Resource Allocation Techniques in Cloud Computing: A Review and Future Directions

Muhammad Faraz Manzoor[1, *], Adnan Abid[1], Muhammad Shoaib Farooq[1], Naeem A. Nawaz[2], Uzma Farooq[1]
[1]Department of Computer Science, University of Management and Technology,
Lahore, Pakistan
[2]Department of Computer Science, Ummal Qura University,
Makkah, Kingdom of Saudi Arabia
F2018288004@umt.edu.pk

Abstract—**Cloud computing has become a very important computing model to process data and execute computationally concentrated applications in pay-per-use method. Resource allocation is a process in which the resources are allocated to consumers by cloud providers based on their flexible requirements. As the data is expanding every day, allocating resources efficiently according to the consumer demand has also become very important, keeping Service Level Agreement (SLA) between service providers and consumers in prospect. This task of resource allocation becomes more challenging due to finite available resources and increasing consumer demands. Therefore, many unique models and techniques have been proposed to allocate resources efficiently. In the light of the uniqueness of the models and techniques, the main aim of the resource allocation is to limit the overhead/expenses associated with it. This research aims to present a comprehensive, structured literature review on different aspects of resource allocation in cloud computing, including strategic, target resources, optimization, scheduling and power. More than 50 articles, between year 2007 and 2019, related to resource allocation in cloud computing have been shortlisted through a structured mechanism and they are reviewed under clearly defined objectives. It presents a topical taxonomy of resource allocation dimensions, and articles under each category are discussed and analysed. Lastly, salient future directions in this area are discussed.**

Index Terms—**Cloud computing; Resource allocation; Resource scheduling; Resource utilization.**

I. INTRODUCTION

Cloud computing has risen as a modern day technology that is based on service oriented architecture so as to provide infrastructure, platform, and software as a service. Resource allocation in cloud computing is about designating the processing tasks to a pool of resources in the cloud infrastructure, which consists of a number of computers. The aim of this contemporary technology is to facilitate the clients with services under pay-per-use payment method. As a new technology, it is facing difficult challenges that need a clear depiction of activities and relationships, keeping in mind the end goal to encourage the strategic advancement and use of cloud computing. Technically, it is a mix of server virtualization technology and other resources alongside different technologies [1].

Resource allocation in cloud computing involves the scheduling and resource provision while keeping in view the available infrastructure, service level agreements, cost, and energy factors. For instance, cloud service provider manages the resources according to the on demand pricing method while ensuring the great Quality of Service (QoS) and user satisfaction [2]. Similarly, the resources have to be assigned in a way that every application gets the required resources without exceeding the limit of cloud environment. In the same way, resource allocation is responsible for handling the issue of the starving of applications by proper resource allocation by enabling the service providers to allocate the resources for each individual module [3].

Cloud computing provides high quality services to the consumer at a very low cost [4]. While, in terms of storage, the data centers give a lot of resources and distributed computing models ready to help on request resources allotment, which prompts the non-ideal resource assignment. Another issue that large data centers face is energy utilization. It has been seen that vitality devours over 20 % of the vast data centers. Reduction in energy utilization can spare resources supplier a major amount of energy and cost [5]. The simplest and efficient way of doing that is to use the hardware resources in an elastic manner and turn off the servers that are not being used. However, this requires a careful planning so that data centers do not run out of resources as requests arrive.

The main aim of this survey is to provide a summary of the resource allocation techniques in cloud computing. So, the four research questions have been developed, which are defined in Table I.

We believe that there exists no comprehensive research study that particularly covers the cloud resources allocation techniques using a topical taxonomy with characteristics as of strategic, target resources, optimization, scheduling and power. Consequently, the mentioned characteristics will assist the authors to know about the variety of information in this domain of study. Articles from different conferences, workshops, and journals were selected on the basis of ranking of their respective journals, conferences, and

Manuscript received 9 March, 2020; accepted 28 June, 2020.

symposiums for detailed review and analysis.

TABLE I. RESEARCH QUESTIONS.

Questions	Motivations
What is the significance of the allocation of resources in cloud computing?	It enables the cloud service providers to manage the resources for each individual module.
What are the existing techniques to allocate resources in cloud environment?	There are many techniques which guarantees to allocate the resources efficiently in cloud computing.
What are the parameters and resources have been considered more during resource allocation?	It analyses important parameters and resources for the service consumer and service providers during allocation of resources.
What are the research gaps in resource allocation in cloud computing domain?	With the help of this review paper the future researchers will clearly understand the need and requirements in future for allocation of resources in cloud computing.

The rest of the article is organized as follows. Section II describes the allocation of resources in cloud computing environment. The taxonomy and subdivisions of the presented techniques in detail are discussed in Section III. Whereas, Section IV discusses the future directions in this area. Lastly, Section V presents the conclusions of the article.

II. RESOURCE ALLOCATION IN CLOUD COMPUTING

Allocation of the resources is the process in which the proper resources are allocated to the tasks required by the consumer so that these tasks are finished proficiently. In cloud computing, it implies designating a virtual machine fulfilling the properties defined by the consumers. Users should submit their task which may have its own time imperatives. The viable way in which these workloads can be allotted to the virtual machines and handled is another type of resource allocation possible technique in the cloud [6]. Simply it is all about defining when a computational action should begin or finish dependent upon: 1) resources assigned, 2) time taken, 3) predecessor actions, and 4) predecessor relationships. In addition, resource allocation in cloud computing includes the resource disclosure, choice, provisioning, application planning, and administration of resources. It involves the decision making of when, what, where, and how much resources should be allocated to the consumer as shown in Fig 1.

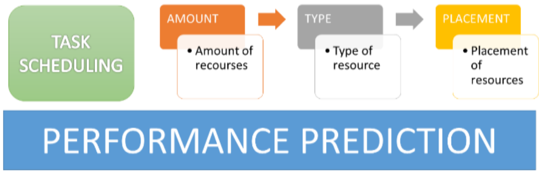

Fig. 1. Cloud resource allocation basic elements.

Figure 2 shows that the allocation of resources is done in the following steps generally: 1) consumer will submit the request to the resource allocator, 2) the request will be added to the queue list, 3) resource allocator informs the allocation unit about the request, 4) allocation unit asks for the requested resources from the Infrastructure as a Service (IaaS), 5) if the resources are available, then IaaS respond back positively, 6) the allocation unit creates a Virtual Machine (VM) from VM pool according to the request, 7) resource allocator is informed after creating a VM, 8) requests are de-queued, and 9) resources are allocated.

