Challenges and Issues of Resource Allocation Techniques in Cloud Computing

Adnan Abid[1], Muhammad Faraz Manzoor[1*], Muhammad Shoaib Farooq[1], Uzma Farooq[1], and Muzammil Hussain[1]

[1] Department of Computer Science, University of Management and Technology, Lahore, Pakistan
[e-mail: f2018288004@umt.edu.pk]
*Corresponding author: Muhammad Faraz Manzoor

Received February 25, 2020; revised April 15, 2020; accepted May 30, 2020; published July 31, 2020

Abstract

In a cloud computing paradigm, allocation of various virtualized ICT resources is a complex problem due to the presence of heterogeneous application (MapReduce, content delivery and networks web applications) workloads having contentious allocation requirements in terms of ICT resource capacities (resource utilization, execution time, response time, etc.). 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. However, there is no published research available in this domain that clearly address this research problem and provides research taxonomy for classification of resource allocation techniques including strategic, target resources, optimization, scheduling and power. Hence, the main aim of this paper is to identify open challenges faced by the cloud service provider related to allocation of resource such as servers, storage and networks in cloud computing. More than 70 articles, between year 2007 and 2020, related to resource allocation in cloud computing have been shortlisted through a structured mechanism and are reviewed under clearly defined objectives. Lastly, the evolution of research in resource allocation techniques has also been discussed along with salient future directions in this area.

Keywords: Cloud computing, Resource Allocation, Resource scheduling, Resource Utilization

1. Introduction

Cloud computing has been around for about two decades and despite the data pointing to the business efficiencies, cost-benefits, and competitive advantages it holds, a large portion of the business community continues to operate without it. Big organizations are using cloud applications to provide the best services to the service consumers and handle the data in excessive amount [79].The main applications of cloud computing systems involve civil infrastructure, healthcare, industrial manufacturing, enery saving systems and transportation [78]. The objectives of this contemporary technology includes on demand service, multi-tenacy, fast and effective virtualization, web based control and interfaces and pay per use payment method. From techniques perspective, cloud computing has evolved from various computing models like, parallel computing, distributed computing and grid computing, and has a further focus on applicability. 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 is the process of allocating available resources to the cloud applications over the internet 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 QoS and user satisfaction[2].

Table 1. Research questions

Questions	Motivations
What is the significance of the allocation of resources in cloud computing?	- Allows the cloud service providers to manage the resources for each individual module in cost efficient manner.
What are the existing techniques to allocate resources in cloud environment?	- Figure out various techniques of resource allocation in cloud computing - Discuss each resource allocation technique
What are the parameters and resources have considered more during resource allocation?	- important parameters(e.g., cost, energy, response time etc.) and resources(servers, storage and networks) for the service consumer and service providers during allocation of resources.
What are the research gaps in resource allocation in cloud computing domain?	- To highlight the research opportunities in different areas of resource allocation in cloud computing

Every cloud service consumer wants a maximum resources for a specific task that can increase the performance and have to be finished on time. In the same way, it is the responsibility of resource allocator 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 at low cost[3][4]. From storage perspective, datacenters provide the resources and distributed computing models facilitate on request resources allocation, which prompts the non-ideal resource assignment. On the other hand energy utilization is another main issue that datacenters face. It has been seen that energy 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 main objective of this study is to provide the challenges faced by cloud service provider while allocating the resources and taxonomy of current advances in resource allocation techniques, while emphasizing on their strengths and weaknesses by employing performance metrics for evaluation. So, the four research questions have been developed which are defined in **Table 1**.

Following is the organization of rest of the article: Section 2 presents related work and provides a comparison of the related surveys with this study. Section 3 presents the selection process for this study section 4 describes the mechanism of resource allocation in cloud computing. The taxonomy and subdivisions of the presented techniques in detail are discussed in section 5. Whereas, Section 6 discusses the evolution and future directions in this area. Lastly, section 7 presents the conclusion of the article.

