ICIC Express Letters
Volume **14**, Number **9**, September **2020**

ICIC International ©2020 ISSN 1881-803X
pp. **909–916**

RESOURCE MANAGEMENT TECHNIQUES IN CLOUD COMPUTING: A STATE OF ART

Ankita Srivastava and Narander Kumar

Department of Computer Science
Babasaheb Bhimrao Ambedkar University (A Central University)
Vidya Vihar, Raebareli Road, Lucknow 226025, India
ankita31srivastava@gmail.com; nk_iet@yahoo.co.in

Received October 2019; accepted January 2020

Abstract. *Cloud computing is the new paradigm that has come into existence because of the tremendous demand for computing facilities. It offers services "pay as you go" which not only has attracted people to take its services personally but also provides IT solutions to small and large-scale industries. Resources are the basic components of cloud computing. The management of these resources has become a very challenging task for cloud providers. Significant research has been done in this field whose main objective deals with the effective management of resources. Various Resource Management (RM) techniques are designed which tackle the different parameters. This paper gives a detailed description of RM and the latest research conducted in this field has been elaborated. It provides a comprehensive review of RM techniques based on distinct parameters. It discusses the various evaluation parameters which are being used in the discussed techniques. Moreover, it presents the challenges existing in this field which needs the kind consideration of the researchers.*

Keywords: Resource management, Taxonomy, Cloud computing, Mobile clouds

1. **Introduction.** With the advancement in Information and Communication Technology (ICT) over the past few years, cloud computing has become an immense platform for hosting and delivering services to users over the Internet. Cloud computing is a computing utility that has infinite capacity, instantaneous scalability and user pays only for the resources they use and only for time duration they are using it [1]. RM and virtualization form the basis of cloud computing. RM is the core function of cloud computing that influences three aspects that are functionality, cost, and performance. In this regard, it needs complex policies and decisions for multi-attribute optimization. These policies are organized as capacity allocation, admission control, energy optimization, load balancing and quality of service guarantees. Admission control restrains the system from receiving the workload infringement of high-level policies. Energy optimization and load balancing can be performed either locally or globally and both are associated to cost. Lastly, the quality of service is associated with addressing requirements and objectives related to consumers [2]. It is the process of apportion of computing, network, energy resources and storage to a set of workloads in order to satisfy specifications and performance objectives of the cloud consumers and infrastructure providers respectively [3].

2. **Related Work.** In recent years, various techniques have been given by the researchers which consider the smooth provisioning, efficiency, maintenance or management of the cloud resources. These techniques are based on priorities or parameters like cost, scheduling methods, task type and the number of processors, throughput and energy. Resource provisioning, allocation, and scheduling are important issues in RM. One of them is optimization based on cost using Optimal Cloud Resource Provisioning (OCRP) which is

DOI: 10.24507/icicel.14.09.909

formulated via a stochastic programming model. The job performed by the OCRP algorithm is the provisioning of the resources which is executed through various stages of provision. Various solutions are provided for the algorithm such as sample average approximations, deterministic equivalent formulation, and Benders decomposition. This methodology presents a balance between resource allocation which is being demanded and resources in reservation [4]. A cloud workload management framework is given in which the workload is clustered via K-means with respect to the weight assigned to them and their QoS requirements and then these clustered workloads are scheduled according to the four scheduling policies based on time, cost, bargaining and compromised cost [5]. An improved version of the round-robin algorithm handles the resource allocation which makes use of dynamic time slice in place of the fixed time slice. It provides an efficient scheduling model which enhances the response time and the system's efficiency [6]. A decomposition method is proposed to answer the resource allocation issue which fulfills the users demands and also provides solution for maximization of profit for cloud providers [7]. Another algorithm for scheduling is elaboration of traditional greedy algorithm which optimizes the workflow scheduling [8]. With the increase in the data centers, power consumption has become one of the challenging issues in RM. A method has been discussed in which an optimal solution is given to maximize the resource utilization and identification of faulty resources is done to avoid misleading schedule. In this, resources are allocated to the workloads and fulfill the Quality of Service (QoS) with reduced Service Level Agreement (SLA) breach rate and maximum utilization of resources with reduced incurred cost. Since the prediction of resource utility is vital for efficient scheduling [9], an approach has been introduced which integrates prediction and feature selection of resource utilization which ultimately result in high performance. This technique improves accuracy significantly and reduces the execution time [10]. A technique Energy Efficient Load Balanced Resource Allocation Method (EELBRAM) is introduced which considers task scheduling and resource allocation in order to satisfy response time. Support vector machine approach has been used for resource allocation which helps in the improvement of the future resource requirement [11].

