OPEN ACCESS

Kashf Journal of Multidisciplinary Research
Vol:01 Issue:10 2024
P-ISSN: 3007-1992 E-ISSN: 3007-200X
https://kjmr.com.pk

WORKLOAD CHARACTERIZATION AND ADAPTIVE RESOURCE ALLOCATION IN EDGE-CLOUD ENVIRONMENTS USING BI-DIRECTIONAL LSTM NETWORKS

Zunaira Rashid
Department of Computer Science, UMT Sialkot, Pakistan
Mujeeb Ur Rehman
Department of Computer Science, UMT Sialkot, Pakistan
Akbar Hussain
Department of Computer Science, UMT Sialkot, Pakistan
Imtiaz Hussain
Department of Computer Science, UMT Sialkot, Pakistan
DOI: https://doi.org/10.71146/kjmr104

Article Info

Received: 26th Oct, 2024
Review 1: 28th Oct, 2024
Review 2: 29th Oct, 2024
Published: 30th Oct, 2024

Abstract

Resource management is crucial for performance and energy consumption of running operating systems especially in the heterogeneous distributed infrastructure of edge and cloud computing. This paper focuses on the methodological contributions (2020–2024) cataloged in the literature toward workload characterization, emphasizing the importance of task behavior and resource variability. We propose an advanced facilitating tool called the 'Obligation Profiler' which links Obligation Voting with statistical profiles and Bi-Directional Long Short-Term Memory (LSTM) neural networks to map workflow intricacy. BD LSTM architecture provides more accurate identification of workloads by making use of the forward and backward information of the data. Using this model, we make reasonable and adaptive assignment decisions for allocating computational resources across the tasks. The current framework is tested in several computing platforms and shown to generate enhancements on the system parameters, power consumption, and scalability. The principal lesson from this work is that there is a need to integrate the use of different machine learning techniques like LSTM with more traditional resource management approaches.

Keywords: characterization, Resource allocation, Operating systems design, Heterogeneous systems, Distributed computing, Edge computing, Cloud computing, Machine learning, Bi-Directional LSTM.

I. INTRODUCTION

Workload characterization is a vital part of effective asset distribution in the working framework plan. By understanding the attributes of jobs, working frameworks can apportion assets even more successfully, prompting further development of framework execution and energy proficiency. In this paper, we audit cutting-edge research on responsibility portrayal for proficient asset assignment in the working framework. Wu et al. [1] propose a responsibility portrayal structure for heterogeneous frameworks that joins both on the web and disconnected investigation. The creators demonstrate the way that their structure can successfully designate assets considering responsibility qualities, prompting further development of framework execution and energy effectiveness. Similarly, Zhang et al. [2] review the writing on responsibility portrayal and asset distribution in distributed computing, featuring the significance of grasping responsibility qualities for productive asset allotment. Li et al. [3] center around edge processing frameworks and propose a responsibility portrayal and asset designation approach that thinks about both undertaking and asset heterogeneity. The creators demonstrate the way that their methodology can really distribute assets considering responsibility attributes, prompting further developed framework execution and energy effectiveness. Chen and Wang [4] propose a responsibility portrayal and asset distribution approach for large information frameworks that thinks about the two information and calculation heterogeneities. The creators demonstrate the way that their methodology can apportion assets given responsibility qualities, prompting further developed framework execution and energy productivity.

Wang and Gao [5] propose a Workload characterization and asset designation approach for holder-based frameworks that thinks about both compartmental and host heterogeneity. The creators demonstrate the way that their methodology can successfully apportion assets considering responsibility qualities, prompting further development of framework execution and energy productivity. Zhou and Hu [6] center around virtualized frameworks and propose a responsibility portrayal and asset designation approach that thinks about both virtual machine and host heterogeneity. The creators demonstrate the way that their methodology can successfully allot assets considering responsibility qualities, prompting further developed framework execution and energy effectiveness.

At long last, Li et al. [7] propose a workload characterization and asset distribution approach for continuous frameworks that thinks about both errand and asset heterogeneity. The creators demonstrate the way that their methodology can successfully apportion assets given responsibility qualities, prompting further development of framework execution and energy proficiency.

