GEORGE BROWN PRESS
JOURNALS, BOOKS & PROCEEDINGS

Economics and Management Innovation

Article

Balance Model of Resource Management and Customer Service Availability in Cloud Computing Platform

Jiaying Huang [1,*]

[1] EC2 Core Platform, Amazon.com Services LLC, Seattle, WA, 98121, United States
* Correspondence: Jiaying Huang, EC2 Core Platform, Amazon.com Services LLC, Seattle, WA, 98121, United States

Abstract: In response to the real-time requirements of resource allocation and service availability dynamic balance in cloud service scheduling, this paper constructs an integrated Markov Decision Process (MDP) scheduling optimization model, comprehensively analyzes the impact of scheduling delay, load balancing efficiency, and virtual machine migration cost on service quality (SLA completion rate and average response time). At the same time, the XGBoost algorithm is used to further mine previous business data and improve the accuracy of business service availability prediction. The model was tested on scheduling and monitoring data for the fourth quarter of operation on a large public cloud platform. The results showed that within a 95% confidence interval, the prediction accuracy reached 93.7% and the resource utilization rate increased by 17.4%. The performance of this scheduling method was evaluated and validated in high concurrency scenarios, and it was found that it can effectively improve the system's availability and scheduling efficiency. This study provides theoretical and engineering practical basis for implementing an intelligent resource scheduling mechanism based on learning and modeling integration.

Keywords: cloud computing; resource management; service availability

1. Introduction

With the rapid development of cloud computing platforms, the imbalance between resource operation and customer service availability is becoming increasingly prominent. The scheduling system needs to meet high-frequency and high load working conditions, while also considering service level and resource utilization. Traditional scheduling models cannot provide optimal scheduling solutions for real-time scheduling and dynamic load changes, and there is an urgent need for models with intelligent prediction to optimize resource allocation schemes, in order to achieve dual improvement in platform performance and user experience.

2. The Theoretical Basis of the Balance between Cloud Computing Platform Resource Management and Customer Service Availability

Cloud computing platforms should achieve a dynamic balance between high efficiency and high availability services within limited resources. To establish a reasonable model, it is necessary to use MDP to model the resource status and behavior development trajectory, and study the task queuing situation and the impact of resource occupancy on service quality through queuing theory. This theoretical model can quantify the effectiveness of scheduling strategies in meeting SLA success rates and average response times [1]. Simultaneously introducing reinforcement learning mechanisms to enable the system to have adaptability and feedback capabilities, in order to achieve effective resource allocation and better service assurance in high concurrency modes.

Received: 21 June 2025
Revised: 03 July 2025
Accepted: 22 July 2025
Published: 28 July 2025

Copyright: © 2025 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

3. Resource Management and Customer Service Availability Analysis in Cloud Computing Platforms

3.1. Key Factors and Obstacles in Resource Management

Resource management is the process of unified and dynamic management of computing, storage, and network resources on cloud computing platforms. This mainly includes resource utilization, task priority, SLA constraints, and scheduling cycles. The scheduling program needs to update the resource status in real time and respond as soon as possible to improve the platform's operational efficiency. However, in actual operation, there are practical problems such as incomplete resources, variable task loads, uneven task loads, and high virtual machine migration costs, which lead to the coexistence of resource idle and service latency, thereby affecting the overall service quality [2]. It is urgent to introduce a predictable and adjustable dynamic management mechanism to address these issues.

3.2. Load Balancing and Scheduling Mechanisms in Cloud Computing Platforms

Ensuring the reliable operation and effective service of cloud service platforms mainly relies on load balancing, with the goal of allocating computing tasks reasonably among multiple nodes to avoid resource competition and the probability of service congestion. There are currently several main allocation methods: static allocation (such as polling allocation or maintaining a minimum number of connections) and dynamic allocation (based on occupied resources and expected load allocation). The former is not flexible enough, while the latter requires a large amount of computation and generates long delay times. To improve real-time scheduling, it is necessary to combine simplified prediction with timely response, complete task priority determination, timely monitoring of resources, and intelligent allocation of strategies. A reasonable scheduling strategy can play a significant role in improving task execution efficiency, response time, and task loss rate.

3.3. Customer Service Availability Assessment and Its Influencing Factors

Customer service availability is an important indicator reflecting the quality of customer service on cloud platforms, mainly including SLA completion, average response time, task completion, etc. The determining factors that affect service availability include resource allocation speed, task accuracy, stability of virtual machine migration, and the system's tolerance to sudden loads. The coverage and response time of monitoring devices directly affect the efficiency of obtaining and scheduling availability data. If the prediction mechanism is too outdated or the scheduling method is too rigid, it is easy to cause service interruption or response expiration, resulting in a decrease in customer satisfaction [3]. Therefore, constructing a precise and controllable availability model is the key prerequisite for enhancing the collaborative optimization of resource management and service assurance.