3. **Resource Management Techniques.** In cloud computing, all the resources which are utilized are virtualized and shared among the different users. Various techniques have been provided by the researchers to address the problem of efficient utilization of the resources. These techniques can be organized based on different parameters. These parameters can be cost, energy, SLA, load balancing and mobile cloud feature. So, there is a requirement for a technique which can handle the parameters and the issues efficiently. Figure 1 depicts the taxonomy of RM techniques. Some of these techniques have been examined below.

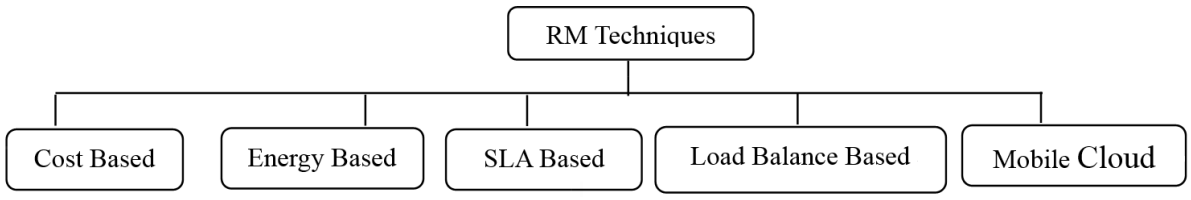

FIGURE 1. Taxonomy of RM

3.1. **Cost-based RM techniques.** In the current scenario, where the cloud users are increasing day by day, their will to get the highest possible QoS in minimum price is increasing day by day. Similarly, cloud providers are also interested in profit maximization. To achieve this aim, various pricing models and economic approaches have been given by various researchers. The main issue in this regard is first, to provide the best services at

a minimum cost for users; secondly, to provide the optimum services to the users with maximum profit within the agreed SLA by cloud providers. Some of the techniques are discussed in detail. A combinatorial double auction model has been proposed in [12]. This performs the efficient allocation of resources on a combinatorial auction model. In this paper, it has been proposed that the double auction method is more efficient than a single-sided auction model as it regulates the providers' monopoly. The model has been evaluated based on economic efficiency and incentive compatibility. The above paper does not consider the resource allocation issue, so a new method has been proposed in which a mechanism for virtual machine allocation and pricing is introduced. It undertakes two technologies: consensus estimate and revenue extraction. The first one provides the optimized target revenue while the latter identifies the winners and distribute the target revenue equally among them. This mechanism achieves truthfulness and envy freeness while delivering stable performance for large scale cloud markets [13]. This motivates the researchers to propose an efficient mechanism which not only provides optimized pricing mechanism for cloud providers and users but also effective management of resources with better QoS and throughput.

3.2. **Energy based RM techniques.** Energy is one of the major concerns in cloud computing. With the increase in the cloud users, the data centers are increasing day by day which is geographically located in different regions. These data centers require a lot of power to work. These centers also release a lot of CO_2 which is hazardous for the environment. So, researchers have given a lot of effort in managing resources with energy optimization. Some of the latest work has been discussed below. An energy-efficient resource RM has been given which provides migration instruction to stack up the virtual machines which result in reducing the number of running physical machines and hence in power consumption. It maintains QoS while delivering the services [14]. Another algorithm that considers energy and SLA violation has been proposed. It is based on learning automata and improves the utilization of the resources and further optimizes the energy consumption. It monitors the variation in user demands to predict the number of physical machines to be used and thus reduces the number of migration and shuts the idle servers. It helps in predicting resource usage in a cloud system [15]. Thus, it can be seen that energy-efficient management of resources is the most required part of the RM.