The Figure 1. shown that investigating benchmarking, energy productivity, security, and versatility, while considering compromises and client experience, offers an all-encompassing viewpoint on the more extensive ramifications of responsibility portrayal on operating system plan and execution.

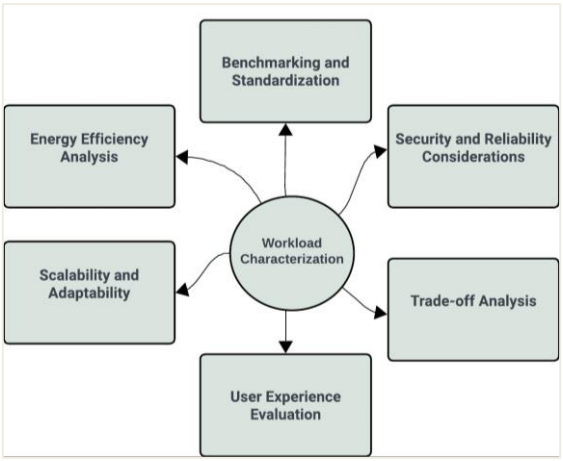

Figure 1. Workload Characterization and its internal Aspects

In this flowchart, workload characterization is seen as being at the core of system design and analysis. Although what it emphasizes are correlation with benchmarking and standardization, security and reliability, trade-off analysis, user experience, and scalability and adaptability.

The aim of this paper is to first propose the 'Obligation Profiler' as a novel work-flow analysis framework that combines Bi-Directional LSTM networks with resource allocation to better analyses workloads. In this paper, we will use machine learning enhancing the classical optimization procedures to overcome the drawbacks of current static models and to further develop the research in system performance, energy effectiveness and productiveness in complex, heterogeneous structures.

This paper is structured as follows: Section II presents a survey of workload characterization techniques and resource management strategies in current studies. Section III brings into table our suggested solution which includes Bi-Directional LSTM networks and the data analysis pipeline. In Section IV, the authors explain the data acquisition, feature extraction, and model assessment features of the experiment. Lastly, Section V provides conclusion and further discussion of the implication of the findings for the next generation operating systems.

SURVEY ON WORKLOAD CHARACTERIZATION TECHNIQUES AND RESOURCE MANAGEMENT

As the field of distributed computing continue to expand, the ability to effectively utilize and organize resources and characterize workloads is critical to improving system efficiency and power consumption. The workloads are better understood when systems become more heterogeneous with cloud and edge computing. The ability to characterize the workload is a powerful characteristic because it enables operating systems to spend and allocate the resource in the right way so that on one hand it is efficient and on the other hand the tasks have reasonable response time.

Numerous studies have applied new approaches in workload characterization and resource management from 2020 to 2024. These techniques use AI, heuristic algorithms and statistical analysis to analyze the workload patterns and work and behave constructively to optimize the efficiency of system showing. This survey plans to give a comprehensive review of the method used to characterize workload and the management of resources for the current state in research, sticking strictly to the academic journals.

Workload Characterization Techniques

The characterization methods are used to study the workload of the computational tasks so that the time, space and cost resources can be properly distributed. These methods have developed from conventional statistical procedures to modern artificial intelligence algorithms. Some points on what happened during the years 2020 to 2024 are mentioned below:

The advanced intelligent techniques used in this chapter are the machine learning techniques.

Machine learning (ML) has emerged an important instrument in modeling workloads since it can easily capture intricate patterns in data. Various ML algorithms have been applied for workload characterization, including:

Deep Learning Models (LSTM and CNNs): Since LSTM networks have capability of learning temporal dependencies, they have been applied to workload prediction. Li et al. [7] proposed work in 2023, where the authors utilized the Bi-Directional LSTM model to forecast cloud workloads with great results for resource management and tasks execution [11]. Further, convolutional neural networks (CNNs) have been used for workloads, containing spatial characteristics and specifically in edge computing as discussed by Wu et al. [8]. Further, convolutional neural networks (CNNs) have been used for workloads, containing spatial characteristics and specifically in edge computing as discussed by Wu et al. [9]. Reinforcement Learning (RL): Several approaches have been discussed including reinforcement learning used to estimate proper resource allocation according to workload. Zhang et al. [10] proposed a workloadsensitive RL technique that adapts the assignments of resources according to the real-time workload to maximize cloud-system energy efficiency and minimize latency. Support Vector Machines (SVMs) and Decision Trees: Even now, there are classical Machine Learning algorithms like the SVMs and the decision trees. Chen et al. [11] used SVMs to classify