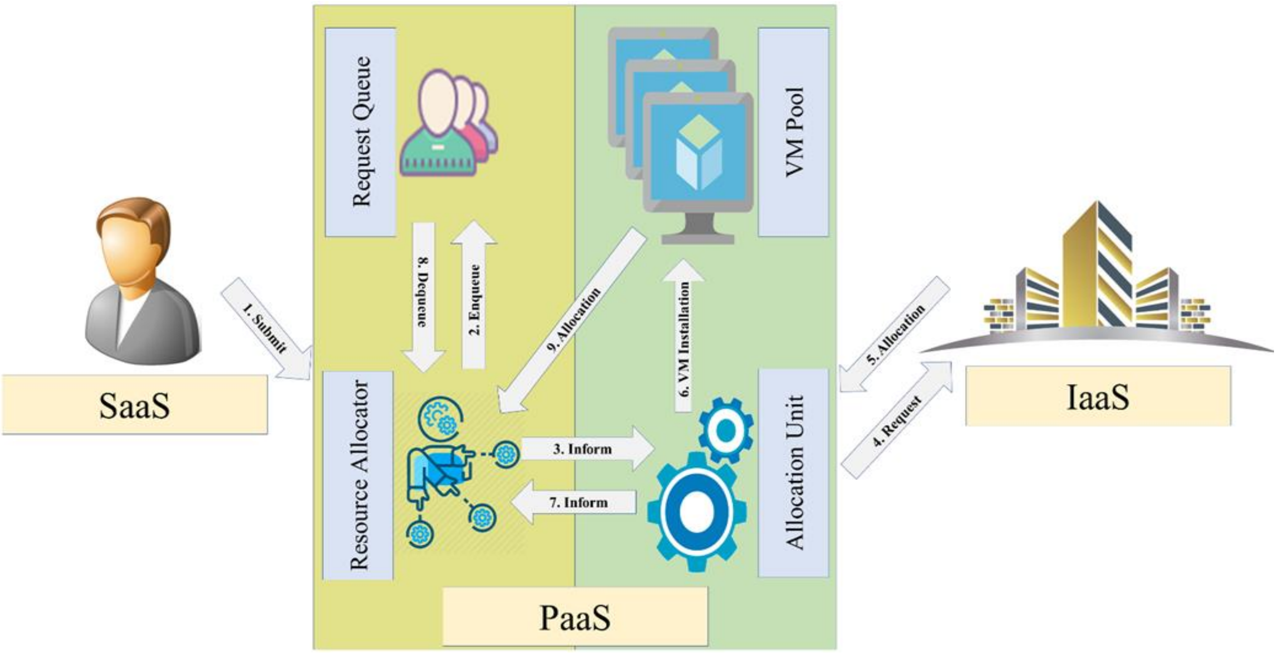

Fig. 2. The basic flow of resource allocation in cloud computing.

One other difficulty in allocating the resources is that both cloud service providers and cloud service consumers do not share their information. For example, cloud service provider will not expose the number and type of resources he has, as such information is not allowed to share in business departments. On the other side, cloud service consumers do not show the type and workload of their application to others, including service providers. For the optimization of the resource allocation for the consumer's request, the cloud service consumers do not publicize his/her request, as they do not precisely recognize what is accessible. Additionally, cloud service providers cannot allot assets in a manner most reasonable to cloud service consumer's applications, in light of the fact that there are no or few insights about their workload patterns.

One of the main challenges in cloud computing is allocation of the resources among the users based on their application usage patterns. These unpredictable requests will run on data center over the internet. Few challenges related to resource allocation in cloud computing that we found are as follows:

– From the cloud service provider point of view, prediction of the consumer's and application demands are very difficult. At the same time, from cloud service consumer's point of view, the job should be completed on time. So, due to limited resources, available and efficient resource allocation techniques are required that support the cloud environment.

– The capability of physical machines should be adequate enough to fulfil the resource needs of all virtual machines.

– Applications running on the VM, and at the same time the consumers, need the networking services with efficient QoS to guarantee the effective delivery of their application data.

– Auction-based resource allocation also faces the challenge of providing the mechanism that proves to be suitable, i.e., the allocation of resources is efficient and its pricing is beneficial for both cloud service provider and cloud service consumer.

– The consumer's Service Level Agreement (SLA) violations need to be minimized while maximizing the utilization of resources because in most of the cases the QoS affects resource allocation techniques in order to minimize the cost.

From the above discussion, the conclusions can be made that the one must take the qualities and attributes of the both components of the cloud computing in consideration to give proficient cloud services and cloud-based application, i.e., reasonable resources are assigned to a suitable application at a fitting time, with the end goal that applications can use the resources successfully.

III. Resource Allocation Techniques

The resource allocation techniques utilize various methodologies for the productive usage of resources to fulfil the consumer requirement. In cloud computing, the resource allocation techniques can be categorized into 1) strategic: satisfying the consumer's ever changing demands, 2) target resources: focusing mainly on requested resources, 3) auction: bidding for the resources, 4) optimization: optimizing the resources, 5) scheduling: prioritizing the task for better performance and 6) power: better resource allocation with less power consumption, as shown in Fig 3. This categorization is further divided into following subheadings. The mentioned techniques are then evaluated on the bases of the following parameters.

There are some important parameters for both cloud service provider (cost, resource utilization, energy, workload, SLA, QoS) and cloud service consumer (execution time, response time, user satisfaction, SLA, QoS) perspectives which should be considered to in the development of the resource allocation techniques:

Cost: One of the most important parameters for cloud service provider which eventually determines whether cloud service provided is bearing high or low cost for providing different services. It is pertinent to mention here that in this article the parameter cost is only for service provider, not for service consumer.

Resource Utilization: All the cloud service providers want to utilize their resources in an optimized manner so that resources do not remain idle. It is important to mention that the efficient resource allocation is effective for environmental safety and also reduces the overall expenditure of data centers.

Power: Power is another important dimension in resource allocation in cloud computing. Energy crisis is increasing day by day, therefore minimizing the utilization of power and energy had become main concern to make cloud services environmentally supportable.

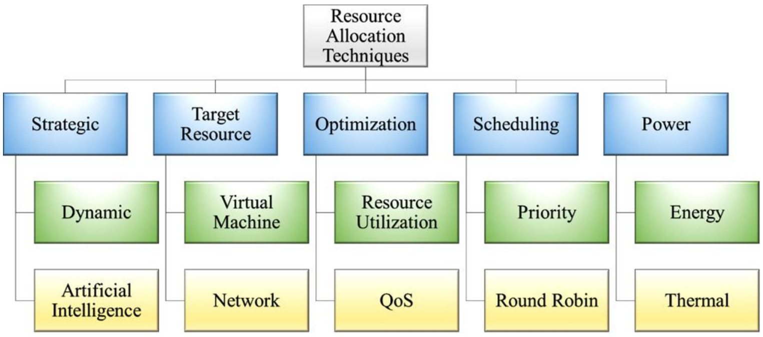

Fig. 3. Taxonomy of resource allocation in cloud computing.

Workload: Workload generally reflects the ability of the system to handle and process the task. Workload should be enough on a system to do the tasks efficiently within the cloud environment. This parameter will determine the amount of workload on empirical setup of resource allocation techniques.

Execution Time: Both cloud service provider and cloud service consumer want the minimum execution time of the task. But, the execution of multiple workloads on single resource will produce interference among these workloads, which leads to poor performance.

Response Time: It is the time taken by the system to answer the request. From the cloud service consumer's prospective, it should be as low as possible. Response time is an important parameter to measure the system performance. Low response time is critical for successful computing.

User Satisfaction: It is the user's comfort towards the cloud service provider. Every cloud service provider wants to satisfy his consumers in every way. By effective allocation of resources in cloud computing, the revenue and user satisfaction can be maximized.

Some of the other parameters are QoS of both cloud service provider and cloud service consumer SLA, fraud preventions, and revenue.

The value of given parameters is represented by the numbers from 1 to 5 where 5 is the highest value and 1 is the lowest value, but high value of parameter is not ideal in every case, e.g., in case of cost, response time, execution time, workload and power, the ideal value should be as low as possible. Similarly, in case of user satisfaction, SLA, resource utilization, fraud prevention, and revenue, the ideal value should be as high as possible.

Consistency Protocol: The following ingredients have been used for the discussion of every feature: 1) feature definition, 2) discussion on considered articles, and 3) valuation of the feature based on the above-mentioned parameters.