3.3. **SLA based RM techniques.** SLA is one of the most important parts of cloud computing services. It is a kind of contract which is signed between the providers and the users which guarantees for delivering the best service which has been agreed upon by both parties. It also contains a penalty clause which deals with the situation in the case of SLA violation. Various researches have addressed this issue which has been discussed below. A method has been introduced which provides a framework for resource provisioning which deals with the problem of under-provisioning and over-provisioning and reduces the number of SLA violations. It has introduced the autonomic provisioning based on the monitor, analysis, plan and execution of the loop. It also considers the cost and helps in minimizing it [16]. A novel method for resource utilization has been given which manages the resources automatically with the agreed SLA. A QoS based RM method is applied which provides self-configuration for resources, self-healing for handling the failures which occur as a sudden change, self-protection against security threats. It also deals with self-optimization to maximize resource utilization. It improves not only the rate of service level agreement violations but also cost and execution time [17]. The above discussed research work not only reduces the SLA violation rate but also considers various other parameters such as latency, cost and execution time. Thus, a multi-objective solution for RM should be given which can maximize the profit, reduce SLA violation rate and reduce energy consumption.

3.4. **Load balance based RM techniques.** Load balancing is a very important term that deals with the sharing of workload among different virtual machines in an optimized way. In this, the task is migrated from one machine to another for an efficient supply of services. This migration of task is governed by different load balancing algorithms. Some of the work in this field is being discussed below. A method has been proposed in which a cloud balancing load mechanism is given. The proposed mechanism can be applied on both the virtual web server and physical machines also [18]. The above study does not consider the processing time. To resolve this issue, a fuzzy-based multi-dimensional resource scheduling algorithm is proposed which schedules and optimizes the workload. It promotes the maximum utilization of resources and minimum processing time. Along with this, a queuing network method is applied for efficient load balance on scheduled resources [19]. With the above study, one can conclude that load balancing is the primary aspect of the efficient utilization of resources.

3.5. **RM techniques for mobile cloud.** Mobile cloud is a technology that deals with the integration of mobile computing with cloud services. With this technology, a user can access the cloud services on his mobile device. With the integration of multiple resources, various challenges have emerged. One of these challenges is to optimize the use of cloud-based resources usage. Various techniques have been proposed by several researchers in this regard. An algorithm has been introduced to optimize energy while performing wireless communication through a dynamic cloudlet-based model [20]. The above work does not consider any other aspect other than energy, so another energy-aware heterogeneous cloud management model has been proposed which not only considers energy but also looks for task mitigation. It applies the task assignment algorithm for assigning the task to mobile clouds and heterogeneous cores. It focuses on task mitigation and reduces energy consumption [21]. As the mobile devices have a constraint for power so to implement effective RM techniques such as load balancing or scheduling becomes a challenging task. It may also suffer various other cloud challenges such as latency, energy, migration, security, and scalability.

4. **Evaluation Parameters.** This section discusses the parameters used in the RM techniques as follows.

4.1. **Throughput.** In the cloud, the performance of the tasks can be evaluated with the help of a parameter known as throughput. It is defined as the number of tasks performed in a given interval of time [17]. If the throughput of the system is high, then it means more tasks are executed in less amount of time and results are being generated fast. The formula for its calculation is given as [17]

$$\text{Throughput} = \frac{\text{Total number of workloads}}{\text{Total time required to execute workloads}} \tag{1}$$

4.2. **Resource utilization.** Resource utilization is the parameter which deals with the utilization of the resources by the tasks. The high resource utilization tends to profit maximization and reduction in energy consumption as a reduction in the number of resources being utilized. Thus, resource utilization, RU can be given as [17]

$$RU = \sum_{i=1}^{n} \frac{\text{Time utilized by resources for workload execution}}{\text{Total uptime of resources}} \tag{2}$$

The rate of resource utilization is given as [19]

$$\text{Rate of Resource Utilization} = \mu_R(t) \cdot D \tag{3}$$

where $\mu_R(t)$ denotes the total demand of a class of user requests at a particular time interval t and D demotes the demand.

4.3. **VM migration time.** VM migration is the movement of the virtual machines from one hardware environment to another. Migration is performed to have load balance, workload consolidation, energy reduction, avoidance of SLA violation. The migration process requires some time which is known as migration time. This time should be a minimum for better performance of the system. It can be calculated as [22]:

$$T_{mk} = \frac{M_k}{B_k} \tag{4}$$

where T_{mk} stands for migration time for server k, M_k stands for memory used by the server k and B_k stands for available bandwidth for server k.