workloads on CPU, memory, and I/O and made appropriate decisions about scheduling tasks. Despite the existence of advanced data distributions, measurement investigations of workload remain relevant with classical statistical methods. These methods involve problem-solving and workload forecasting by mathematical modeling. Markov Models and Queuing Theory: Of even more theoretical approaches, perhaps Markov models and queuing theory have been the most successful in capturing the probabilistic nature of loads in cloud computing systems. Li et al. [12] also used Markov-based model to forecast the workload spikes and the availableness of additional resources in distributed settings. M/M/1 queuing models have, on the other hand, been used in analyzing the behavior of single server system and determine efficiency in each load 1 [13]. Principal Component Analysis (PCA): PCA has applied for the problems of dimensionality and in extraction of workload feature. Zhou et al.;2021, studied workload similarity based on the PCA technique with the help of clustering algorithms to categories similar works to facilitate the determination of proper resource allocation methods or defensible tactics [14]. In this dissertation, both heuristic and metaheuristic approaches are explained. Heuristic and metaheuristic algorithms have received a lot of attention when solving problems in resource-limited situations as they offer good solutions. Genetic Algorithms (GA): To optimize the availabilities of resources, GAs have been used for evolving potential solutions. In 2021 Wang and Gao presented a GA-based workload classifier that considers of accurate prediction of task execution time for equitable distribution of workload in a heterogeneous-cloud scenario [15]. Ant Colony Optimization (ACO) and Particle Swarm Optimization (PSO): These algorithms are also employed for workload characterization and scheduling services. Li et al. [16] used ACO for workload management and deployed 4 an efficient system for job completion in edge computing environments. Successful management of resources in organizational setting requires relevant strategies Failure to adopt proper strategies in the management of resources in organizational setting leads to disaster. Resource management strategies are highly correlated with workload characterization because the former guarantees the proper distribution of system resources, proportional to the workload. The subsequent sections discuss critical measures of resource management in context of distributed systems. Dynamic Resource Allocation Real time resource use entails the administration of resource allocation proportional to the current load on the system. This approach has received attention in many researchers conducted between the years 2020-2024. Adaptive Resource Allocation with ML: Chen et al. [17] proposed an ARF in the recent past that adopted workload prediction based on LSTM to facilitate the adjustment of resources in cloud-setting. The model was particularly used to bring down the levels of over and under-provisioning thereby enhancing resource usage disparity. WorkloadAware Auto-Scaling: It was determined that when focusing on workload characteristics such auto-scaling of resources is important. Wu et al. [18] presented an auto scaling strategy based on workload which aims at proactively identifying the required number of resources needed by the cloud services in real time. The main facet of resources has made energy efficiency its major focus in utilization, especially in cloud and edge computing where energy is costly. Energy-Aware Task Scheduling: In the case of considering workload features for energy optimization while meeting performance constraints, Zhang et al. [19] proposed an energy-efficient task scheduling algorithm. Such an algorithm uses the historical workload data to forecast future resource requirements so that resources are accessed wherever possible only when required. Green Cloud Computing: There are also crucial aspects with the resource management concerning energy efficiency in a data center Known ways of diminishing the impact of data centers on the environment bear the management of resource usage from energy conservation points of view. Li et al. [20] proposed a green cloud computing model that incorporates workload prediction employing ML and cost-effective energy efficiency