2. Related Work

There is no study has been conducted on resource allocation techniques in cloud computing that discusses the strategic, target resources, optimization, scheduling and power in terms of cost, resource utilization, energy, workload, execution time, response time, user satisfaction, and QoS. One of the highlights of this study is that it also discusses how these resources have evolved over the past decade. Though there have been studies which focus on one or more aspects, yet this survey is considerably different from other studies. **Table 2** highlights the differences between this study and other relevant studies in the field of resource allocation in cloud computing.

Table 2. Comparison with other related works

Authors	**Sukhpal and Chana[6]**	**Lavanya and Shoba [7]**	**Hameed et.al [8]**	**Mohamaddiah et.al. [9]**	**Bhavani and Guruprasad[10]**	**This Research**
Focus of Study	Resource Scheduling Technqiues	Resource Scheduling Techniques	Energy Efficient Techniques	Resource Allocation and Monitoring	Resource Allocation Techniques	Resource Allocation Techniques
Evaluation of techniques	✓	-	-	-	-	✓
Taxonomy presented	✓	-	✓	✓	✓	✓
Evolution	-	✓	-	-	-	✓
Year	2016	2016	2016	2014	2014	2020

Sukhpal and Chana [6] have discussed resource scheduling techniques using systematic literature review method. The authors have presented issues faced by cloud service provide and cloud service consumer in comprehensive manner, but the evolution and trends in scheduling techniques remain unaddressed. Also, Lavanya and Shoba [7] presented a review on resource allocation and resource scheduling in cloud computing in systematic literature manner. Authors have discussed the evolution of resource scheduling techniques but evaluation of presented techniques was not discussed. On the other hand, Hameed et.al [8] summarized the energy efficient available techniques presented in the existing literature. The highlight of their study was that they have analyzed the advantages and disadvantages of techniques. However, the authors did not evaluate the techniques based on the resource allocation technique parameters. Subsequently, Mohamaddiah et.al. [9] conducts a study in resource allocation and monitoring in the cloud computing environment. They describe resource allocation and monitoring issues in cloud computing and finally solution approach for resource allocation and monitoring but the evaluation and evolution of resource allocation techniques was not discussed. Similarly, Bhavani Guruprasad [10] provides a survey of some of the models and solutions for the resource allocation problem in the cloud computing environment. While the authors have discussed research challenges in this domain but they did discuss evaluation and evolution of resource allocation techniques in cloud computing.

3. Selection Process

3677 initial studies have been produced using the search process, from which 290 are relevant and 77 are selected for primary study. The selection process consists of three steps to acquire independent assessments as shown in **Fig. 1**.

Pre stage(Search string) *:* it is an initial step for the formulation of the search string, we define the searching keywords, depending on the topic and the formulated research questions (shown in Table 1). For the connection of the keyword the logical operators AND and OR have been used. After few tests, the following search string has been developed which gives us the adequate amount of relevant research studies: (Cloud) AND (Computing) AND (Resource)) OR (Strategic) OR (Target Resource) OR (Optimization) OR (Scheduling) OR ((Energy) AND (Power))

Title based search: Papers that are irrelevant are manually excluded, based on the title in the first stage. Only 290 papers remained after this stage.

Abstract based search: At this stage, abstracts of the selected papers in the previous stage are studied and the papers are categorized for the analysis along with research approach. 181 papers remained after this stage.

Full text based analysis: At this stage the empirical quality of the studies has been evaluated. The total of 77 papers were extracted from 181 papers for primary study. Following questions are defined for the conduction of final data extraction.

- Is the literature search expected to have covered all studies related to the research domain?
- Did the reviewers evaluate the quality of the involved research?
- Were the fundamental studies sufficiently described?