A. *Strategic*

There is a great development in the adoption of cloud computing because of the promising features for both service consumers and service providers. To satisfy the consumers every day, new requirements for the strategic-based resource allocation techniques have been utilized. The strategic resource allocation is further categorized in: 1) dynamic resource allocation: technique is used by the cloud service provider to predict the nature of consumers, their demands, 2) artificial intelligence: imitate nature to schedule tasks among resources.

1. *Dynamic*

The workload of the task submitted to the cloud infrastructure by the consumers changes continuously. In order to address the individual requirements of each task, cloud service provider has to use dynamic techniques to allocate resources. There are few techniques which use leasing as a fundamental resource provisioning for advance reservation, e.g., a technique for predicting varying run-time overheads associated with utilizing virtual machines was proposed in [7], enabling us to proficiently support advance reservations. In addition, Ch. Li and L. Y. Li [8] proposed technique in which the resources of cloud service provider were leased by the Software as a Service (SaaS) provider, and also SaaS was leased to consumers. The aim of the SaaS providers is to reduce the resources cost they leased from resource providers and maximize the profit they earned from the consumers. Strategies that determine the proper set of lease(s) for preemption to minimize the effect of pre-empting virtual machine have been proposed in [9].

On the other hand, there is a work for dealing with the high priority tasks first. A. T. Saraswathi, Y. R. A. Kalaashri, and S. Padmavathi Dr. [10] introduced a technique for the tasks of high priority. This technique ignores the formation of the latest virtual machines to implement the recently arrived jobs. The proposed approach suspends a low priority job to complete the high priority job. Subsequently, after the completion of high priority job, it resumes the suspended job on any of the virtual machine. This strategy has minimal overhead to execute all jobs creating and comparing another virtual machines.

2. *Artificial Intelligence*

Strategic-based resource allocation techniques are greatly influenced by artificial intelligence. In cloud computing, the artificial intelligence is an area that concentrates on developing the techniques that act and work like humans for resource allocation. There is an algorithm named "Fuzzified Genetic Algorithm" that combines basic artificial intelligence models to achieve better results. M. Shojafar, S. Javanmardi, S. Abolfazli, and N. Cordeschi [11] proposed a technique where they combined fuzzy model and genetic algorithm and introduced a new technique called "FUGE". They use fuzzy logic to find the most suitable task by representing tasks as chromosomes and genes. Similarly, the authors in [12] utilize fuzzy logic to allocate resources to the consumer's task. The consumer tasks are categorized based on different parameters like memory, expected time, and bandwidth. The resources are also categorized based on disk space, network bandwidth, and CPU cycle. The input values are then converted within the range from 0 to 1 through fuzzification. After that, they are submitted to a neural network, which comprises of three layers: 1) input layer, 2) hidden layer, and 3) output layer. The neural network determines the mappings of the cloud resources to the consumer tasks. The fuzzy range values are changed into their original values in defuzzification phase. This proposed technique has improved the overall performance of the system.

Moreover, Ant Colony Optimization is used in the existing techniques for resource allocation and optimization. The authors in [13], satisfy the demands of cloud computing infrastructure that estimate the required bandwidth by predicting the available resources in advance. Ch. Li and L. Li [14] proposed an iterative technique for resource allocation for both SaaS and IaaS. This optimization technique for efficient resource allocation is compared with different existing algorithms that showed improved overall performance of the system.

The analysis in Table II shows that in the context of dynamic resource allocation techniques A. T. Saraswathi, Y. R. A. Kalaashri, and S. Padmavathi technique [10] is relatively efficient as it is less costly, utilization of resources is high, workload and execution time is low. On the other hand, in the context of artificial intelligence-based resource allocation, L. Ying, Q. P. Rui, and X. Jie [13] proposed a technique that is more efficient than the other techniques as consumption of energy, workload, and execution time is low.

TABLE II. COMPARISON OF STRATEGIC RESOURCE ALLOCATION TECHNIQUES.

Dynamic Resource Allocation					
Reference No.	Cost	Energy	Resource Utilization	Work load	Execution time
[7]	4	-	4	4	2
[8]	1	-	-	3	3

[9]	-	3	4	4	-
[10]	2	-	4	1	1
Artificial Intelligence Resource Allocation					
Reference No.	Cost	Energy	Resource Utilization	Work load	Execution time
[11]	-	4	3	-	-
[12]	2	-	3	-	4
[13]	-	2	-	1	2
[14]	1	-	4	-	5

B. *Target Resource*

The target resource type characteristic shows the central resource for which the technique is originally developed. The target resource allocation attributes determine the type and level of resource allocated to consumer task. Two types of target resource allocation-based techniques are discussed in this paper: 1) Virtual Machine and (2) Network. The present resource allocation techniques allocate the resources to the task at various granularity levels. Likewise, (1) virtual machine allocation shows the virtual machine’s position on physical machine. Moreover, (2) network failure in cloud data center could happen due to the inefficient resource allocation, logical segmentation of physical machines, and scheduling.

1. *Virtual Machine*

Virtualization technology is the main reason that makes the cloud computing affordable and practical. Cloud computing resources need high investments as they are very costly, so the cloud service provider wants to adopt the techniques that allocate the virtual machine to the task efficiently. Few techniques have been proposed to meet the job requirements and avoid the delay, e.g., T. S. Somasundaram *et al.* [15] introduced a novel technique named “Care Resource Broker” (CRB) that fulfils the task requirement. The proposed technique improves the reaction time, throughput and highlights the reasons of task scheduling failure because of unavailability of needed computing resources. CRB executes services to outline and administration of VM-based resources and satisfies the consumer’s task requirements by deploying a required number of resources.

Likewise, J. Machina and A. Sodan [16] proposed a technique that describes the execution of the tasks as a component of allocated cache size, regardless of whether the cache is dynamically partitioned. On the other hand, there have been few techniques that used artificial intelligence concepts to allocate the virtual machines to the consumer’s task. S. Kundu, R. Rangaswami, A. Gulati, M. Zhao, and K. Dutta [17] proposed a new system in which they use neural networks to show application execution in virtualized shared frameworks utilizing numerous resource nods. Likewise, in [18], the authors use machine learning to show the execution of application based on low level matrices in order to find the most suited configuration to maximize the throughput.

Ch.-H. Lin, Ch.-T. Lu, Y.-H. Chen, and J.-Sh. Li explored the behavior of the virtual machine according to their actual service for resource usage optimization in [19]. For the solution of the virtual machine allocation problem, an ideal threshold was introduced according to the probability distribution function so that for every virtual machine the total CPU utilization is reduced and the VM execution SLA is accomplished. On the other hand, Sh. Zaman and D. Grosu [20] proposed an auction-based VM allocation technique in which they use combinational auction that considers the consumers’ requests while provisioning decisions made. This method proved to be an efficient method because the resources will be matched to the consumers having a highest bid limit.

2. *Network*

According to the authors in [21], just 54 % of the IT expert overview about utilization of cloud services demonstrated that they include network operation personnel, which influences the utilization of system best practices and consideration regarding the health of overall traffic delivery. 28 % of review respondents trusted that troubleshooting and monitoring packet traces between VMs are needed, also 32 % trusted that troubleshooting and monitoring traffic data from virtual switches are required. With the passage of time, researchers realized that there is also a need of network aware resource allocation techniques to optimize the allocation of the resources to the consumer’s task. Like the authors solve the problem in [22] where the consumer processed the multiple jobs simultaneously on different servers. Connection requests are represented as a virtual network where nodes are virtual machines and edges are physical network paths. The objective of maximizing the profit was achieved by resolving the optimization issue of the virtual network. The issue of offering the best virtual network with IP over a wavelength-division multiplexing (WDM) was considered in [23]. The authors proposed their limitations according to the capacity, flow conversation constant, and propagation delay.