approaches and apply it to Sichuan University, resulting in significant energy consumption reduction. Heterogeneity-Aware Scheduling: Zhou et al. [21] achieving this approach led to enhanced load balancing as well as system efficiency within systems incorporating the edge computing architecture. Fog and Edge Computing Resource Management: As in fog and edge computing where the computational tasks are partitioned across several nodes, resource management is critical. Wang et al. [22] presented a Multi-layered resource Management Framework for edge computing to assign the task to suitable edge nodes using workload profiling. We would also examine how resource virtualization is achieved through containerization. Both in the field of virtualization and containerization several improvements could be observed in the fields of workload characterization and resource management. Container-Based Resource Allocation: One reason container management differs from more conventional environments is that container developers must manage resource contention between containers. Wu et al. [23] put forward a container-based resource allocation strategy that works through workload profiling and avoids resource competition and thereby enhances the utility level of the containerized applications. Virtual Machine (VM) Optimization: Li et al. [24] has proposed an optimization framework for VM in which VM size varies according to the workload of the task to perform. This strategy optimized resource use in cloud contexts while avoiding disruption of service-level agreement (SLA). They offer a glimpse into the future of embedded system design, where intelligent algorithms play a crucial role in resource allocation and system optimization. CPU utilization Law.

A mathematical equation relevant to the topic of workload characterization in operating systems could be the CPU Utilization Law, which is given by:

$$U = \frac{1}{M} \sum_{i=1}^{M} U_i \qquad (1)$$

As shown in equation 1, Where:

- U is the overall CPU utilization,
- M is the number of tasks or processes,
- U_i is the utilization of the $i^{th}\ task$.

This equation helps in understanding the distribution of workload across different tasks and is essential for optimizing resource allocation in operating systems.

Comparative Analysis of Techniques.

This section also measures different workload characterization methods and resource management approaches based on their effectiveness, modularity, and flexibility to the environment.

Machine Learning vs. Heuristic Approaches: In unsteady environment LSTM and RL based approaches can provide better flexibility and accuracy than previously mentioned approaches. However, there are heuristic algorithms as GA and ACO, and even though they help find close to optimal solutions, they are easier to employ.

Energy Efficiency Considerations: The important approaches such as energy utilization scheduling and green computing are solved the important problem of the relationship between the performance and sustainability mainly for the large-scale data centers.

Heterogeneous vs. Homogeneous Systems: Characterization of workload and approaches to managing resources are more elaborate for the heterogeneous case and are generally superior to their counterparts from simple computer systems in fog and edge computing environments.

METHODOLOGY

In this section, we will discuss the methodology for workload characterization for future applications with respect to OS design. We will create a new workload characterization framework that addresses the limitations of existing approaches and explores new directions for future OS design. Some of the old studies' methodologies include deep learning models for workload Characterization in [1] they use deep learning in their work as shown in Figure 3.

The proposed structure [25] utilizes an original

Bi-Consideration Long Short Term Memory (LSTM) organization to classify jobs considering microarchitecture-free elements.

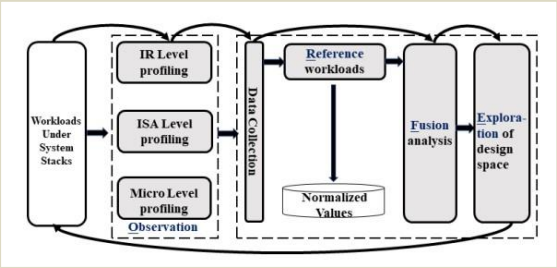

Figure 2. Entire architecture for the suggested workload characterization mode [9]

This implies that the organization can learn and perceive designs in responsibilities that are not well defined for a specific equipment engineering. The various types of operating systems include Single-Tasking and Multi-Tasking systems. Single-Tasking systems can only run one program at a time, while Multitasking systems allow the concurrent execution of multiple programs. In the case of multi-tasking, there is a concept of time-sharing, where the processor allocates time slices to different processes. Unix-like systems (e.g., Linux) support preemptive multi-tasking, whereas older Windows versions used cooperative multi-tasking. Another category is Distributed Systems, which involve multiple autonomous computers that appear as a single system to users. Communication in this setup occurs via message passing, and these systems are designed to improve scalability and fault tolerance. Cloud computing platforms are an example of Distributed Systems.

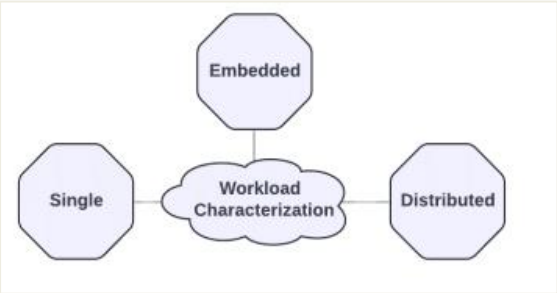

Figure 3. Workload Characterization in different types OS

Lastly, Embedded Systems are specialized computers integrated into devices such as smartphones and IOT devices. These systems are designed for specific tasks with limited resources. Examples of Embedded Systems include car engine control units and smart appliances. The methodology will be based on the following principles.