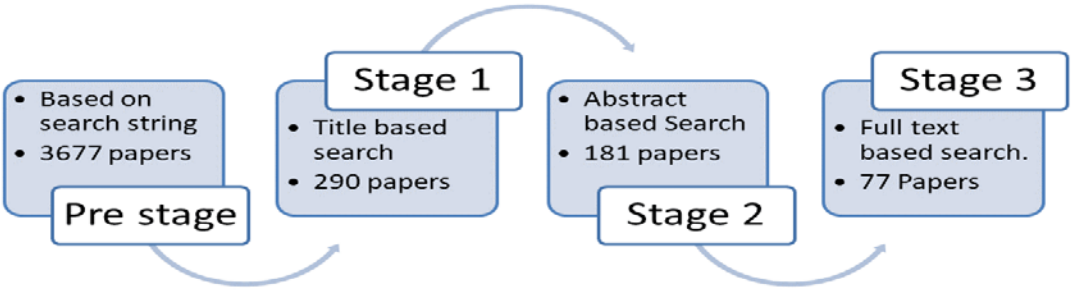

Fig. 1. Selection process

4. Resource Allocation in Cloud Computing

In cloud computing, resource allocation is the process in which virtual machine is designated to fulfill the properties define by the consumers. 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 [11]. Simply, it is all about defining when a computational action should begin or finish dependent upon: 1) resource allocation 2) time taken 3) action of the predecessor 4) relationships of the predecessor. The general process of allocation of resources is shown in **Fig. 2**.

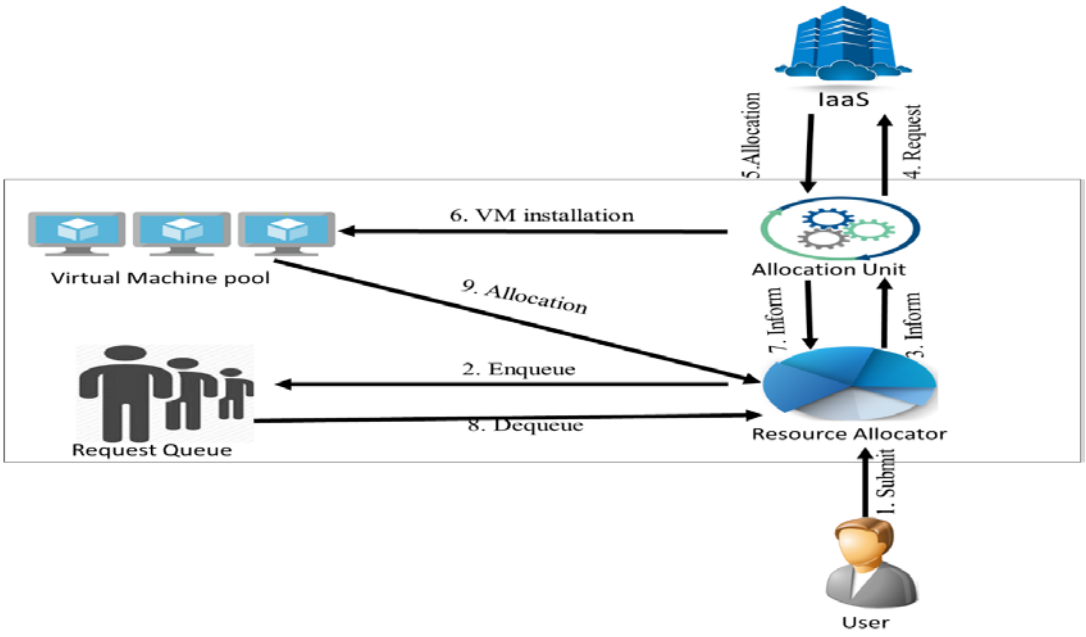

Fig. 2. General flows of resource allocation in cloud computing

Moreover, the resources in data center of cloud service provider are progressively heterogeneous and comprise of changing generations of infrastructural hardware because of innovative headway. Like, nowadays in industries multicore CPU's with large cache memory are used. Furthermore, in data centers more energy efficient CPUs such as the

ARM [12] or Atom[13] CPUs are ready to be used. Moreover, other subsystems are advancing, also; such advancements incorporate Memristor [14], Phase Change Random Access Memory[15], Solid State Drive[16] storage and memory subsystems, also change in network architecture like B-Cube[17] and D-Cell[18]. To keep up their services up to date, Cloud service provider needs to embrace these changes in technology, yet this promotes the problem of heterogeneity. So, it is important to make use of the regularly expanding heterogeneity for cloud service provider and cloud service consumer to accomplish their objectives: most extreme use of resources and cost execution.