3.4. Balance Strategy between Resource Allocation and Service Quality

In the cloud computing platform, the contradiction between resource allocation and service quality runs through the entire scheduling life cycle. To achieve an effective balance between the two, optimization needs to be carried out from three dimensions: intelligence, dynamics and differentiation of the resource allocation strategy. On the one hand, the platform should classify tasks based on the type of business load (such as real-time computing, big data analysis, IO-intensive, etc.), give priority to ensuring the computing resource occupation of core tasks, and ensure that high-priority services have stable performance responses; On the other hand, during periods of resource shortage, the task scheduling sequence should be flexibly adjusted in combination with the task completion deadlines, historical behavior patterns and user credit systems to avoid SLA defaults caused by resource contention.

To enhance system reliability, multi-layer redundancy mechanisms can be deployed. For instance, backup computing units can be configured for critical nodes, or the "hot standby + cold migration" mechanism can be utilized to ensure the rapid recovery of faulty nodes. Meanwhile, predictive fault perception and fault-tolerant mechanisms are introduced. Through real-time monitoring of the operational status of resources and the accumulation trend of tasks, potential bottlenecks are identified in advance to achieve preventive migration and load diversion.

At the scheduling strategy layer, combined with the weight mechanism, targeted optimization is carried out for tasks of different SLA levels: higher resource guarantee weights are set for high-level services, and resource occupation is appropriately compressed for tasks with tolerable delays. The platform can also integrate machine learning models such as XGBoost to predict high-load Windows and performance bottlenecks, achieving pre-allocation of resources. This ensures resource utilization while optimizing service quality and customer experience, ultimately achieving a win-win situation of resource efficiency and service reliability.

4. Balance Model between Resource Management and Customer Service Availability

4.1. Design Concept of Balance Model

The concurrent service requests from multiple tenants test the resource supply and service availability of cloud computing platforms. Traditional scheduling strategies overly emphasize resource utilization while neglecting service quality, resulting in task delays or violations of SLA requirements. In order to solve this contradiction, this article constructs a framework theoretical model that balances "benefits costs" in order to obtain a balanced resource allocation strategy from resource allocation and service availability. The design concept of "state discrimination decision making real-time adjustment" is adopted to integrate resource status, task load, and SLA requirements together. The scheduling behavior is optimized with the following utility function as the objective:

$$F = \lambda_1 \cdot SLA + \lambda_2 \cdot Util - \lambda_3 \cdot Cost \quad (1)$$

Among them, SLA represents the service completion rate, $Util$ is the resource utilization rate, $Cost$ is the scheduling cost, and λ_1 is the weight coefficient. The design concept is to combine service quality with platform efficiency, and to simultaneously consider both service quality and platform efficiency in scheduling algorithms, creating a dynamically adjustable unified scheduling model and achieving optimal resource scheduling and service quality assurance.

4.2. Interaction between Resource Management and Service Availability

The feedback relationship between resource management and service capabilities in cloud computing platforms is closely related: resource scheduling directly affects the SLA satisfaction rate and service latency of services, and the dynamic situation of services in turn affects the scheduling priority of resource scheduling modes. When the system load increases, the reliability rapidly decreases, and the platform will trigger a high optimization task guarantee mechanism; However, stable resource management systems tend to release excess resources to maintain overall utilization. From Table 1, it can be analyzed that there is a close correlation between the core performance status of the service platform and the overall service availability changes [4]. Therefore, based on this relationship, the system can freely regulate the mode of resource allocation to achieve self-tuning functionality.

Table 1. Analysis of the Interactive Characteristics between Resource Status and Service Availability.

Resource status characteristics	Changes in usability performance	Management response mechanism
CPU utilization is too high	Response delay increases	Initiate task diversion and resource reallocation mechanism
The task queue continues to accumulate	SLA completion rate decreases	Enhance the scheduling weight of high priority tasks
High resource idle rate	SLA maintains stability	Dynamically release resources to a low priority task pool

The coordination mechanism provides the basis for scheduling variables and objective functions in the balance model from a practical point of view.

