where this issue is particularly evident. If real-time data processing fails to keep up with the speed, cloud computing platforms cannot quickly grasp the current workload, resource utilization status, or system operation status, which in turn affects the appropriate allocation of work nodes[4]. Delay may also lead to delayed response of the system to load fluctuations, increasing the risk of task timeouts or resource allocation errors. Therefore, real-time data latency not only reduces the system's response

ISSN: 2710-0588

efficiency, but also weakens the cloud computing system's ability to adjust in a constantly changing environment.

3.3. Unstable data quality

In cloud computing platforms, the unstable quality of data is manifested as the possibility of interference, information omission, or insufficient accuracy of the data sources relied upon during task allocation and scheduling. Given that cloud computing systems must gather massive amounts of information from various channels, such as resource utilization status, task execution status, system burden, and other factors, any fluctuation in data quality may have a direct impact on the effectiveness of worker allocation. Data quality issues often manifest as incomplete or distorted information, which may be caused by sensor damage, communication link interruptions, errors during data transmission, or system component failures. The instability of this data can cause misjudgments in task allocation algorithms, thereby interfering with the efficiency of resource allocation and task execution. Fluctuations in data quality may also lead to incorrect assessments of task importance, which can have adverse effects on the system's load forecasting and scheduling strategies.

3.4. Complex Task Scheduling and Resource Management

With the continuous expansion of cloud computing platforms, the types of resources, task categories, and their interdependencies involved are becoming increasingly complex. Task assignment is no longer solely based on resource availability for worker allocation, but requires comprehensive consideration of numerous factors, such as the urgency of the task, the actual supply of resources, the computational requirements of the task, and the constraints of network transmission bandwidth[5]. The complexity and difficulty of resource management are manifested in numerous aspects, with each type of resource possessing its unique attributes and mutually constraining and competing with each other. The real-time changes in system workload require task assignment strategies to quickly adapt and adjust, which requires precise selection under numerous constraints. Many cloud computing tasks are limited by strict completion times, which forces dispatch algorithms to have high real-time and fast response capabilities. Complex task dispatch algorithms often mean that more computing resources and time are needed to complete the decision-making process.

4. Optimization Strategy for Cloud Computing Worker Allocation Based on Real-time Data

4.1. Application of Multi Objective Optimization Algorithm

Multi objective optimization aims to find an equilibrium among multiple objectives in order to obtain a set of "Pareto optimal solutions", that is, to avoid adverse effects on other objectives as much as possible while improving one objective. By using such strategies, a coordination can be achieved between various goals, thereby improving the overall operational efficiency of the system[6]. Multi objective optimization algorithms typically use evolutionary computation methods such as genetic algorithm, particle swarm optimization (PSO) algorithm, ant colony algorithm, etc. These algorithms draw on the principles of natural evolution and rely on iterative search to seek the optimal or near optimal solution. Taking genetic algorithms as an example, by mimicking processes such as natural selection, gene recombination, and mutation, the optimal resource allocation plan is sought; The particle swarm optimization strategy explores the ideal solution to multi-objective optimization problems by simulating the dynamics of particle populations. Before applying these strategies, it is necessary to conduct quantitative analysis and model construction on each objective to ensure that the optimization path and constraints of each objective are clear. The optimization process is shown in the multi-objective optimization flowchart in Figure 1.