4.4. **Energy consumption.** The techniques devised for the management of resources should consume minimum energy. The latest research focuses on the minimum consumption of energy. Power consumption can be calculated as [23]

$$PC_j(t) = i \cdot PC_{jMax} + (1 - i) \cdot PC_{jMax} \cdot u_j(t) \tag{5}$$

where i is the fraction of power consumption when the host is in idle state, PC_{jMax} is the maximum power consumption of host, $u_j(t)$ is the CPU resource utilization of the physical machine on time t. Total energy consumption of a physical machine j in a time t can be given as:

$$E_j = \int_0^t PC_j(t)dt \tag{6}$$

4.5. **SLA violation.** In the cloud computing environment, a Service Level Agreement (SLA) is an agreement on the service requirement that is being agreed upon between cloud users and cloud providers. SLA violation is the breach in the agreement for which the party which has breached the agreement has to pay the fine to another party. The equation for SLA violation, S_v is given as [22]:

$$S_v = T_{AH} \cdot P_D \tag{7}$$

where T_{AH} stands for SLA violation time per active host and P_D stands for performance degradation due to the migration of the virtual machines. T_{AH} and P_D can be calculated as:

$$T_{AH} = \frac{1}{H} \sum_{j=1}^{H} \frac{T_{V_j}}{T_{A_j}} \tag{8}$$

where H is the number of hosts, T_{V_j} is the total time for SLA violation during which server j's utilization was 100%, and T_{A_j} is the total active time of the server j.

$$P_D = \frac{1}{N} \sum_{i=1}^{N} \frac{D_i}{C_i} \tag{9}$$

where N is the number of virtual machines, D_i is the estimated performance degradation of virtual machine i and C_i is the total capacity demanded by the virtual machine i during its execution time.

4.6. **Profit.** In the cloud computing environment, the maximization of profit and revenue generation is the major concern for cloud providers. The profit can be calculated as [24]:

$$\text{Profit} = \text{Total Revenue} - \text{Total Expenditure} \tag{10}$$

$$\text{Total Revenue} = \sum_{j=1}^{n} v_j \cdot r_j \tag{11}$$

$$\text{Total Expenditure} = c_v \sum_{j=1}^{n} v_j \cdot a_j \tag{12}$$

where v_j is the allocation vector given by user u_j and c_v is the cost incurred to run the allocation vector, r_j is the price paid by the user u_j for the allocation vector and a_j is the number of computing resources requested for each virtual machine by user u_j [25].

4.7. **Execution time.** Various techniques have been designed to reduce execution time with the help of resource scheduling, load balancing. The execution time can be calculated as [5]:

For homogeneous workloads,

$$\text{Execution Time } (T_E) = \text{Remaining Jobs} \cdot \text{Runtime of Jobs} \tag{13}$$

For heterogeneous workloads,

$$\text{Execution Time } (T_E) = \sum_{j=1}^{n} W_j \text{ Runtime} \tag{14}$$

where W_j Runtime stands for runtime for each workload j.

5. **Challenges in Resource Management.** Some of the major challenges of RM have been discussed here. These challenges demand detailed study and research which helps in the formation of an effective cloud computing system.

5.1. **Resource provisioning.** Resource provisioning deals with the allocation of resources of service providers to the consumer. The provisioning among different applications should maintain the elasticity of the application while considering application-specific SLA. The algorithm for provisioning should be designed in such a manner that it should handle the upcoming dynamic data for CPU allocation. A lot of work is done in this regard but still, a proper mechanism is required to develop a resource prediction model that can facilitate scaling concerning the variable needs of the users.

5.2. **Resource allocation.** Cloud providers need to give extra attention while allocating resources to the workload such as the resources should be allocated in such a way that it should not violate the SLA. One of the major concerns in this is the allocation and scheduling of resources efficiently to acquire the QoS goals as defined in SLA [7]. As the resource demands are quite dynamic, it becomes the key responsibility of cloud providers to observe and analyze the demands and allocate the resources accordingly. It demands an efficient algorithm as it affects cloud performance and cost.

5.3. **Resource scheduling.** Resource scheduling is the process in which it is determined which task or activity should be first performed based on the given SLA or QoS [5]. An efficacious scheduling algorithm is required to find the best cloud resources for the application or the workload so that it can lead to better resource utilization and thus enhances the resource usage ratio. There is a demand of expeditious and efficacious algorithm to tackle the variable workload which needs an efficient distribution of resources.