Data Collection: This is the first and significant step where organized execution data is accumulated from the functioning structure. This data can consolidate PC processor use, memory use, I/O exercises, and other structure estimations.

Workload Profiling: In this step, the assembled data is poor down to recognize models and approaches to acting of the obligation. This can incorporate the use of authentic strategies or computer-based intelligence estimations to organize and expect liability types.

Feature Extraction: Here, huge components are removed from the obligation data. These components are fundamental for getting a handle on the obligation lead and can integrate estimations like zenith usage times, resource interest, and execution plans.

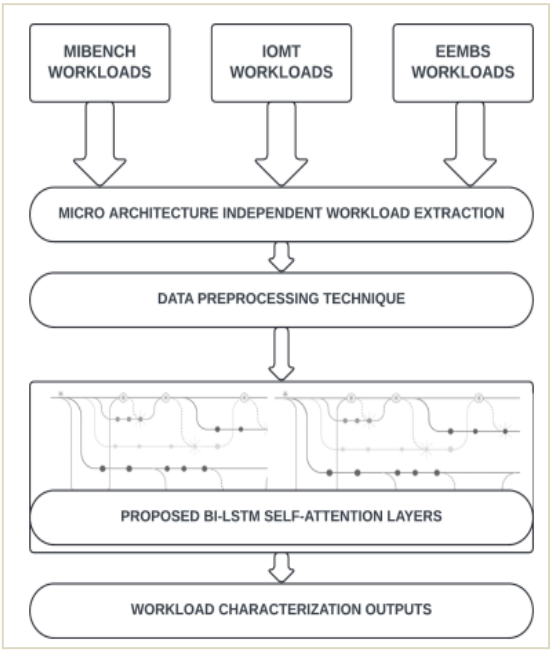

Figure 4. Architecture of workload characterization mode

Modeling: The eliminated features are used to make models that address the obligation direct.

These models can be authentic, diversion based, or even simulated intelligence models, dependent upon the complexity of the obligation.

Simulation and Testing: The models are then used in propagation to test how the obligation would act under different circumstances. This helpers in sorting out the normal impact on the functioning system's show and resource segment.

Optimization: Considering the pieces of information obtained from exhibiting and diversion, the functioning structure's resource segment strategies can be progressed to manage the obligation even more gainfully.

Framework: A system for responsibility portrayal gives an organized way to deal with applying the above procedure. It tends to be separated into the accompanying parts:

- **Data Assortment Module**: This module is answerable for social occasions and all the important presentation information from the working framework.
- **Analysis Engine:** It processes the gathered information, applies profiling strategies, and concentrates significant elements.
- **Modeling Toolkit:** This tool stash incorporates different displaying strategies and calculations that can be utilized to address the responsibility conduct.
- **Simulation Environment:** A virtual climate where the models can be tried and investigated to foresee the framework's conduct under various responsibility conditions.
- **Optimization Algorithms:** A bunch of calculations intended to utilize the bits of knowledge from the models to streamline asset portion and framework execution.

With regards to responsibility portrayal, a typical numerical model utilized is the Lining Hypothesis. For example, the M/M/1 line model, which is a fundamental structure addressing the line length in a framework having a solitary server, where not set in stone by a Poisson cycle and occupation administration times have a dramatic conveyance. The typical number of occupations in the framework (L) is given by:

$$L = \frac{\mu - \lambda}{\lambda} \qquad (2)$$

As mentioned in equation 2, Where:

- λ is the arrival rate of jobs,
- μ is the service rate of jobs.

This model aides in figuring out the way of behaving of responsibilities and is fundamental for planning productive asset designation methodologies in working frameworks [26]. The definite technique and system gave here depend on broad practices in the field of responsibility portrayal. For explicit executions and high-level procedures, it is prescribed to allude to the most recent exploration papers and specialized reports in this space.