Most of the techniques based on network aware resource allocation focus on resources and communication within the data centers. X. Meng, V. Pappas, and L. Zhang [24] proposed a new network aware virtual machine placement technique with the aim to reduce total network traffic by optimal virtual machine placement. To reduce the cost and network traffic, they allocate virtual machines with large communication near each other. Similarly, J. Dong *et al.* [25] proposed a network aware virtual machine allocation in reflection of multiple resource requirements in the data center to minimize the energy consumption of the data center. The two types of resources are considered for the virtual machine placement at the same time: 1) the network resources optimization and 2) the physical resource allocation.

Table III shows that the virtual machine-based resource allocation technique proposed by T. S. Somasundaram *et al.* [15] is more suitable as it cost, response time, and execution time of a task is low, whereas resource utilization and user satisfaction is high. Subsequently, in the context of network, the technique proposed by G. Sun, V. Anand, H.-F. Yu, D. Liao, and L. Li [22] is relatively efficient as the cost, workload, and execution time is low and also resource utilization is high.

TABLE III. COMPARISON OF TARGET RESOURCE ALLOCATION TECHNIQUES.

Reference No.	Cost	Energy	Workload	Resource Utilization	Response Time	Execution time	User Satisfaction
Virtual Machine							
[15]	2	-	-	5	1	2	5
[16]	4	3	-	-	-	2	3
[17]	4	-	-	-	3	1	2
[18]	4	-	3	-	1	-	1
[19]	1	1	-	4	2	-	2
[20]	2	-	2	5	-	-	2
Network Resource Allocation							
Reference No.	Cost	Energy	Workload	Resource Utilization	Response Time	Execution time	User Satisfaction
[22]	1	-	2	5	3	2	3
[23]	2	-	4	1	4	-	1
[24]	1	3	2	-	2	3	2
[25]	-	1	-	2	1	-	2

C. *Optimization*

The main aim of optimization is to improve the throughput by increasing the use of physical and virtual resources. This will result in the cloud service providers to maximize their profits by facilitating maximum consumers and lessening operational uses by affiliating the workload on lesser machines. The current Resource Allocation Techniques (RAT) target the multiple optimization objectives, such as 1) Resource Utilization: efficient resource usage for safety of the environment and decreases in datacenter operational consumption and 2) The Quality of Service (QoS): targets the accomplishment of consumers' multiple satisfaction matrices, such as latency (delay in communication), CPU speed, stability, memory, etc. Non-acquiescence with execution measurements can increase violation of service performance levels. SLA between cloud service providers and cloud service consumers define the Quality of Service details.

1. *Resource Utilization*

As the use of cloud computing is increasing day by day, the load on servers has also increased. The primary point of researchers has been to utilize maximum resources while keeping up low power consumption. One of the most popular algorithms to optimize the resources utilization is generic algorithm. X. Lu, J. Zhou, and D. Liu [26] proposed an enhanced versatile generic algorithm to find the suitable solution for all the tasks taking place in real time. Similarly, CPU usage with a generic algorithm to upgrade the resources of the virtual machine was analysed in [27]. On the other hand, less consumption of power is also considered for better utilization of resources. For instance, R. Lee and B. Jeng [28] proposed a theoretical solution which offers a technique for dynamic volume provisioning which lower the energy consumption cost. The simulation depends on real time indications and recommendations acquired from sites, e.g., Google. Moreover, power of server can be managed at the ensemble layer. Subsequently, equal distribution of workload can also optimize the resource utilization. Z. Abbasi, G. Varsamopoulos, and S. K. S. Gupta [29] proposed an algorithm for effective distribution of workload. In the proposed method, users select the active servers and a particular threshold is set so that it does not ascend past a specific incentive in server to ensure the right activities [30], [31]. The authors represented the Electronic Eligibility Verification Service (EEVS) and Dynamic Voltage and Frequency Scaling (DVFS) in [32] whose main focus is to reduce the whole energy consumed in a cloud during resource utilization, but the algorithm compromised the time duration and power consumption.

On the other hand, there are also few techniques with the aim of better virtual machine utilization and placement. A. Wolke, B. Tsend-Ayush, C. Pfeiffer, and M. Bichler [33] proposed an active, but stationary "Bin packing" heuristic technique which made a positive impact on resource utilization, but could not cover the overloading due to migration.

2. *Quality of Service*

The violation of SLA can make a great impact on the satisfaction level of cloud service consumers. SLA is a contract that regulates the QoS between the cloud service provider and cloud service consumer. It also includes the service cost with the level of QoS balanced by the cost of the service [34]. The cloud service provider should develop its system to fulfil QoS demands of every component in the cloud [35], [36]. There are some QoS-based cloud service consumers-oriented resource allocation techniques that try to fulfil their requirements and some of the techniques are cloud service provider oriented, which try to fulfil their requirements which could make a negative effect on cloud service consumer satisfaction. In [37], the authors focused on the QoS parameters, such as Cost and satisfaction of consumer, but did not focus on the QoS requirements cloud service consumers. On the other hand, few resource allocation techniques focused on the satisfaction of both cloud service provider and cloud service consumer. Such technique was proposed by L. Wu, S. K. Garg, and R. Buyya [38]. They focused on QoS perimeters of both cloud service provider and cloud service consumer through with goal of limiting the SLA violations and infrastructure price. The technique shows good results by decreasing price 50 % with utilization of fewer virtual machines and optimizing the means to limiting the SLA violations. V. C. Emeakaroha, I. Brandic, M. Maurer, and I. Breskovic [39] proposed a scheduling-based heuristics resource allocation technique in which they used to avoid SLA violation penalties by utilizing different SLA parameters, such as quality, availability and responsibilities, for the development of the application. There are restricted applications in real world system based on the parameters they considered in their research, while consumers would be keener on execution parameters, e.g., reaction time and handling time.

There are few techniques which consider the SLA parameters for the allocation of the resources. A. Kumar, E. S. Pilli, and R. C. Joshi [40] proposed a technique named “EARA”. It is an efficient agent-based resource allocation technique that considers various SLA parameters, such as memory, bandwidth, and execution time in which few agents gather accessible resource data to allocate it according to the consumer’s demands depending on the SLA arrangement.

The analysis in Table IV shows that in the context of maximum resource utilization techniques the proposed technique in [33] is more efficient as its cost, energy consumption is low, whereas resource utilization is high. On the other hand, the proposed technique in [38] is proved to be efficient as demands of both Cloud Service Provider (CSP) and Cloud Service Consumer (CSC) are high and execution time is low.

TABLE IV. COMPARISON OF OPTIMIZATION RESOURCE ALLOCATION TECHNIQUES.

Resource Utilization					
Reference No.	Cost	Energy	Resource Utilization	Workload	Execution Time
[26]	4	-	4	2	-
[27]	3	2	4	-	2
[28]	1	1	5	-	5
[29]	3	2	5	-	4
[32]	-	2	5	-	5
[33]	1	2	4	-	3
Quality of Service(QoS)					
Reference No.	Cost	QoS demands of CSP	Resource Utilization	QoS demands of CSC	Execution time
[37]	1	3	-	2	-
[38]	-	4	2	4	2
[39]	1	4	-	4	-
[40]	2	5	-	2	-

D. *Scheduling*

Resource scheduling is critical research area in cloud environment because of the substantial resource cost and execution time. Various resource parameters and scheduling criteria are considered in various categories of resource scheduling techniques. Resource scheduling becomes more difficult because both cloud resource provider and cloud service consumer do not want to share their data with each other. To schedule and execute the workloads, cloud service providers also consider unpredictable resources. This paper categorized resource scheduling techniques in two subdivisions: 1) Priority base: schedule the resources based on different parameters like memory, network bandwidth, and required CPU time; 2) Round Robin: RR works using time slice which can be explained as a small entity of time.