On the other hand cloud service provider's face few challenges during allocation of resources among the user's tasks based on their application usage patterns. Few of them are as follows:

i. The prediction of consumer's application requirement is very difficult for cloud service provider and at the same time consumer wants to complete the task on time. So, efficient resource allocation techniques are required to overcome this problem.

ii. The physical machines should be capable enough to fulfill the resource needs of every virtual machine running on it and at the same time the consumers need the networking services with efficient QoS to guarantee the effective delivery of their application data.

iii. The service provider has to schedule the availability of the resources in case of a job that can take time more than usual. Thus, there is need of a technique that can handle the interruption and also switch the job to the available resource

iv. Energy efficient resource allocation is one of the open challenge in cloud computing because of the increasing energy costs and the need to minimize greenhouse gas emissions and also to reduce the total energy consumption, communications and storage.

It is clear from above discussion that qualities and properties of both cloud service provider and cloud service consumer should be taken into consideration to provide efficient services so that adequate resources are allocated to a suitable task with an objective that task should be completed on time and cloud service provider gain maximum profit.

5. Resource Allocation Techniques

Resource allocation in cloud computing involves decision making by cloud service provider with respect to what, when, how much and where to allocate the available resources to the task. Normally, users determine the amount and type of the resources for the request, and in response, the service providers allocate the requested resources in their data centers. For the efficient execution of applications, the type and the numbers of resource containers should be sufficient to meet the constraints defined by the user (i.e., job completion time deadline) and also should match the workload characteristics.

The analysis in this study reveals that 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) optimization: optimizing the resources, 4) Scheduling: prioritizing the task for better performance and 5) power: better resource allocation with less power consumption. The brief taxonomy of resource allocation techniques is shown in **Fig. 3**.

Following are the different parameters to evaluate the various resource allocation techniques from both cloud service provider (resource utilization, workload, cost, energy, SLA, QoS) and cloud service consumer(response time, user satisfaction, execution time,

SLA, QoS) perspective:

Resource Utilization: One of the objective of the cloud service provider is to utilize all the resources efficiently in order to prevent them to remain idle. Also, maximum resource utilization is important for environmental safety and it helps to maximize the profit.

Workload: it is the capability of the system to process a task. It is important for the cloud service provider to make sure that the workload should be enough on a system to complete the tasks on time. This parameter will determine the amount of workload on empirical setup of resource allocation techniques.

Cost: This parameter will determine the profit or loss of cloud service provide for providing different services. It is important to mention that in this study this parameter will only addresses the profit and loss of cloud service provider not the cloud service consumer.

Energy: it is important for cloud service provider to minimize the utilization of energy to make the cloud services environmentally supportable keeping ever increasing energy crisis in view.

Response Time: system performance can be determined by the response time of system to a task. While cloud service provider wants it to be as low as possible, it will help also important for successful computing.

User Satisfaction: Every cloud service provider wants to satisfy his consumers effective allocation of resources so that the revenue and user satisfaction can be maximized

Execution Time: minimum execution time is ideal for both cloud service provider and cloud service consumer. But it is important to mention that the multiple workloads on single system can cause interference among workload which will eventually lead to poor performance.

Some of the other parameters from cloud service provider and cloud service consumer are QOS and SLA to evaluate various resource allocation techniques.