5.4. **Resource pricing.** Cloud price modeling is the biggest challenge and a term coined for this is "cloudonomics" [26]. A proper and efficient pricing model is required to maximize the profit of cloud providers while satisfying the end-user need. There are various stakeholders in the cloud network like end-users, service providers, and infrastructure providers. They have different goals like profit, investment, cost or revenue and different constraints like technology or budget [4]. These conflicting thoughts and values demand a good pricing model. Various services like video on demand need special pricing models as bandwidth is a very important aspect of this. In addition to higher bandwidth, Quality of Service (QoS) also plays a very important role. A good pricing model is the requirement

of cloud computing which enables the execution of the cloud industry in an economic manner.

5.5. **Energy.** The demand for an energy-efficient cloud is growing at a fast pace. Generally, the approaches which are undertaken in this are the identification of task which handles the optimization of the used hardware performance but hardware are not the only components that consume energy and moreover, it is not possible every time to replace the already running servers with the new servers which may use less energy [9]. Various load balancing and scheduling techniques should be introduced beforehand to suffice the demand for energy. This approach requires the migration and placement of virtual machines which may hinder the efficiency and throughput of the servers in addition to the quality of service constrained via SLA. The efficient approach is the demand of cloud computing which not only reduces the power consumption but also increases the efficiency and throughput with the required SLA.

5.6. **Security.** Security is a major concern in cloud computing. The security issues can be in communication level, computation level or SLA level. In the communication level, security problems can be in-network which generally addresses the confidentiality and integrity of data. In the application level, the application needs to be secured to avoid any change in it. At the host level, security is required by the operating system. Computational level includes security challenges in virtualization. Similarly, data also needs to be secured when they are in rest or stored in the data center so that it should remain confidential and maintain integrity [11]. Data needs security when it is communicated among cloud entities and it has to be performed with a secured communication channel. A secured mechanism is required for enhanced and reliable RM.

6. **Conclusions.** With the advent of utility computing, demand for cloud computing is increasing day by day in the professional and personal aspects of life. The major challenges should be addressed before its generalized or application-specific usage. The above study shows the recent development in the field of cloud computing which forms the basis for further research. It also segregates the RM techniques which help concern various evaluation parameters. These parameters may be either conflicting or non-conflicting. So a technique should be designed in such a manner that if it is complimenting one parameter, then it should not deteriorate other. Moreover, these techniques may be either a single objective or multi-objective. A multi-objective technique should be designed which should have the ability to manage the conflicting issues. Besides, the challenges and opportunities are identified which shows a vast scope of research in this field. Moreover, to implement an efficient private or hybrid cloud in the mobile devices has become a challenging task because of the infrastructure and architectural difference. These devices suffer from latency, bandwidth allocation, response time, throughput, security. So, a novel and efficient technique is required to tackle the mobile cloud issues with the kind consideration of RM parameters.

REFERENCES

[1] M. Armbrust et al., *Above the Clouds: A Berkeley View of Cloud Computing*, University of California, Berkeley, Technical Report, No. UCB/EECS-2009-28, pp.1-23, 2009.
[2] D. Marinescu, *Cloud Computing: Theory and Practice*, 2nd Edition, Elsevier, 2013.
[3] F. Nzanywayingoma and Y. Yang, Efficient resource management techniques in cloud computing environment: A review and discussion, *International Journal of Computers and Applications*, vol.41, no.3, pp.165-182, 2018.
[4] S. Chaisiri, B. S. Lee and D. Niyat, Optimization of resource provisioning cost in cloud computing, *IEEE Trans. Services Computing*, vol.5, no.2, pp.164-177, 2012.
[5] S. Singh and I. Chana, QRSF: QoS-aware resource scheduling framework in cloud computing, *The Journal of Supercomputing*, vol.71, no.1, pp.241-292, 2015.