1. *Priority*

Cloud service provider determines the priority among the different consumer demands while allocating the resources in cloud computing. The parameters that considered in priority-based resource allocation are time, cost, no of processor request. For instance, Sotomayor, R. S. Montero, I. M. Llorente, and I. Foster [41] proposed a virtual aware resource allocation technique for parallel workloads to make the response time more efficient. In proposed technique, virtualization technology splits the computing capacity into the following two levels: Virtual machine with high CPU priority and Virtual machine with low CPU priority. Furthermore, Ch. S. Pawar and R. B. Wagh [42] introduced a new SLA-based technique to schedule the resources dynamically that considers SLA parameters like memory utilization, processor time, and network bandwidth to enhance resource utilization and reduce resource contention. Zh. Lee, Y. Wang and W. Zhou [43] proposed a scheduling algorithm named “dynamic priority scheduling algorithm” (DPSA) to solve scheduling problem in service request. In DPSA, consumer tasks are categorized on the bases of their specific requirements into task units after they are received and analysed to schedule specifically and give the productive service.

Few techniques used the priority-based resource allocation to reduce the overhead and completion time. X. Wu, M. Deng, R. Zhang, B. Zeng, and Sh. Zhou [44] introduced QoS aware task scheduling technique which considers the task priority to decrease the completion time. The proposed technique processes the priority based on task properties and stores it in a task queue based on recognized priority. It identifies the execution time of each task and resources are allocated to task which takes the minimum time.

2. *Round Robin*

In recent times, Round Robin has been the most common and widely used scheduling algorithm which makes it suitable to allocate resources to the task efficiently. The RR algorithm is developed for the allocation of the time of CPU among the arranged tasks. Similarly, with every task lineup in a queue list, whereas CPU time is distributed among the tasks. Few researchers have worked on the improvement of the CPU response time. A. Abdulrahim, S. Abdullahi, and J. B. Sahalu [45] introduced a round robin-based technique, which allots the time quantum to each task on the ready queue by calculating a dynamic time quantum of jobs based on the average burst time on a queue list. This technique reallocates the essential time to complete the task in case the time quantum is not enough for completion. This mechanism improved the system performance by minimizing the task’s context switches. Similarly, M. Mishra and A. Khan [46] introduced a Round Robin-based technique, which compares the algorithm static time quantum to the remaining burst time of a task to the after the first allocation. The CPU will reallocate the time quantum to the task if the remaining time is less than one-time quantum, otherwise it is sent to a waiting queue.

Similarly, .S. N. Rao, N. Srinivasu, S. V. N. Srinivasu, and G. R. K. Rao [47] introduced a Round Robin-based technique that calculates the time quantum from the average burst time of tasks on the ready queue that can be adjusted dynamically to the tasks that need more time than the allotted time quantum. Furthermore, A. Noon, A. Kalakech, and S. Kadry [48] proposed a new Round Robin-based technique that calculates the dynamic time quantum without any task arrangement depending on the tasks arrival time in the ready queue and assigns the time quantum. Subsequently, in [49], the authors introduced a Round Robin-based technique that calculates the time quantum dynamically based on Priority else the algorithm executed the task according to the Shorter Job First technique.

Table V shows that a priority-based resource allocation technique proposed in [43] is efficient as its cost, response time, and workload are low, whereas resource utilization and user satisfaction are high. Subsequently, in the context of round robin algorithm-based techniques, the technique presented in [46] is more suitable as resource utilization and user satisfaction are high and response time and execution time are low.

TABLE V. COMPARISON OF SCHEDULING BASED RESOURCE ALLOCATION TECHNIQUES.

Priority							
Reference No.	Cost	Response time	Resource Utilization	Workload	Execution time	SLA	User Satisfaction
[41]	-	2	-	-	-	2	3
[42]	4	1	5	2	-	-	2
[43]	2	2	4	2	-	-	5
[44]	-	2	-	4	1	2	4
Round Robin							
Reference No.	Cost	Energy	Resource Utilization	Response time	Execution time	SLA	User Satisfaction
[45]	4	-	-	2	1	1	4
[46]	4	2	4	2	2	-	5
[47]	-	-	2	1	2	-	3
[48]	2	1	-	2	1	2	4
[49]	3	-	-	1	1	-	3

E. *Power*

Power consumption has become a critical aspect as hosting centers and data centers consume significant amount of power. Without any proper resource utilization technique, poor resource utilization and hotspot problems may occur. Proper resource allocation techniques can not only reduce the power consumption, but also save operational cost. In the paper, it is extensively characterized into two subdivisions, i.e., 1) energy-aware: expect to expand benefit level execution measures under power dissemination and power utilization requirements and 2) thermal-aware allocation: predicts the thermal impacts of a task placement and the resource allocation depending on the anticipated thermal impact.

1. *Energy Aware*

Energy consumption and resource allocation in cloud computing are very interrelated. Efficient resource allocation is projected to produce financial advantages, as well as the environmental harmony. Energy aware resource allocation techniques have become very successful in managing the problems due to power utilization in data centers. Few techniques have managed to decrease the power consumption by proper placement of VMs. For instance, S. E. Dashti and A. M. Rahmani [50] proposed Practical Swarm Optimization to migrate virtual machine dynamically for the improvement of resource allocation and acquire more advantages in the data center. This technique guarantees QoS and minimum response time by proposing a new heuristics technique to allocate resource dynamically, with balancing the load of virtual machines. Y. Gao, H. Guan, Zh. Qi, Y. Hou, and L. Liu [51] proposed Ant Colony system-based resource allocation technique to solve the problem of virtual machine placement. The main objective of the technique is to provide accurate solution that simultaneously decreases the resource wastage and power consumption. The comparison of the introduced technique with the existing techniques shows that the proposed technique can compete professionally with other techniques.

There are few techniques that promote green computing while allocating the resources to the tasks in cloud computing. For instance, N. J. Kansal and I. Chana [52] proposed a resource allocation technique to manage the resources of cloud and maximize their utilization efficiently. The main aim of the proposed technique is to reduce energy consumption of cloud without influencing the performance of task. This resource utilization technique is introduced to locate the most suitable work hub node based on Artificial Bee Colony meta-heuristic technique. The energy utilization is decreased with the contention among processor utilization and memory. There are two workload types, 1) CPU and 2) memory intensive, that are considered in this technique. These workloads are carefully associated in order to avoid the conflicts among the resources. Therefore, when minimizing energy consumption and carbon emission, this technique makes contributions to the green computing and is helpful in increasing the satisfaction of cloud user. Furthermore, in [53], the authors proposed a resource allocation technique known as GRMP-Q protocol to decrease the energy utilization of the user servers and computers. The aim of the protocol is to migrate the maximum workload towards the servers and allot the durations of CPU to the consumers.