ISSN: 2710-0588

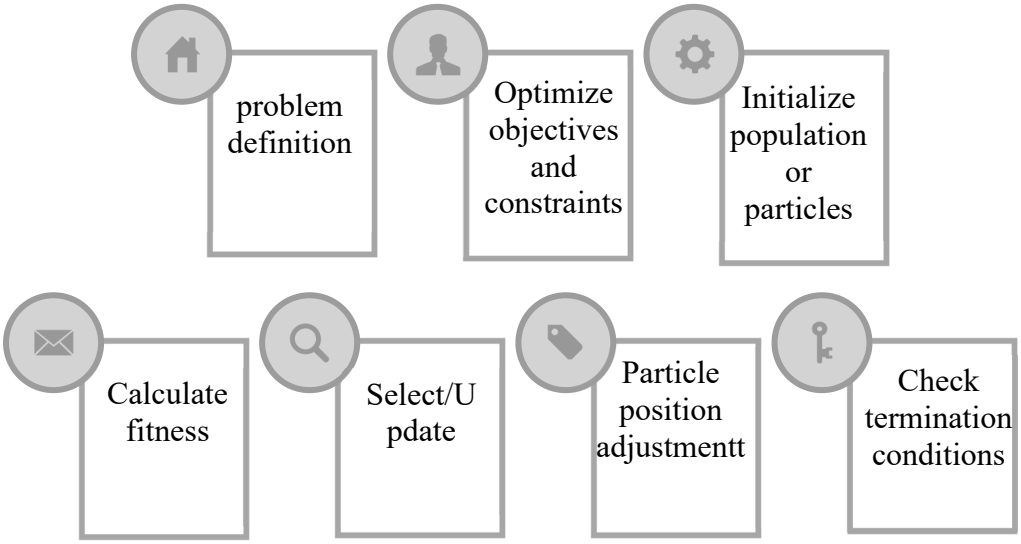

Figure 1. Multi objective optimization flowchart

In specific applications, a scheduling model for task allocation is first constructed, which converts the time nodes, resource consumption ratios, and energy efficiency indicators of task execution into mathematical formulas. Optimization algorithms will comprehensively evaluate multiple resource allocation strategies under given constraints[7]. By continuously adjusting algorithm parameters and optimizing the weights of objectives, the most suitable worker allocation scheme for the current cloud computing environment is ultimately obtained.Multi objective optimization technology can also adjust the importance level of tasks in real time, ensuring that core tasks are prioritized for execution, thereby enhancing the robustness and rapid response characteristics of cloud computing platforms. Through multi-objective optimization algorithms, the optimal solution can be found in complex environments with multiple objectives and constraints, optimizing the efficiency of cloud computing resource allocation, ensuring the stability of system operation, and significantly improving the efficiency of resource allocation.

4.2. Improve data transmission and processing flow

In the allocation of cloud computing workers based on real-time data, the optimization of data transmission and processing flow is crucial, as it directly affects the system response speed, task scheduling accuracy, and resource utilization efficiency. The key is to introduce advanced data compression and deduplication methods. Compression technology can effectively reduce data volume, alleviate bandwidth pressure, and accelerate transmission rate; The deduplication technology can identify and eliminate duplicate information, avoid ineffective transmission, and further enhance the efficiency of system operation[8]. Secondly, updating the protocol for data transmission is also necessary. Choosing new protocols such as HTTP/2 or QUIC can reduce transmission latency, optimize multiplexing and header compression, reduce latency caused by frequent requests, and enhance data transfer efficiency in cloud computing. Thirdly, improving data processing speed cannot be ignored. Using cutting-edge technologies such as the Internet of Things and edge computing, data processing tasks will be migrated to the edge of the network to achieve local pre-processing of data, and only core data will be uploaded to the cloud, effectively reducing the pressure on the cloud computing platform and accelerating processing speed. At the application level, edge computing can preliminarily screen, clean and summarize the data, and then send the refined data to the cloud for in-depth analysis. The optimization of data storage and retrieval is equally important. By utilizing distributed storage architecture and intelligent management strategies, large-scale datasets

ISSN: 2710-0588

can be efficiently processed and accessed, avoiding processing delays caused by storage bottlenecks. Distributed storage disperses data to numerous nodes, reducing the pressure on individual nodes and improving the speed of data retrieval and system stability.

Table 1. Optimization Measures for Data Transmission and Processing Flow

optimization measures	target	Specific techniques and methods	Optimization effect
Data compression and deduplication	Reduce data volume and minimize redundancy	Data compression and deduplication technology	Reduce bandwidth requirements and improve transmission efficiency
Optimization of Data Transmission Protocol	Improve transmission speed and reduce latency	HTTP/2, QUIC protocol, header compression, multiplexing	Reduce transmission latency, improve bandwidth utilization, and optimize data flow
Edge computing and Data Preprocessing	Reduce cloud load and improve processing efficiency	Delegate data processing tasks to edge nodes for preprocessing	Reduce cloud burden, improve data processing speed and efficiency
Distributed Storage and Intelligent Management	Optimize storage and access speed	Distributed storage system, intelligent data management strategy	Improve data storage access speed and enhance system stability

From Table 1, we can clearly see how different optimization technologies can work together from data compression and de duplication, protocol optimization, edge computing and distributed storage, so as to improve the efficiency of data transmission and processing on the whole.