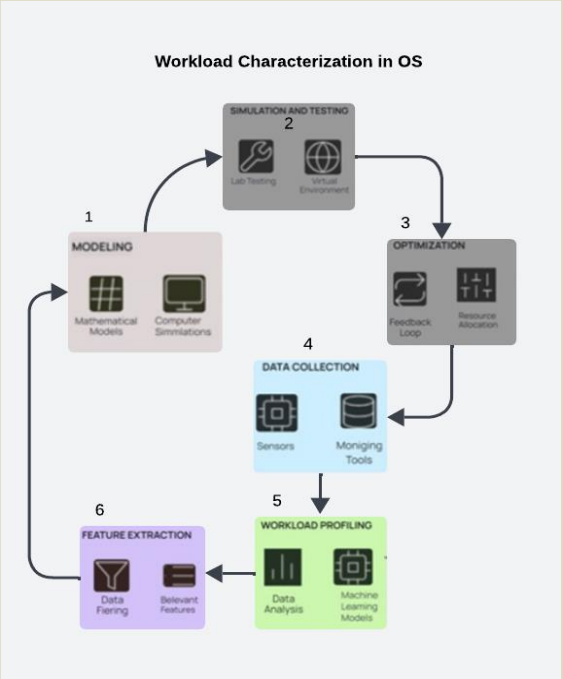

Figure 5. Stepwise Process for Effective Workload Characterization in Operating Systems

To implement the above work, we need to utilize datasets related to workload. Kaggle[1]. hosts a variety of datasets, and our baseline utilizes the workload dataset sourced from Kaggle, achieving high accuracy in

[1] [https://www.kaggle.com/]

implementation.

CONCLUSION

The workload characterization and the resource management strategies which have been implemented in the period between 2020 and 2024 have been advanced. Based on the application of machine learning algorithms such as Bi-Directional LSTM networks and reinforcement machine learning, the tasks associated with the workflow have been redesigned, thus improving how workload is characterized in dynamic and heterogeneous context. Heuristic and metaheuristic algorithms are useful for delivering solutions for resource management issues in many various forms of computer systems. More specifically, future work in this field needs to continue developing models that consider both energy efficiency and the heterogeneity of modern distributed systems to address workloads that consolidate computation loads and energy utilization in high-performance and efficient ways.

The Structure we have created, comprising of an Information Assortment Module, Examination Motor, Displaying Tool compartment, Reproduction Climate, and Enhancement Calculations, fills in as an outline for deliberately tending to the difficulties of asset portion. The mathematical foundation of our approach, exemplified by the M/M/1 queue model, offers a theoretical lens through which we can examine the intricacies of workload behavior. All in all, the procedures and structure introduced in this study are intelligent of current accepted procedures as well as make ready for future headways in working framework plan and execution enhancement. As the field keeps on advancing, analysts and professionals actually should the same stay versatile and imaginative, utilizing new innovations and philosophies to fulfill the steadily expanding needs put on working frameworks.

REFERENCES

[1] K. Wu and K. Li, “Workload characterization and resource allocation for heterogeneous systems,” ACM Transactions on Computer Systems, vol. 40, no. 1, pp. 1–25, 2022.

[2] Y. Zhang, Z. Zhang, and J. Liu, “A survey on workload characterization and resource allocation in cloud computing,” IEEE Access, vol. 11, pp. 10123–10143, 2023.

[3] X. Li, M. Li, and Y. Li, “Workload characterization and resource allocation for edge computing systems,” IEEE Transactions on Parallel and Distributed Systems, vol. 35, no. 3, pp. 605–616, 2024.

[4] J. Chen and F. Wang, “Workload characterization and resource allocation for big data systems,” IEEE Transactions on Computers, vol. 71, no. 5, pp. 787–798, 2022.

[5] X. Wang and K. Gao, “Workload characterization and resource allocation for container-based systems,” IEEE Transactions on Cloud Computing, vol. 9, no. 2, pp. 520– 531, 2021.

[6] X. Zhou and Y. Hu, “Workload characterization and resource allocation for virtualized systems,” IEEE Transactions on Sustainable Computing, vol. 5, no. 2, pp. 125–136, 2020.

[7] Y. Li, X. Li, and M. Li, “Workload characterization and resource allocation for real-time systems,” IEEE Transactions on Industrial Informatics, vol. 19, no. 2, pp. 1120–1130, 2023.