2. *Thermal Aware*

Usage of the data center has increased substantially, and that leads to the increase in power consumption. Power consumption directly affects the temperature of the physical machines in the data centers due to which the performance and reliability of the system is affected. That is why there is a need of techniques that control the temperature of the systems. Thermal aware resource allocation techniques predict the thermal effect of job employment and the resource allocation [54]. Few thermal aware techniques consider the decreasing of the workload. For instance, A. Beloglazov and R. Buyya [55] proposed the decentralized approach to reallocate the virtual machines to the physical machines keeping essential parameters of QoS in sight between consumers and providers. They proposed an energy efficient technique according to the network bandwidth, CPU, and RAM optimization for the allocation of the virtual machine. They considered the present physical nodes' temperature in resource reallocation process for thermal optimization. The main aim of the technique was to minimize error-proneness and cooling system load, and also decrease the workload of the overheated nodes to prevent hotspots. Similarly, R. Ayoub, K. Indukuri, and T. S. Rosing [56] proposed the dynamic scheduling temperature aware workload technique. They transfer the workload from the cold nodes to the hot nodes. The proposed technique worked on two levels: 1) socket level and 2) core level. They present the temperature predictor which employees the band limited property.

There is a scheduler which deals with the jobs between sockets at the socket level and takes the fan speed, performance, and temperature as an input. The scheduler decides the predicted temperature of every core at the core level and transfers the task from the hot core to cold core. A similar technique was introduced by Y. Kodama *et al.* [57] for power efficiency in data centers. The authors demonstrated that there is a noteworthy irregularity in the temperatures of CPU that is a reason for disparity in the power utilization of fans. According to the authors, that rearrangements and the scheduling of nodes decrease the power utilization of the fans.

The analysis in table VI shows that energy-based resource allocation technique in [53] is relatively efficient as the parameters such as cost, energy, and workload have low value and resource utilization and user satisfaction have high value. On the other hand, thermal aware resource allocation technique in [57] has low cost, less energy consumption, low execution time, and high user satisfaction.

TABLE VI. THE COMPARISON OF POWER-BASED RESOURCE ALLOCATION TECHNIQUES.

Energy Aware							
Reference No.	Cost	Energy	Workload	Resource Utilization	Response time	Execution time	User Satisfaction
[50]	2	2	-	2	4	-	2
[51]	2	1	-	2	-	-	2
[52]	3	1	3	3	-	2	3
[53]	1	1	2	4	-	-	4
Thermal Aware							
Reference No.	Cost	Energy	Workload	Resource Utilization	Response time	Execution time	User Satisfaction
[55]	1	2	4	-	3	-	2
[56]	-	2	-	3	2	-	3
[57]	2	1	-	-	-	1	4

IV. FUTURE DIRECTION

Cloud computing provide services that enables to allocate VM to various tasks according to the demand of the cloud consumers. Based on the existing research, there are still few aspects that should be covered in the domain of resource allocation in cloud computing.

Strategic: The strategic resource allocation techniques are utilized to increase or decrease the allocation of the resources according to the ever-changing demands of the cloud consumers. The strategic-based resource allocation has already covered the aspects of fulfilling the fluctuating demands of the user's by prediction and employment of the artificial intelligence to allocate the resources.

One prominent area of future research is to find the details for detection of resource and workload for improved mappings for the execution and scheduling of jobs. To this end, workloads should be executed efficiently, so as to be flexible, scalable, and optimal, thus avoiding under and over utilization of resources.

Also, the use of artificial intelligence algorithms in resource allocation decreases the error chances and failure rate to nearly zero, better precision and accuracy are accomplished for resource allocation in cloud computing. However, at the same time, artificial-based resource allocation should also focus on cost factor and improve the technique to make it suitable for larger systems as well.

Target Resource: target resource-based resource allocation characterizes the specific resource for which the allocation technique is designed. The aspects of the VM position on a physical machine and the network aware resource allocation have been covered already.

The network aware resource allocation should focus on the minimization of the communication among the VMs which belong to various sub-data centers (or servers). The techniques should focus mainly to satisfy the consumer requirements and reduce the communication cost by finding the shortest path among the VMs.

Optimization: Consumer's demands are increasing almost every day, which requires a proper resource allocation technique to fulfil these demands. Optimization-based resource allocation solves this issue by guaranteeing of QoS to the cloud users. The researchers have already proposed techniques for better resource utilization and the guarantee of the QoS in optimization-based resource allocation.

The optimization-based resource allocation techniques should focus to improve the efficiency of the techniques in terms of total profit and to improve consumer satisfaction levels by considering the SLA negotiation procedure in cloud computing environments. Moreover, there is a need to focus on the penalty limitation by considering system failures.

Scheduling: The scheduling of resources guarantees the effective and efficient resource utilization, and early identification of resource capacity. The work on the allocation of the resources to the task based on their priority, cost, and CPU time (RR) have been already done in scheduling-based resource allocation.

For the implementation of the different resource scheduling algorithms, different scheduling criteria have to be reassessed. Also, based on the existing research, we felt the need to test the resource scheduling algorithms on real environment. We realized that dynamic resource scheduling is an open issue.

Power: Power-based resource allocation techniques aims to reduce the energy consumption, less heat generation, and wastage of resources. The techniques on less energy consumption and heat generation have already been proposed in power-based resource allocation.

There is a need of a comprehensive research on power-based resource allocation technique, most importantly with respect to green optimization of the data center. Low power energy-efficient hardware equipment is used by the low energy aware technologies for the reduction of the energy utilization and peak power consumption.

V. Conclusions

This research presents a structured literature survey based on resource allocation techniques in cloud computing. This study helps understanding different resource allocation techniques on the basis of their schemes, the problems addressed, and the results of their approaches that are used by the different researchers in a contextualized manner. Apart from presenting a summary of the selected articles under proper heads, it also presents promising future directions in the field of resource allocation in cloud computing.

This research paper also concludes that efficient resource allocation technique should meet criteria like cost, energy, response time, execution time, workload, resource utilization, user satisfaction, and SLA. The techniques discussed in this research paper must be beneficial to the cloud users in terms of quality of service and also to the cloud service providers in terms of profit. The study helps us to conclude that recent developments in storage infrastructure, networks and wireless communications, and virtualization have influenced the research being conducted in the field of cloud computing. Moreover, it discusses the advantages and disadvantages of various resource allocation strategies in literature and highlights future directions. The future research directions involve improved usage of artificial intelligence in scheduling and optimization of resource allocation strategies. Additionally, it is also recommended that an extensive research is needed on power-based resource allocation schemes, especially with regard to green optimization of the data center. Similarly, mobility patterns need to be investigated further to improve the task assignment and resource allocation in cloud computing. Lastly, it is envisaged that the services of cloud computing will become an integral part of almost all types and scales of information systems.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] R. Buyya, Ch. Sh. Yeo, S. Venugopal, J. Broberg, and I. Brandic, “Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility”, *Future Generation Computing System*, vol. 26, no. 6, pp. 599–616, Jun. 2009. DOI: 10.1016/j.future.2008.12.001.

[2] S. Gong, B. Yin, Z. Zheng, and K.-Y. Cai, “Adaptive multivariable control for multiple resource allocation of service-based systems in cloud computing”, *IEEE Access*, vol. 7, pp. 13817–13831, 2019. DOI: 10.1109/ACCESS.2019.2894188.

[3] S. H. H. Madni, M. Sh. A. Latiff, Y. Coulibaly, and Sh. M. Abdulhamid, “Recent advancements in resource allocation techniques for cloud computing environment: A systematic review”, *Cluster Computing*, vol. 20, no. 3, pp. 2489–2533, 2017. DOI: 10.1007/s10586-016-0684-4.

[4] Q. Qi and F. Tao, “A smart manufacturing service system based on edge computing, fog computing, and cloud computing”, *IEEE Access*, vol. 7, pp. 86769–86777, 2019. DOI: 10.1109/ACCESS.2019.2923610.

[5] A. Berl, E. Gelenbe, M. Di Girolamo, G. Giuliani, H. De Meer, M. Dang, and K. Pentikousis, “Energy-efficient cloud computing”, *The Computer Journal*, vol. 53, no. 7, pp. 1045–1051, 2010. DOI: 10.1093/comjnl/bxp080.

