STUDY ON RESOURCE SCHEDULING AND ALLOCATION FOR MULTI-OBJECTIVE-BASED VIRTUAL MACHINES

K. KRISHNAKUMAR, N. MALARVIZHI, T. KUMANAN

1Assistant Professor, Department of Computer Science and Engineering, Meenakshi Institute of Higher Education and Research, Chennai, Tamilnadu, India.

2Professor, Department of Computer Science and Engineering, School of Computing, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, Tamil Nadu, India.

3Professor, Department of Computer Science and Engineering, Meenakshi Academy of Higher Education and Research, Chennai, Tamil Nadu, India.

1kkrishnakumar86@gmail.com, 2drnmalarvizhi@gmail.com

ABSTRACT

Over the Internet, the way of storing and access to data and programs is termed cloud computing. This does not require a dedicated server in residence. Cloud computing provisions data access anywhere, any time. Google Drive, Apple I Cloud, Amazon are some of the big retailers in Cloud. Users go on with pay and use service in Cloud. Many researchers have developed various meta-heuristic algorithms to optimize capital distribution of resources. The virtual resource allocation method includes construction/destruction of virtual machines dynamically, without affecting the execution of the application. Therefore, assigning resources at the virtual machine level provides several distinct advantages such as consistency in the distribution and migration of virtual resources. In this paper various reviews related to study on resource scheduling and allocation for multi-objective-based virtual machines has been studied.

Keywords: Resource scheduling, Virtual Machines, Optimization, Resource distribution

I. INTRODUCTION

Serverless computing is the latest development in cloud computing; for example, the Lambda function offered by Amazon Web Services. These functions can run either in the public or private Cloud. The concern here is that although users go on the pay-per-use model, the actual billing does not depend on usage and allocation (Eivy & Weinman (2017)). Therefore most of the time customer pays more than the necessity. Hence resource allocation and scheduling is a challenge that prevails in the existing Cloud. In serverless, until a function is called, no resource is allocated or changeable. However, this depends on the behaviour and execution of the workload. The services are to be tunable, testable, resilient, and monitored (Eivy & Weinman (2017)). For instance, the AWS Lambda pricing offer one million hits on 400,000 GB seconds free per month. Though the scheme seems to be a lot of free usages, it is only the one needed. Therefore cloud billing is confusing for the customer, and the need to be wary of the price. The virtual machine infrastructure cost for 150 hits per second is $200 per month. In the case of 30k, it is $18k per month. Hence it is necessary for 100% utilization and customer satisfaction (Eivy & Weinman (2017)). Virtual machines make scheduling choices by sending the tasks to the cloud environment[16][17][18], which is carried out based on the existing device state, as shown in Figure 1.
The rest of this paper is organized as follows: Section 2 discusses the multiobjective schemes in cloud and section 3 describes the detailed review on the resource scheduling and allocation in cloud. Section 4 presents the conclusion.

**II. MULTIOBJECTIVE SCHEMES IN CLOUD**

Abazari et al. (2019) proposed a secure and performance-aware scheduling model to minimize the security threats assigning virtual machine tasks, verifying their execution order and their attack response policy in a heterogeneous infrastructure. Though the proposed model had a reasonable completion time, the algorithm had limitations due to its absence in dynamic scheduling. Therefore it is necessary to develop an efficient algorithm from the multiobjective perspective that performs dynamic workflow scheduling.

Zuo et al. (2016) proposed a Multiobjective Scheduling based on Ant Colony Optimization (MOSACO) performing task scheduling and optimizing the private and public cloud-based computing resources. The algorithm considered the completion time, cost, QoS under deadline constraints. Simultaneously, the algorithm showed only a considerable improvement task completion time and cost.

Janagoudar et al. (2020) described a multiobjective scheduling for estimation of host load using machine learning algorithms. Considering multi-tenant data centers resource monitoring is a tedious task. The algorithms considered the parameters such as memory, CPU utilization, and the workload to classify whether virtual machine was over/underloaded. Resource allocation with less energy consumption is a critical requirement today. The proposed mechanism incorporated with the logistic regression classifier performs server classification with various metrics. Simultaneously, the algorithm was compared only with worst fit scheduling and was better than the worst fit in resource provisioning and utilization.

Srichandan et al. (2018) defined a hybrid approach-based scheduling by utilizing a biologically inspired bacterial foraging algorithm and genetic algorithm. The scheduling was performed in a heterogeneous cloud where the algorithm solved the multiobjective optimization issue considering the reduce of makespan and energy consumption. Mutation and crossover technique achieve local and globally optimal solutions. In contrast, the similarity of each factor was not examined. The algorithm did not increase the convergence rate and also took additional time for crossover and mutation. Thus the algorithm showed few improvements in the service-oriented manufacturing domain.