4.3. Establishing Mathematical Models and Algorithms for Equilibrium Models

This article uses a mathematical model based on state behavior return to systematically describe the mapping relationship between resource management and availability, with MDP as the theoretical basis. In this model, system state S represents the current resource load, service request volume, and task queue situation; Behavior A represents specific scheduling strategies, including resource allocation and task migration operations. Minimize resource utilization and task latency while ensuring sufficient service availability. The optimization objective function is defined as:

$$\pi^* = arg\max_{\pi} E[\sum_{t=0}^{\infty} \gamma^t R(S_t, A_t)] \quad (2)$$

Among them, π is the scheduling strategy, γ is the discount factor, and $R(S_t, A_t)$ is the immediate reward value for taking action A_t in state S_t. Based on a comprehensive consideration of factors such as SLA completion efficiency, resource efficiency, and migration costs, the use of policy iteration algorithms can achieve optimal scheduling route migration through continuous state updates, achieve optimal allocation of platform resources and services, and maintain the best balance between platform resources and services in high concurrency states.

4.4. Feasibility Analysis and Implementation Path of the Balanced Model

To verify the effectiveness of the resource allocation and service performance balance model, this article analyzes it from four aspects: system architecture adaptability, computational cost, policy response efficiency, and platform compatibility [5]. The model is based on the MDP process and can be directly integrated into existing mainstream cloud platform scheduling engine systems. The platform can use the measured information obtained from the monitoring system to automatically extract state information features and implement real-time response using pre training methods, thereby greatly reducing deployment and iteration costs.

The implementation path mainly includes five stages: state data collection, state space modeling, strategy training, policy deployment, and system feedback optimization. At each stage, a closed loop is formed through independent and collaborative optimization of data and strategies. As shown in Figure 1, the complete running process of the model is presented.

Figure 1. Implementation flowchart of resource management and service availability balance model.

This model has excellent system scalability and robustness, and can establish a robust and fast coordination management system for large cloud computing platforms, fully demonstrating its application effectiveness and system stability under high loads.

5. Model Optimization and Practical Application

5.1. Improving the Accuracy of Service Availability Prediction through Machine Learning

In order to optimize resources in advance and avoid risks, this article uses the XGBoost algorithm to construct a service availability prediction model, which makes predictions before scheduling resources. Select characteristic attributes such as CPU usage, memory occupancy, workload, and request response time, and use time series data obtained through sliding windows as input to predict the completion rate of SLA in the future. XGBoost is based on gradient boosting decision tree, and the optimization objective function is:

$$L = \sum_{i=1}^{n} l(yi, \hat{y}i) + \sum_{k=1}^{K} \Omega(f_k) \quad (3)$$

Among them, l is the prediction error term, and $\Omega(f_k)$ is the model complexity regularization term. During model training, 80% data fitting and 20% validation were used, resulting in a prediction accuracy of 93.7% at a 95% confidence interval. From the perspective of important features, load change rate and response slope are the most important. In contrast, compared to traditional threshold determination methods, the XGBoost method can better distinguish the nonlinear trend of the load, thereby improving the active response and prediction accuracy of the scheduling system.

5.2. Case Study: Analysis of Application Effectiveness of a Cloud Computing Platform

To further verify the practical effectiveness of the resource management and customer service availability balance model, this paper selects a leading domestic public cloud computing platform as the research object and focuses on its production environment operation data in the fourth quarter of 2024 for empirical analysis. The platform has an average of over 80,000 active virtual machines per day, serving users in multiple key industry scenarios such as finance, e-commerce, manufacturing, and government affairs. It features large task volumes, high concurrent access, and complex resource states, effectively demonstrating the adaptability and universality of the model in complex business environments.

The data adopted in this experiment is derived from the historical records of the platform's scheduling center and performance monitoring system, covering approximately 5 million pieces of task scheduling and resource status data. The indicators include but are not limited to: CPU/ memory / I/O utilization rate, task queue length, average response time, SLA completion rate, virtual machine migration frequency, etc. The data has undergone anonymous desensitization processing to ensure security and representativeness. Two system environments were deployed in the experiment: the traditional scheduling mode and the MDP + XGBoost joint optimization scheduling model proposed in this paper. Each was run for two weeks as the comparison window period to ensure the stability and comparability of the experimental data.

During the model deployment process, the historical data is first modeled by sliding Windows, and the service availability prediction model is constructed through XGBoost. This model is mainly characterized by CPU utilization, load growth rate, request response slope, etc., to predict the SLA compliance rate within the next 5 minutes and provide pre-evaluation information for the scheduling system. Subsequently, the scheduling engine invokes the dynamic resource allocation model based on Markov Decision Process (MDP) in real time according to the prediction results to achieve task priority determination, dynamic resource reallocation and virtual machine migration optimization.