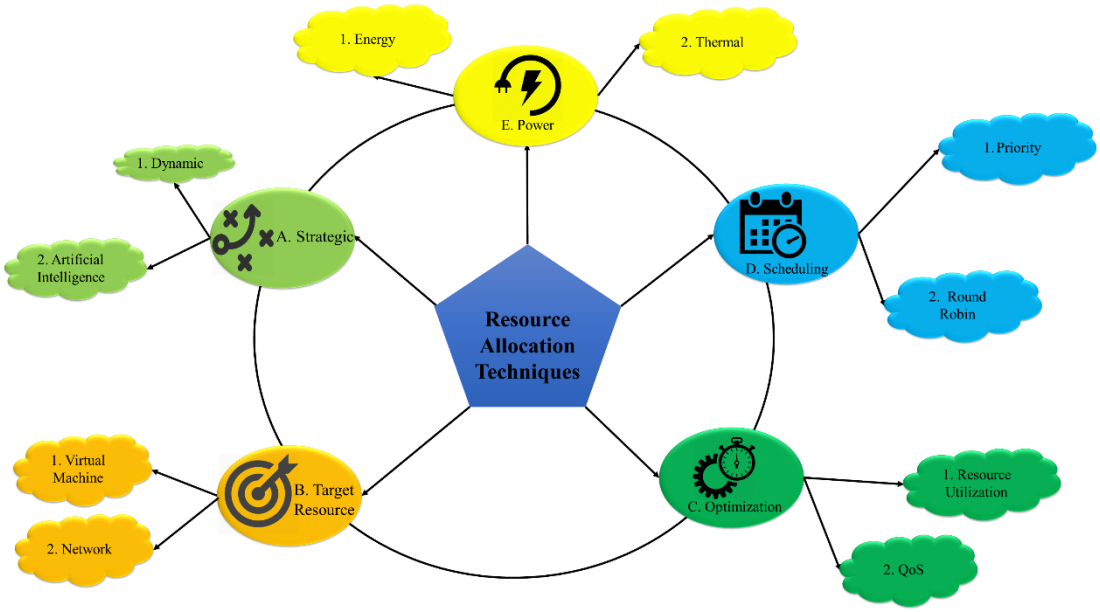

Fig. 3. Taxonomy of resource allocation techniques

Different types of notations are used to evaluate the resource allocation techniques in previously published literature in this domain. It was not possible to determine the

dependency of one parameter to other by those notations . There have been few reviews published related to resource allocation techniques in cloud computing. They used different type of notation in order to evaluate the proposed techniques. The drawback from that evaluation is that it does not show the dependency of one parameter to other. In this interrelated notations are used to show the dependencies of parameters.
Notation ↑ represents the high and ↓ represents the low value of parameter. It is noticeable that high value of parameter is not ideal in every case for instance value should be low for the parameters like response time, cost , energy, workload and execution time.

5.1 Strategic

Promising features for both cloud service provider and consumer have led the great development in the adoption of cloud computing. Cloud service provider utilizes strategic based resource allocation techniques for the fair comparison among the available resources, prediction of consumer's demand and selecting the most suitable service for the task. It can be further categorized as: 1) Dynamic Resource allocation: to predict the nature of consumer's task, 2) Artificial intelligence: imitate nature to schedule tasks among resources.

5.1.1 Dynamic

Dynamic resource allocation techniques allow the service provider to address the each task requirements as the requirements of each task submitted by the consumer is different. There are few techniques which use leasing as a fundamental resource provisioning for advance reservation. For instance, C.Guo [19] proposed a technique to support the advance reservation of resources for consumer's task by predicting varying run-time overheads associated with utilizing virtual machine. While the proposed technique processes the task in a small time but cost become high for the cloud service provider. Also, Sotomayor [20] proposed a technique in which SaaS provider leases the resources and then further lease them to the consumer to increase the profit. The proposed technique was unable to minimize the execution time of task. Similarly, C.Li[21] proposed a technique that determine the proper set of lease(s) for preemption to minimize the effect of pre-empting virtual machine. The highlight of this technique was low execution time and maximize resource utilization but the cost of resources allocation increases.

Table 3. Dynamic resource allocation techniques parameters

Author(s) name	Cost	Energy	Resource Utilization	Workload	Execution time
C.Guo[19]	↑	–	↑	–	↓
B.Sotomayor[20]	↓	–	–	–	↑
C.LI[21]	↑	–	↑	–	↓
M.Salehi[22]	↓	–	↑	↓	↓

On the other hand a technique for the high priority task was proposed by M.Salehi et.al. [22]. The proposed technique implements the latest jobs first without considering the formation of latest virtual machines and high priority jobs are completed first by suspending the low priority jobs. It resumes the processing of low priority jobs after the completion of high priority jobs. This technique enabled the service provide to lower the overheads to execute all the tasks. The evaluation of dynamic resource allocation techniques is shown in **Table 3**.