4.3. Real time data-driven dynamic adjustment

The dynamic adjustment mechanism driven by real-time data relies on high efficiency in data collection, processing, and analysis. During the operation of the system, the running status of each work node is continuously tracked and recorded, involving multiple dimensions of information such as task completion, workload, and network bandwidth usage. Information collection relies on various sensors, monitoring software, and network communication protocols, and is quickly transmitted to the cloud data processing center. In the data processing center, conduct in-depth analysis of the collected information and optimize resource allocation plans based on real-time system operation. If it is found that a certain work node is overloaded or response delay is exacerbated, the system relies on real-time data feedback to quickly locate the source of the problem and flexibly adjust task allocation strategies. The tasks undertaken by performers with higher workloads can be reassigned to nodes with lower workloads to ensure smooth completion of various tasks within the specified time. This flexible task allocation mechanism effectively prevents delay issues caused by overload or failure of individual execution nodes. The system has the ability to flexibly adjust task allocation strategies based on real-time information changes. For example, during task execution, the system may receive updated task priority information, which provides real-time adjustment basis for task allocation, enabling the system to change the execution order of tasks in a timely manner and prioritize important tasks. As shown in Table 2, there is a close relationship between real-time information and adjustment strategies. The system implements corresponding dynamic adjustment strategies based on the different categories of real-time

ISSN: 2710-0588

information, thereby improving the efficiency of task allocation and resource allocation, and ensuring that the cloud computing platform operates efficiently under various workloads and environmental conditions.

Table 2. Real time Data and Adjustment Strategies

Real time data items	data type	adjustment
Task Progress	Proportion or time	Dynamically adjust priority based on task progress
Workload situation	CPU and memory utilization rate	Transfer tasks to workers with lighter workloads
Network bandwidth usage	Bandwidth Utilization	Optimize network bandwidth allocation to avoid network bottlenecks
network delay	milliseconds	Improve data processing speed and efficiency
Task priority change	numerical value	Adjust the task scheduling order based on real-time priority,

4.4. Enhance the intelligence and adaptability of task scheduling

The efficient intelligent scheduling system in cloud computing environment is committed to optimizing resource allocation and task processing efficiency. The system integrates five core components: data collection, task prediction, scheduling decision-making, resource management, and feedback adjustment, as shown in Figure 2.