The deployment results show that the overall SLA completion rate of the platform has increased from 85.1% before deployment to 93.7%, the average response time has decreased from 420ms to 323ms, and the task backlog rate has decreased by 19.6%. The resource utilization rate increased from 52.6% to 70.0%, with an increase rate of 17.4%. The frequency of virtual machine migration has decreased from 15 times per day to 9 times per day, a reduction of 40%, significantly lowering the costs of system oscillation and data migration. The detailed comparison is shown in the following table 2:

Table 2. Comparison of Platform Core Performance Indicators Before and After Model Deployment.

Indicator Name	Pre deployment values	Value after deployment	Change amplitude
SLA completion rate	85.1%	93.7%	↑ 8.6%
Resource utilization rate	52.6%	70.0%	↑ 17.4%
Average response time	420ms	323ms	↓ 23.1%
Virtual machine migration frequency	15 times/day	9 times/day	↓ 40.0%

From the perspective of scheduling behavior, the model effectively curbs the resource contention caused by task accumulation and optimizes the task scheduling sequence. The sensitive identification of sudden high loads by the prediction mechanism has also significantly improved the scheduling response speed of high-priority tasks. When resources are tight, the scheduling system actively compresses the delay of low-priority tasks, releases resources for core tasks, and ensures the continuity of critical services. In addition, the feedback from the satisfaction survey on the user side also confirmed the practicality of the model: the average feedback delay complaint rate of platform customers decreased by approximately 27.5%, and the satisfaction rate of work order responses increased by 12.3%.

To sum up, the resource scheduling balance model based on the integration of MDP and XGBoost proposed in this paper has good cross-platform portability and generalization ability. When facing the requirements of multi-tenant, high concurrency and high complexity resource scheduling, this model can effectively achieve the dynamic optimal allocation of resources and the stable guarantee of service availability, providing important theoretical and engineering references for the subsequent construction of an intelligent adaptive scheduling system.

6. Conclusion

This paper aims at the core problems of low resource allocation efficiency and insufficient service availability faced by cloud computing platforms in a high-concurrency environment, and proposes a resource management and scheduling optimization strategy that integrates the Markov Decision Process (MDP) and the XGBoost prediction model. By introducing state assessment, task priority judgment and intelligent prediction mechanisms, a scheduling model with real-time performance, adaptability and controllability characteristics has been constructed. Empirical analysis shows that the deployment of this model in the actual platform not only significantly improves the SLA completion rate and resource utilization rate, reduces response delay and migration cost, but also effectively

enhances the stability of the system and the continuity of customer service. Compared with the traditional static strategies, the method proposed in this paper shows stronger intelligence and robustness under dynamic load conditions, providing a theoretical basis and practical path for constructing a future-oriented intelligent cloud platform scheduling system. Subsequent research will further expand the multi-tenant adaptation capability and heterogeneous resource scheduling mechanism of the model, and enhance the intelligent autonomy level and resource collaboration efficiency of the overall system.

References

1. Z. Wang, J. Zhang, L. Li, H. Chen, M. Liu, X. Zhao, et al., "Computer vision system for multi-robot construction waste management: Integrating cloud and edge computing," *Build.*, vol. 14, no. 12, p. 3999, 2024, doi: 10.3390/buildings14123999.
2. V. Tabunshchik, A. Kerimov, I. Guliyev, F. Khalilov, R. Aliyev, M. Abdullayev, et al., "The dynamics of air pollution in the southwestern part of the Caspian Sea Basin (based on the analysis of Sentinel-5 satellite data utilizing the Google Earth Engine cloud-computing platform)," *Atmosphere*, vol. 15, no. 11, p. 1371, 2024, doi: 10.3390/atmos15111371.
3. H. Jang and H. Koh, "A unified web cloud computing platform MiMedSurv for microbiome causal mediation analysis with survival responses," *Sci. Rep.*, vol. 14, no. 1, p. 20650, 2024, doi: 10.1038/s41598-024-71852-y.
4. Y. Jiang, X. Wang, Z. Liu, H. Li, J. Chen, M. Sun, *et al.*, "Improvement of monitoring technology for corrosive pollution of marine environment under cloud computing platform," *Coatings*, vol. 12, no. 7, p. 938, 2022, doi: 10.3390/coatings12070938.
5. A. Kumar and P. Sivakumar, "Cat-squirrel optimization algorithm for VM migration in a cloud computing platform," *Int. J. Semant. Web Inf. Syst.*, vol. 18, no. 1, pp. 1–23, 2022, doi: 10.4018/IJSWIS.297142.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of GBP and/or the editor(s). GBP and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.