[6] Ch. Guo, H. Wu, K. Tan, L. Shi, Y. Zhang, and S. Lu, “Dcell: A scalable and fault-tolerant network structure for data centers”, *ACM SIGCOMM Computer Communication Review*, vol. 38, no. 4, pp. 75–86, 2008. DOI: 10.1145/1402958.1402968.

[7] B. Sotomayor, R. S. Montero, I. M. Llorente, and I. Foster, “Resource leasing and the art of suspending virtual machines”, in *Proc. of the 11th IEEE International Conference on High Performance Computing and Communications*, 2009, pp. 59–68. DOI: 10.1109/HPCC.2009.17.

[8] Ch. Li and L. Y. Li, “Optimal resource provisioning for cloud computing environment”, *The Journal of Supercomputing*, vol. 62, no. 2, pp. 989–1022, 2012. DOI: 10.1007/s11227-012-0775-9.

[9] M. A. Salehi, B. Javadi, and R. Buyya, “Resource provisioning based on preempting virtual machines in resource sharing environments”, *The Journal of Concurrency and Computation: Practice and Experience*, pp. 1–21, 2013. DOI: 10.1002/cpe.3004.

[10] A. T. Saraswathi, Y. R. A. Kalaashri, and S. Padmavathi Dr., “Dynamic resource allocation scheme in cloud computing”, *Procedia Computer Science*, vol. 47, pp. 30–36, 2015. DOI: 10.1016/j.procs.2015.03.180.

[11] M. Shojafar, S. Javanmardi, S. Abolfazli, and N. Cordeschi, “FUGE: A joint meta-heuristic approach to cloud job scheduling algorithm using fuzzy theory and a genetic method”, *Cluster Computing*, vol. 18, no. 2, pp. 829–844, 2015. DOI: 10.1007/s10586-014-0420-x.

[12] V. V. Kumar and K. Dinesh, “Job scheduling using fuzzy neural network algorithm in cloud environment”, *Bonfring International Journal of Man Machine Interface*, vol. 2, no. 1, pp. 1, 2012. DOI: 10.9756/BIJMMI.1064.

[13] L. Ying, Q. P. Rui, and X. Jie, “Computing resource allocation for enterprise information management based on cloud platform ant colony optimization algorithm”, *Advanced Materials Research*, vols. 791–793, pp. 1232–1237, 2013. DOI: 10.4028/www.scientific.net/AMR.791-793.1232.

[14] Ch. Li and L. Li, “Efficient resource allocation for optimizing objectives of cloud users, IaaS provider and SaaS provider in cloud environment”, *The Journal of Supercomputing*, vol. 65, no. 2, pp. 866–885, 2013. DOI: 10.1007/s11227-013-0869-z.

[15] T. S. Somasundaram, B. R. Amarnath, R. Kumar, P. Balakrishnan, K. Rajendar, R. Rajiv, G. Kannan, G. R. Britto, E. Mahendran, and B. Madusudhanan, “CARE Resource Broker: A framework for scheduling and supporting virtual resource management”, *Future Generation Computer Systems*, vol. 26, no. 3, pp. 337–347, 2010. DOI: 10.1016/j.future.2009.10.005.

[16] J. Machina and A. Sodan, “Predicting cache needs and cache sensitivity for applications in cloud computing on cmp servers with configurable caches”, in *Proc. of the IEEE International Symposium on Parallel & Distributed Processing*, 2009, pp. 1–8. DOI: 10.1109/IPDPS.2009.5161233.

[17] S. Kundu, R. Rangaswami, A. Gulati, M. Zhao, and K. Dutta, “Modeling virtualized applications using machine learning techniques”, in *Proc. of 8th ACM SIGPLAN/SIGOPS Conference on Virtual Execution Environments*, 2012, vol. 47, pp. 3–14. DOI: 10.1145/2151024.2151028.

[18] J. Wildstrom, P. Stone, E. Witchel, and M. Dahlin, “Machine Learning for on-line hardware reconfiguration”, in *Proc. of the 20th International Joint Conference on Artificial Intelligence*, 2007, vol. 7, pp. 1113–1118.

[19] Ch.-H. Lin, Ch.-T. Lu, Y.-H. Chen, and J.-Sh. Li, “Resource allocation in cloud virtual machines based on empirical service traces”, *International Journal of Communication Systems*, vol. 27, no. 12, pp. 4210–4225, 2014. DOI: 10.1002/dac.2607.

[20] Sh. Zaman and D. Grosu, “A combinatorial auction based mechanism for dynamic VM provisioning and allocation in clouds”, *IEEE Transactions on Cloud Computing*, vol. 1, no. 2, pp. 129–141, 2013. DOI: 10.1109/TCC.2013.9.

[21] J. Frey, “Network management and the responsible, virtualized cloud”, research rep., 2011.

[22] G. Sun, V. Anand, H.-F. Yu, D. Liao, and L. Li, “Optimal provisioning for elastic service oriented virtual network request in cloud computing”, in *Proc. of 2012 IEEE Global Communications Conference (GLOBECOM)*, 2012, pp. 2517–2522. DOI: 10.1109/GLOCOM.2012.6503495.

[23] T. D. Wallace, A. Shami and C. Assi, “Scheduling advance reservation requests for wavelength division multiplexed networks with static traffic demands”, *IET communications*, vol. 2, no. 8, pp. 1023–1033, 2008. DOI: 10.1049/iet-com:20070500.

[24] X. Meng, V. Pappas, and L. Zhang, “Improving the scalability of data center networks with traffic-aware virtual machine placement”, in *Proc. of 2010 IEEE INFOCOM*, 2010, pp. 1–9, 2010. DOI: 10.1109/INFCOM.2010.5461930.

[25] J. Dong, X. Jin, H. Wang, Y. Li, P. Zhang, and Sh. Cheng, “Energy-saving virtual machine placement in cloud data centers”, in *Proc. of the 13th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing*, 2013, pp. 618–624. DOI: 10.1109/CCGrid.2013.107.

[26] X. Lu, J. Zhou, and D. Liu, “A method of cloud resource load balancing scheduling based on improved adaptive genetic algorithm”, *Journal of Information & Computational Science*, vol. 9, no. 16, pp. 4801–4809, 2012.

[27] S. Ravichandran and E. R. Naganathan, “Dynamic scheduling of data using genetic algorithm in cloud computing”, *International Journal of Computing Algorithm*, vol. 2, no. 1, pp. 11–15, 2013. DOI: 10.20894/IJCOA.101.002.001.003.

[28] R. Lee and B. Jeng, “Load-balancing tactics in cloud”, in *Proc. of the IEEE International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, 2011, pp. 447–454. DOI: 10.1109/CyberC.2011.79.

[29] Z. Abbasi, G. Varsamopoulos, and S. K. S. Gupta, “Thermal aware server provisioning and workload distribution for internet data centers”, in *Proc. of the 19th ACM International Symposium on High Performance Distributed Computing*, 2010, pp. 130–141. DOI: 10.1145/1851476.1851493.

[30] P. B. Galvin, “VMware vSphere vs. Microsoft Hyper-V: A technical analysis”, Corporate Technologies, CTI Strategy White Paper, 2009.

[31] Q. Zhang, M. F. Zhani, Sh. Zhang, Q. Zhu, R. Boutaba, and J. L. Hellerstein, “Dynamic energy-aware capacity provisioning for cloud computing environments”, in *Proc. of the 9th ACM international conference on Autonomic computing*, 2012, pp. 145–154. DOI: 10.1145/2371536.2371562.