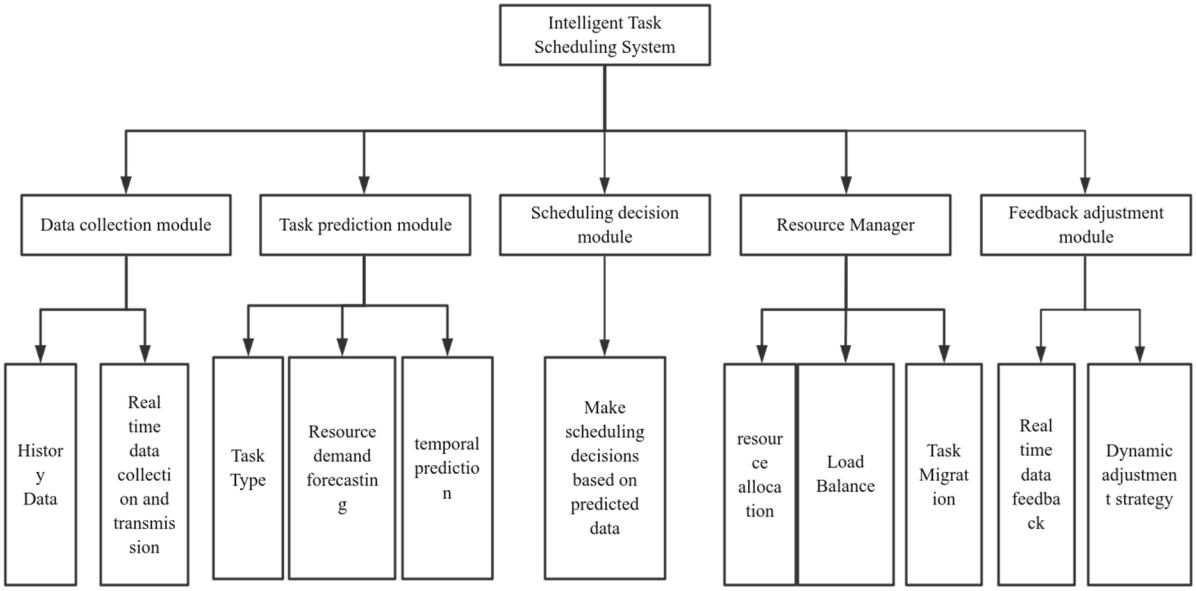

Figure 2. Architecture diagram of intelligent task scheduling system

When the system is started, the data collection module is responsible for real-time tracking of node work status, collecting information on task progress, resource utilization, and network latency, and transmitting this information to the cloud through monitoring devices. The task prediction module will conduct in-depth analysis of accumulated data and use machine learning techniques to predict the computational complexity and completion time of tasks, in order to ensure the accuracy and efficiency of task allocation. Based on the estimated results,

ISSN: 2710-0588

the scheduling decision module will automatically plan the task allocation scheme and optimize the allocation according to the urgency and resource requirements of the tasks. The resource management module is responsible for dynamically adjusting resources on the cloud platform, ensuring that workloads are transferred from busy nodes to idle nodes, and preventing task delays caused by resource shortages. The feedback adjustment module continuously monitors task execution and adjusts scheduling strategies based on actual situations. Through this series of scheduling methods, the system can adapt in real-time to changes in workload, task priority, and resource requirements, thereby improving the processing efficiency and response speed of cloud computing systems.

5. Conclusion

In the research of cloud computing worker allocation optimization based on real-time data, multi-objective optimization algorithms, improved data transmission and processing flow, real-time data-driven dynamic adjustment, and enhanced intelligence and adaptability of task scheduling are adopted to effectively improve resource utilization efficiency and task response speed. With the continuous development of technology, cloud computing environments have become more complex, and traditional static scheduling strategies can no longer meet the current requirements for efficient and flexible resource management. Therefore, by flexibly adjusting and intelligently upgrading real-time data, challenges such as uneven system load distribution, lagging data processing, and excessive resource consumption can be efficiently addressed. In the future, with the integration and application of high-end technology, resource allocation optimization in the field of cloud computing will move towards a higher level of intelligence and automation, laying a more solid foundation for the stable operation, scalability, and operational efficiency of cloud computing.

References

[1] Alanagh A Y ,Firouzi M ,Kenari R A , et al.Introducing an adaptive model for auto-scaling cloud computing based on workload classification[J].Concurrency and Computation: Practice and Experience, 2023,35(22):

[2] Min C ,Yaoyu L ,Xupeng W , et al.Energy-aware intelligent scheduling for deadline-constrained workflows in sustainable cloud computing[J].Egyptian Informatics Journal,2023,24(2):277-290.

[3] Youssef S ,Soufiane J ,Said K E , et al.Reducing energy footprint in cloud computing: a study on the impact of clustering techniques and scheduling algorithms for scientific workflows [J]. Computing, 2023,105(10):2231-2261.

[4] Youssef S ,Soufiane J ,Said K E , et al.Reducing energy footprint in cloud computing: a study on the impact of clustering techniques and scheduling algorithms for scientific workflows [J]. Computing, 2023, 105(10):2231-2261.

[5] Buyya R ,Ilager S ,Arroba P .Energy-efficiency and sustainability in new generation cloud computing: A vision and directions for integrated management of data centre resources and workloads [J]. Software: Practice and Experience,2023,54(1):24-38.

[6] Ahmad Z ,Acarer T ,Kim W .Optimization of Maritime Communication Workflow Execution with a Task-Oriented Scheduling Framework in Cloud Computing[J].Journal of Marine Science and Engineering,2023,11(11):

[7] Mirsaeid S H .A survey study on task scheduling schemes for workflow executions in cloud computing environment: classification and challenges[J].The Journal of Supercomputing, 2023, 80(7): 9384-9437.

[8] VermaP ,MauryaK A ,YadavS R .A survey on energy-efficient workflow scheduling algorithms in cloud computing[J].Software: Practice and Experience,2023,54(5):637-682.