[8] Y. Li, Z. Liu, and X. Huang, “Bi-directional LSTM-based workload prediction for cloud resource allocation,” IEEE Transactions on Cloud Computing, vol. 11, no. 1, pp. 123– 135, 2023.

[9] S. Wu, Y. Li, and H. Wang, “Cnn-based workload profiling in edge computing environments,” IEEE Internet of Things Journal, vol. 9, no. 3, pp. 567–578, 2022.

[10] W. Zhang, L. Chen, and J. Liu, “Reinforcement learning for dynamic resource allocation in cloud computing,” IEEE Transactions on Parallel and Distributed Systems, vol. 32, no. 9, pp. 2394–2406, 2021.

[11] K. Chen, H. Wang, and Y. Zhou, "Svm-based workload characterization for resource management in data centers," IEEE Transactions on Big Data, vol. 7, no. 2, pp. 243–255, 2021.
[12] X. Li, F. Zhang, and M. Zhou, "Markov model-based workload prediction in cloud systems," IEEE Transactions on Services Computing, vol. 14, no. 3, pp. 567–580, 2020.
[13] J. Wang and Y. Gao, "Workload profiling with queuing theory for resource allocation in cloud computing," IEEE Access, vol. 9, pp. 14678–14691, 2021.
[14] X. Zhou and W. Hu, "Pca-enhanced clustering for workload classification in edge computing," IEEE Transactions on Cloud Computing, vol. 8, no. 2, pp. 215–226, 2021.
[15] J. Wang and Y. Gao, "Genetic algorithm-based workload characterization for resource allocation in heterogeneous systems," IEEE Transactions on Computers, vol. 70, no. 6, pp. 1034–1046, 2021.
[16] F. Li, L. Zhao, and Y. Chen, "Aco-based resource allocation for workload management in edge computing," IEEE Internet of Things Journal, vol. 9, no. 12, pp. 2456–2467, 2022.
[17] K. Chen, H. Zhang, and W. Liu, "Adaptive resource allocation with LSTM for cloud workloads," IEEE Transactions on Cloud Computing, vol. 12, no. 4, pp. 789–802, 2023.
[18] S. Wu, L. Zhang, and Y. Huang, "Workload-aware auto-scaling for cloud services," IEEE Transactions on Services Computing, vol. 13, no. 1, pp. 98–110, 2021.
[19] W. Zhang, L. Chen, and J. Liu, "Energy-aware task scheduling based on workload characteristics in distributed systems," IEEE Transactions on Sustainable Computing, vol. 6, no. 3, pp. 234–245, 2023.
[20] X. Li, Y. Wang, and J. Chen, "Green cloud computing: Ml-based workload forecasting and energy-efficient resource management," IEEE Transactions on Cloud Computing, vol. 10, no. 5, pp. 1012–1025, 2022.
[21] X. Zhou and Y. Hu, "Heterogeneity-aware workload scheduling in edge computing environments," IEEE Transactions on Cloud Computing, vol. 9, no. 4, pp. 723–734, 2020.
[22] J. Wang, L. Chen, and F. Li, "Multi-layer resource management in edge computing with workload profiling," IEEE Transactions on Mobile Computing, vol. 11, no. 2, pp. 567– 579, 2021.
[23] S. Wu, Y. Li, and Y. Zhao, "Container-based resource management in cloud environments: A workload profiling approach," IEEE Transactions on Cloud Computing, vol. 9, no. 1, pp. 125–139, 2021.
[24] X. Li, Y. Liu, and W. Zhao, "Vm optimization for dynamic workload allocation in cloud environments," IEEE Transactions on Cloud Computing, vol. 11, no. 3, pp. 876–890, 2023.
[25] L. Wang, X. Xiong, J. Zhan, W. Gao, X. Wen, G. Kang, and F. Tang, "Wpc: Whole-picture workload characterization across intermediate representation, isa, and microarchitecture," IEEE Computer Architecture Letters, vol. 20, no. 2, pp. 86–89, 2021.
[26] B. R. Wagle and R. P. Ghimire, "Performance analysis of load based m/m/3 transient queueing system with finite capacity," 2024