[32] Y. Ding, X. Qin, L. Liu, and T. Wang, “Energy efficient scheduling of virtual machines in cloud with deadline constraint”, *Future Generation Computer Systems*, vol. 50, pp. 62–74, 2015. DOI: 10.1016/j.future.2015.02.001.

[33] A. Wolke, B. Tsend-Ayush, C. Pfeiffer, and M. Bichler, “More than bin packing: Dynamic resource allocation strategies in cloud data centers”, *Information Systems*, vol. 52, pp. 83–95. 2015. DOI: 10.1016/j.is.2015.03.003.

[34] S. Son, G. Jung, and S. Ch. Jun, “An SLA-based cloud computing that facilitates resource allocation in the distributed data centers of a cloud provider”, *The Journal of Supercomputing*, vol. 64, no. 2, pp. 606–637, 2013. DOI: 10.1007/s11227-012-0861-z.

[35] S. Singh and I. Chana, “QoS-aware autonomic resource management in cloud computing: A systematic review”, *ACM Computing Surveys (CSUR)*, vol. 48, no. 3, article no. 42, 2016. DOI: 10.1145/2843889.

[36] S. Iqbal, M. L. M. Kiah, B. Dhaghighi, M. Hussain, S. Khan, M. K. Khan, and K.-K. R. Choo, “On cloud security attacks: A taxonomy and intrusion detection and prevention as a service”, *Journal of Network and Computer Applications*, vol. 74, pp. 98–120, 2016. DOI: 10.1016/j.jnca.2016.08.016.

[37] F. I. Popovici and J. Wilkes, “Profitable services in an uncertain world”, in *Proc. of the 18th IEEE/ACM Conference on Supercomputing*, 2005, p. 36. DOI: 10.1109/SC.2005.58.

[38] L. Wu, S. K. Garg, and R. Buyya, “SLA-based resource allocation for software as a service provider (SaaS) in cloud computing environments”, in *Proc. of the 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*, 2011, pp. 195–204. DOI: 10.1109/CCGrid.2011.51.

[39] V. C. Emeakaroha, I. Brandic, M. Maurer, and I. Breskovic, “LA-aware application deployment and resource allocation in clouds”, in *Proc. of the 35th annual IEEE computer software and applications conference workshops*, 2011, pp. 298–303. DOI: 10.1109/COMPSACW.2011.97.

[40] A. Kumar, E. S. Pilli, and R. C. Joshi, “An efficient framework for resource allocation in cloud computing”, in *Proc. of 4th IEEE International Conference on Computing, Communications and Networking Technologies*, 2013, pp. 1–6. DOI: 10.1109/ICCCNT.2013.6726596.

[41] B. Sotomayor, R. S. Montero, I. M. Llorente, and I. Foster, “Virtual infrastructure management in private and hybrid clouds”, *IEEE Internet Computing*, vol. 13, no. 5, 2009. DOI: 10.1109/MIC.2009.119.

[42] Ch. S. Pawar and R. B. Wagh, “Priority based dynamic resource allocation in Cloud computing”, in *Proc. of the IEEE International Symposium on Cloud and Services Computing*, 2013, pp. 311–316. DOI: 10.1109/ISCOS.2012.14.

[43] Zh. Lee, Y. Wang, and W. Zhou, “A dynamic priority scheduling algorithm on service request scheduling in cloud computing”, in *Proc. of the IEEE International Conference on Electronic and Mechanical Engineering and Information Technology*, 2011, vol. 9, pp. 4665–4669. DOI: 10.1109/EMEIT.2011.6024076.

[44] X. Wu, M. Deng, R. Zhang, B. Zeng, and Sh. Zhou, “A task scheduling algorithm based on QoS-driven in Cloud Computing”, *Procedia Computer Science*, vol. 17, pp. 1162–1169, 2013. DOI: 10.1016/j.procs.2013.05.148.

[45] A. Abdulrahim, S. Abdullahi, and J. B. Sahalu, “A new improved round robin (NIRR) CPU scheduling algorithm”, *International Journal of Computer Applications*, vol. 90, no. 4, pp. 27–33, 2014. DOI: 10.5120/15563-4277.

[46] M. Mishra and A. Khan, “An improved round robin CPU scheduling algorithm”, *Journal of Global Research in Computer Sciences*, vol. 3, no. 6, pp. 64–69, 2012.

[47] G. S. N. Rao, N. Srinivasu, S. V. N. Srinivasu, and G. R. K. Rao, “Dynamic time slice calculation for round robin process scheduling using NOC”, *International Journal of Electrical and Computer Engineering*, vol. 5, no. 6, 2015. DOI: 10.11591/ijece.v5i6.pp1480-1485.

[48] A. Noon, A. Kalakech, and S. Kadry, “A new round robin based scheduling algorithm for operating systems: Dynamic quantum using the mean average”, *International Journal of Computer Science Issues*, vol. 8, no. 1, 2011. arXiv: 1111.5348.

[49] R. Mohanty, H. S. Behera, K. Patwari, M. Dash, and M. L. Prasanna, “Priority based dynamic round robin (PBDRR) algorithm with intelligent time slice for soft real time systems”, *International Journal of Advanced Computer Science and Applications*, vol. 2, no. 2, 2011. DOI: 10.14569/IJACSA.2011.020209.

[50] S. E. Dashti and A. M. Rahmani, “Dynamic VMs placement for energy efficiency by PSO in cloud computing”, *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 28, no. 1–2, pp. 97–112, 2016. DOI: 10.1080/0952813X.2015.1020519.

[51] Y. Gao, H. Guan, Zh. Qi, Y. Hou, and L. Liu, “A multi-objective ant colony system algorithm for virtual machine placement in cloud computing”, *Journal of Computer and System Sciences*, vol. 79, no. 8, pp. 1230–1242, 2013. DOI: 10.1016/j.jcss.2013.02.004.

[52] N. J. Kansal and I. Chana, “Artificial bee colony based energy-aware resource utilization technique for cloud computing”, *Concurrency and Computation: Practice and Experience*, vol. 27, no. 8, 2015. DOI: 10.1002/cpe.3295.

[53] R. Yanggratoke, F. Wuhib, and R. Stadler, “Gossip-based resource allocation for green computing in large clouds”, in *Proc. of the 7th IEEE International Conference on Network and Service Management*, 2011, pp. 1–9.

[54] A. Beloglazov, R. Buyya, Y. Ch. Lee, and A. Zomaya, “A taxonomy and survey of energy-efficient data centers and cloud computing systems”, *Advances in computers*, vol. 82, pp. 47–111, 2011. DOI:

10.1016/B978-0-12-385512-1.00003-7.

[55] A. Beloglazov and R. Buyya, “Energy efficient resource management in virtualized cloud data centers”, in *Proc. of the 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing*, 2010, pp. 826–831. DOI: 10.1109/CCGRID.2010.46.

[56] R. Ayoub, K. Indukuri, and T. S. Rosing, “Temperature aware dynamic workload scheduling in multisocket CPU servers”, *IEEE transactions on Computer-aided design of integrated circuits and systems*, vol. 30, no. 9, pp. 1359–1379, 2011. DOI: 10.1109/TCAD.2011.2153852.

[57] Y. Kodama, S. Itoh, T. Shimizu, S. Sekiguchi, H. Nakamura, and N. Mori, “Imbalance of CPU temperatures in a blade system and its impact for power consumption of fans”, *Cluster computing*, vol. 16, no. 1, pp. 27–37, 2011. DOI: 10.1007/s10586-011-0174-7.