[6] P. Pradhan, P. K. Behera and B. N. B. Ray, Modified round robin algorithm for resource allocation in cloud computing, *International Conference on Computational Modeling and Security*, vol.85, pp.878-890, 2016.
[7] C. Li, Y. C. Liu and X. Yan, Optimization-based resource allocation for software as a service application in cloud computing, *Journal of Scheduling*, vol.20, no.1, pp.103-113, 2017.
[8] H. Dai, Y. Yang, J. Yin, H. Jiang and C. Li, Improved greedy strategy for cloud computing resources scheduling, *ICIC Express Letters*, vol.13, no.6, pp.499-504, 2019.
[9] B. K. Dewangana, A. Agarwal, M. Venkatadri and A. Pasricha, Self-characteristics based energy-efficient resource scheduling for cloud, *Procedia Computer Science*, vol.152, pp.204-211, 2019.
[10] G. Kaur, A. Bala and I. Chana, An intelligent regressive ensemble approach for predicting resource usage in cloud computing, *Journal of Parallel and Distributed Computing*, vol.123, pp.1-12, 2019.
[11] A. Kumari, J. K. R. Sastry and K. R. Rao, Energy efficient load balanced optimal resource allocation scheme for cloud environment, *International Journal of Recent Technology and Engineering*, vol.8, pp.146-153, 2019.
[12] P. Samimia, Y. Teimourib and M. Mukhtara, A combinatorial double auction resource allocation model in cloud computing, *Information Sciences*, vol.357, pp.201-216, 2016.
[13] B. Yang, Z. Li, S. Jiang and K. Li, Envy-free auction mechanism for VM pricing and allocation in clouds, *Future Generation Computer Systems*, vol.86, pp.680-693, 2018.
[14] D. M. Bui, Y. Yoon, E. N. Huh and S. Jun, Energy efficiency for cloud computing system based on predictive optimization, *Journal of Parallel and Distributed Computing*, vol.102, pp.103-114, 2017.
[15] M. Ranjbari and J. A. Torkestani, A learning automata-based algorithm for energy and SLA efficient consolidation of virtual machines in cloud data centres, *Journal of Parallel and Distributed Computing*, vol.113, pp.55-62, 2018.
[16] M. G. Arani, S. Jabbehdari and M. A. Pourmina, An autonomic approach for resource provisioning of cloud services, *Cluster Computing*, vol.19, no.3, pp.1017-1036, 2016.
[17] S. S. Gill, I. Chana, M. Singh and R. Buyya, CHOPPER: An intelligent QoS-aware autonomic resource management approach for cloud computing, *Cluster Computing*, vol.21, no.2, pp.1203-1241, 2018.
[18] S.-L. Chen, Y.-Y. Chen and S.-H. Kuo, CLB: A novel load balancing architecture and algorithm for cloud services, *Computers & Electrical Engineering*, vol.58, pp.154-160, 2017.
[19] V. Priya, C. S. Kumar and R. Kannan, Resource scheduling algorithm with load balancing for cloud service provisioning, *Applied Soft Computing Journal*, vol.76, pp.416-424, 2019.
[20] K. Gaia, M. Qiu, H. Zhao, L. Tao and Z. Zong, Dynamic energy-aware cloudlet-based mobile cloud computing model for green computing, *Journal of Network and Computer Applications*, vol.59, pp.46-54, 2016.
[21] K. Gai, M. Qiu and H. Zhao, Energy-aware task assignment for mobile cyber-enabled applications in heterogeneous cloud computing, *Journal of Parallel and Distributed Computing*, vol.111, pp.126-135, 2018.
[22] S. Mustafa, K. Bilal, S. U. Rehman Malik and S. A. Madani, SLA-aware energy efficient resource management for cloud environments, *IEEE Access*, vol.6, pp.15004-15020, 2018.
[23] W. Zhu, Y. Zhuang and L. Zhang, A three-dimensional virtual resource scheduling method for energy saving in cloud computing, *Future Generation Computer Systems*, vol.69, pp.66-74, 2017.
[24] N. C. Luong, P. Wang, D. Niyato and W. Yonggang, Resource management in cloud networking using economic analysis and pricing models: A survey, *IEEE Communications Surveys & Tutorials*, vol.19, no.2, pp.954-1001, 2017.
[25] S. Zaman and D. Grosu, A combinatorial auction-based mechanism for dynamic VM provisioning and allocation in clouds, *IEEE Trans. Cloud Computing*, vol.1, no.2, pp.129-141, 2013.
[26] J. Weinman, Cloudonomics: A rigorous approach to cloud benefit quantification, *Journal of Software Technology*, vol.14, no.4, pp.10-18, 2011.