**III. RESOURCE SCHEDULING AND ALLOCATION IN CLOUD: A REVIEW**

An increase in cloud applications increases the cloud system complexity. Jang et al. (2019) proposed a novel VM scheduling using the loan and redeem system. The technique presents two types of boosting during execution. These are i) under credit for non-boosted execution and ii) boost credit for boosted execution. The credit rating is reflected by the characteristics of the application running on a virtual process. The credit decides to boost and

---

Figure 1: An example of Queue based Dynamic Resource Allocation

---
reflects on the application characteristics. The mechanism was run on a Xen hypervisor 4.6 and compared with a credit scheduler. The system was evaluated for response time, network bandwidth, utilization. The scheduler improved the I/O performance for a cloud system. However, the deduction time of boost credits was controlled, and the scheme achieved 39% less response time. Therefore a dynamic solution for cloud virtual machines remains a challenge.

Jeevitha & Athisha (2020) addressed energy consumption, utilizing the Dynamic Voltage Frequency Scaling approach. It is the combination of the shortest job first and round-robin with vibrant quantum. Concentrating on the cloud data center energy efficiency, the scheduling minimized the process waiting time and starvation. However, the load balancing scheme and factors that were not included in the study are significant drawbacks.

Dou et al. (2017) proposed a dynamic scheduling based on energy awareness in enhancing the QoS during bigdata execution in cloud environments. The technique considered VM migration and two dynamic VM migration phases focused on reducing execution time and service price. The first phase of all the VM was active and then on high energy-consuming server migration. Thus saving energy cost and providing users with discount prices. In the second phase, the optimization reduces the execution time migrating to a high-performance server, reducing the execution time. However, the method was not intensive and efficient for real-world data centers and required improvement in meeting the personalized requirements.

Santhosh & Manjaiah (2016) proposed scheduling algorithms based on user priority and deadline. The user priority is categorized into high, medium, and low, having three queue sets. The queue stores the job which do not require a specific CPU. Using first-come first scheduling, user requests are arranged, and the deadline is processed. Multiple jobs with the same deadline are selected using the FCFS method. Jobs are arranged to ascend to the deadline, and low priority is moved to medium and then to high. However, the job has to wait until it reaches priority is a major drawback here.

Quang-Hung et al. (2017) presented a particular case of scheduling problems using parallel workloads. It utilizes the EminTRE-LDTF algorithm to allocate virtual machines from multiple resources to fixed interval time. They were of non-preemption to physical machines meeting the resource requirements. The algorithm fixes a start time and duration of the virtual machine to reduce the overall busy time and minimize physical server energy consumption. Though the algorithm reduced the energy consumption, it does not deal with job requests received from heterogeneous physical servers.

Sui et al. (2019) proposed a virtual machine scheduling model to reduce power consumption and performance interference among virtual machines. Virtualization saves the energy of cloud datacenters. Apart from this, virtual machines are deployed on physical hosts that perform computation as an independent unit. Considering a single virtual machine, the physical resources compete, causing degradation due to interference among the virtual machines. However, the method produced only less improvement on the CPU utilization preventing individual servers from highlighting.

Mansouri et al. (2019) presented a hybrid method that includes the combination of PSO algorithms and fuzzy systems to increase cloud throughput and load balance. Several parameters have been obtained from this hybrid algorithm in terms of makespan and imbalance degree. The proposed algorithm seems to provide a significant result compared to the other existing algorithms. Several mechanisms are involved with the resource allocation methodologies that had been presented in the cloud environment.

Based on a profound research methodology, an optimized resource scheduling algorithm was proposed in the IaaS cloud system infrastructure to establish the optimization for cloud scheduling. An Improved Genetic Algorithm (IGA) has utilized a technique of IGA for minimal genes and presented the Dividend Policy scheme in Economics to select the optimum allocation for virtual machines’ demands. Park et al. (2011) demonstrated a resource allocation strategy based on market (RAS-M) in which bulk data centers use the advanced resource consumption and also the cloud users could be able to receive services with higher Quality of Service (QoS).

Qi et al. (2020) focused on cyber-physical systems, which gained popularity in critical areas such as healthcare, traffic control, etc. Many enterprises utilize cyber-physical systems to handle and implement distributed computing resources. The authors find an optimal virtual machine scheduling for cloud-based cyber-physical systems enhancing the QoS. The non-dominated sorting genetic algorithm is utilized to find the optimal
scheduling strategy. The performance of the algorithm was evaluated in terms of energy consumption, resource utilization, and downtime. Though the algorithm minimized the downtime and maximized the resource utilization, the authors point on multiple user requirements in concern to QoS has to be taken into consideration.

Tian et al. (2018) proposed optimal energy-efficient scheduling for the virtual machine and deducing the lower-bound VM migration problem. Though the scheduling algorithms show efficiency, the authors have marked the open issues related to cloud scheduling. These include: i) finding a near-optimal solution is NP-complete, ii) Load balancing and energy efficiency need to be combined for producing an integrated solution, iii) users’ multiple requests have to be considered.

IV. CONCLUSION

Cloud Computing system having many workloads, including several assets owned by various entities, the user need not own the cloud infrastructure. Cloud computing's primary purpose is to allow customers to optimally schedule tasks with the least running time to achieve load balancing with the highest quality of service. Allocation of resource is considered integral component in cloud computing since it directly affects cloud computing performance. In this paper, a detailed literature survey on the experts' work towards resource scheduling and allocation for various multiobjective based virtual machines..

REFERENCES