5.1.2 Artificial Intelligence

In cloud computing the artificial intelligence based resource allocation techniques act and work like humans for resource allocation. Keeping the impact of artificial intelligence on cloud computing, authors have started to develop resource allocation techniques based on artificial intelligence such as, Shojafar et.al [23] introduced a new technique name FUGE by merging genetic algorithm and fuzzy model. This technique helps the cloud service provider to find the suitable tasks by representing them as chromosomes. Resource utilization of resources were high but technique did not as much efficient in terms of energy. Also, Kumar and Dinesh[24] proposed fuzzy logic based resource allocation technique. Various parameters such as memory, expected execution time and bandwidth are used to categorize the consumer's task. Fuzzification technique then converts the values from 0 to 1 and then submitted to neural network. The neural network consists of three layers input layer, hidden layer, 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 de-fuzzification phase. This proposed technique has improved the overall performance of the system.

On the other hand, Wei et al. employed the application of game theory across a range of inter-dependent tasks with associated time and cost constraints to optimize resource allocation [25]. The technique was implemented in two phases: in the first phase each agent solves the optimization problem locally and in the second phase an evolutionary global optimization algorithm is used to estimate a final optimal solution while following to QoS constraints resulting in overall improved efficiency.

Table 4. Artficial intelligence resource allocation techniques parameters

Author(s) name	Cost	Energy	Resource Utilization	Workload	Execution time
M. Shojafar[23]	–	↑	↑	–	–
V.V. Kumar[24]	↓	–	↑	–	↓
Wei[25]	↓	↓	↑	–	–
Y. Liang[26]	–	↓	–	↓	↓

Moreover, Y. Liang [26], proposed a technique to allocate the resource efficiently by estimating the bandwidth and predicting the available resources in advance. This technique reduces the energy of resources and execution time of task but the cost factor remain unaddressed. The evaluation of Artificial intelligence based resource allocation techniques is shown in **Table 4**.

5.2 Target Resource

Target resource based resource allocation techniques determine the type of resources that are required for task. In this study two type of resources are discussed and analyzed based on the previously published literature. 1) Virtual Machine: Virtual machine allocation shows the virtual machine's position on physical machine, (2) Network: Network failure in cloud datacenter could happen due to the inefficient resource allocation, logical segmentation of physical machines and scheduling

5.2.1 Virtual Machine

Cloud computing is efficient, affordable and reliable only because of the virtualization technology. Resource services to make the cloud computing functional are costly so the cloud service providers are always looking for the techniques to allocate the resources to consumer's task efficiently. There are few techniques developed by the various authors to meet the job requirements and avoid the delay, for instance, Thamarai et.al [29] proposed a techniques named Care Resource Broker (CRB) to improve the throughput, reaction time of the resources. This technique completes the task efficiently by allocating the exact number of resources. According the results presented by the authors, CRB is cost, resource efficient but authors did not consider energy factor for technique development. Likewise, a technique to execute the consumer's task as a component of assigned cache size, irrespective of whether the cache is dynamically partitioned or not was proposed by Machina et.al [30]. The proposed technique has successfully reduced the execution time of task but cost was high.
On the other hand, few authors have used algorithms of artificial intelligence to allocate the resources to the consumer's task. Like, Kundu et.al [31] show the execution of tasks by utilizing in virtualized framework by using neural networks. They proposed technique has a great effect on response time and execution time but the cost remain unaddressed.
Similarly, Wildstrom et.al [32] employed machine learning algorithms to increase the throughput. They discover the appropriate configurations by showing the execution of application on low level matrices.

Table 5. Virtual machine resource allocation techniques parameteres

Author(s) name	Cost	Energy	Workload	Resource Utilization	Response Time	Execution time	User Satisfaction
Thamarai[29]	↓	–	–	↑	↓	↓	↑
J. Machina [30]	↑	↑	–	↓	–	↓	↑
S. Kundu [31]	↑	–	–	–	↓	↓	↑
J. Wildstrom [32]	↑	–	↑	–	↓	–	–

The proposed technique has minimized the response time but cost of resources was increased. The evaluation of virtual machine resource allocation techniques is shown in **Table 5**.

5.2.2. Network

About 54% IT experts discussion over the cloud services utilization reveals that they include network operation staff, which effects the utilization of systems best practices and consideration concerning the health of complete traffic delivery**[34]**. While, Almost 28% thought that there is a need of monitoring and troubleshooting packets among the virtual machines, 32% thought that monitoring and troubleshooting traffic data from virtual switches is needed. With the technology evolving authors realized that network based resource allocation techniques are also important to optimize the processing of consumer's task. Such as, G.Sun et.al. [35] proposed a technique to conumser is able to process various tasks on multiple servers at the same time. Connection requests are represented as a virtual network where nodes are virtual machines and edges are physical network paths. The aim of the proposed technique was to increase the profit of cloud service provider but important

factor of execution time of task was not considered while developing the technique.

Table 6. Network based resource allocation techniques

Author(s) name	Cost	Energy	Workload	Resource Utilization	Execution time
G. Sun [35]	↓	–	↓	↑	–
X.Meng[36]	↓	↑	↓	–	–
M. Alicherry[37]	↑	–	–	↑	↓
A. Aldhalaan[38]	↓	↓	–	–	–

On the other hand, X.Meng et.al. [36] proposed a technique to deal with the issue of offering the best virtual network with an IP over a wavelength-division multiplexing (WDM**).** The proposed technique was effective against the flow conversation constant and propagation delay but important factors such as execution time and resource utilization ignored**.** Furthermore few techniques have been proposed that uses the distance of datacenters as their main parameter. For instance, Alicherry et.al [37] introduced a network aware resource allocation of virtual machine and selection of data center. They concentrated on the improvement of the completion time of a job by decreasing the distance between the data centers. The proposed technique has reduced the execution time but cost become higher. Aldhalaan and Menasc´ [38] proposed technique in which their aim is not to reduce the cost but maximizing the revenue. The main highlight of this technique is that the user will have to pay more if the virtual machine of a request are allocated near to each other. The evaluation of network resource allocation technique is shown in **Table 6**.

5.3 Optimization

Optimization of resource allocation aims for various objectives such as 1) efficient resource pricing to produce fiscal benefits for service providers and service consumers, 2) efficient resource utilization for environmental safety and reduction in data center operational expenditures, and 3) QoS-based objectives such as makespan/response time minimization. Optimization based resource allocation techniques also improved the throughput by increasing the use of virtual and physical resources. This study categorizes the optimization based resource allocation techniques into 1) Resource Utilization: proper allocation of resources among the tasks for efficient resource utilization and at the same time minimizes the operational consumption of data centers. 2) The Quality of Service(QoS): the main objective of these techniques to accomplish various requirements of customers such as latency, stability, CPU speed. Execution measurements can increase violation of service performance levels without any agreements between service provider and service consumer.

5.3.1 Resource Utilization

The main aim of the resource utilization based resource allocation techniques is to utilize maximum resources and keeping the power consumption as low as possible. So that the resources do not remain idle. Few authors have employed generic algorithm to maximize the maximum resources such as, Xin Lu [39] proposed a technique to find the appropriate resources for every task taking place in real time by enhancing the versatile generic algorithm. The authors successfully maximize the resource utilization but did not address the energy factor. Similarly, Ravichandran and Naganathan [40] proposed a generic algorithm based technique to upgrade the resources of virtual machines. The highlights of the proposed

technique were the low execution time and energy consumption but the cost remains high.
On the hand. Authors have also proposed power efficient resource allocation techniques. Such as, a technique for dynamic volume provisioning to decrease the cost of energy consumption was proposed byLee and Jeng[41]. Real time indications and recommendation acquired from various sites such as Google determine the simulation type. While the authors have reduced the cost and energy consumption but could not handle the execution time. Also, Abbasi [42] proposed a technique to optimize the resource allocation by distributing equal workload among the resources. User chooses one of their active server and a specific threshold has been set so that it does not arise past a specific incentive in server, to ensure the smooth workflow. While the authors have reduced the energy consumption of resources but cost remains high. The evaluation of resource utilization based resource allocation techniques is shown in **Table 7**.

Table 7. Resource utilization based resource allocation technique parameters

Author(s) name	Cost	Energy	Resource Utilization	Workload	Execution Time
X. Lu [39]	↓	–	↑	↓	–
S. Ravichandran [40]	↑	↓	↑	–	↓
R.Lee [41]	↓	↓	↑	–	↑
Z.Abbasi[42]	↑	↓	↑	–	↑

5.3.2 Quality of Service

QoS-based resource allocation strategies are intended to maintain the service-level objective and parameters desired by the consumer in the SLA document. Service Level Agreement(SLA) has a great impact on satisfaction level of both cloud service provider and cloud service consumer. It regulates the Quality of service(QoS) between service provider and service consumers. It also comprises the cost of service with the level of QoS balanced by the service costs [43]. Every cloud service provider wants to adopt a technique that fulfills the QoS demands of consumers while keeping the cost low [44] [45]. On the other hand QoS based resource allocation focus on other areas such as response time, makespan, and throughput, as it naturally allows admission control procedures and input scheduling.
Authors have proposed techniques to accomplish QoS demands of service provider while there are few techniques developed to fulfill service consumer QoS demands. For instance, Popovici et.al[46] proposed service provider oriented resource allocation techniques to lower the cost and accomplish the QoS demnds(latency, stability) but did not focus on the Qos perimeters of 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 as, technique was Wu et.al[47] proposed a technique to focus on QoS demands of both service provider and service consumer. The goal of the authors was to lower the infrastructure price and SLA violations. While the authors have satisfied the QoS demands of service provider and service consumer but the execution time remains unaddressed. Similarly, a scheduling-based heuristic resource allocation technique was proposed by Emeakaroha et.al [48] to avoid SLA violation consequences by using different SLA parameters for application development. In their research they have considered the various parameters to limit the applications in real world system, but consumers would be more keen on parameters like, reaction time and handling time. The proposed technique have lower the cost and accomplish the QoS demands of both but did not address the execution time and resource utilization.
There are few techniques which consider the SLA parameters for the allocation of the resources.Ergu et.al[49] proposed a resource allocation technique in which the resources are

allocated using pairwise comparison of jobs according to their ranks and the analytic hierarchy procedure given the accessible resources and consumer's preferences like completion time, job cost and network bandwidth. The authors unable the focus on the important factors such as cost and resource utilization. The evaluation of QoS based auction based resource allocation techniques is shown in **Table 8**.

Table 8. Qos based resource allocation techniques parameters

Author(s) name	QoS demands of CSP	Cost	Resource Utilization	QoS demands of CSC	Execution time	User Satisfaction
F.I. Popovici [46]	↑	↓	–	↓	–	↓
L. Wu.[47]	↑	–	↓	↑	–	↑
V.C. Emeakaroha [48]	↑	↓	–	↑	–	↑
D.Ergu [49]	↓	–	–	↑	↓	↑

5.4 Scheduling

Scheduling based resource allocation techniques is crucial in cloud computing because of the significant resource cost and execution time. Different scheduling area and resource allocation parameters are considered in different categories of resource scheduling techniques. Scheduling based resource allocation can be divided into two subdivisions. 1) Priority: various parameters such as time, cost, no of processor requests of tasks are considered while allocating resources. 2) Round Robin: assigns resources on basis of first come first serve for fixed time called as time quantum.

5.4.1 Priority

Priority based resource scheduling technique assign the priority among the task based on the time, cost, no of processor request during resource allocation. Authors have proposed scheduling resource allocation tecniques based on parallel workload, SLA. such as, Xiaocheng et.al [50] proposed a technique for efficient response time of the system based on the virtual aware resource allocation. This technique divides the computing capacity into two levels, high priority virtual machine and low priority virtual machine. The authors have only focused on response time and execution time, and factors like cost resource utilization are ignored. Also, Pawar et.al [51] proposed a technique that considers important SLA parameters such as, processor time, memory utilization and network bandwidth. The proposed technique successfully allocates the resource dynamically and enhance resource utilization but the response time remains unaddressed. Futhermore, Lee.et.al[52] proposed a technique to solve the issue of scheduling in service request named dynamic priority scheduling algorithm (DPSA). This technique categorized the tasks into task units based on their specific requirements. The results displayed by the authors show that this technique gives the productive service and schedule the tasks efficiently but resource utilization factor was not considered.

On the other hand, X.Wu [53] proposed a task scheduling technique to decrease the execution time of the task based on the QoS. This technique analyzed the priority of the task based on its properties and stores it in the task queue. The proposed technique enables to reduce the execution time but the response time still remains the same.The evaluation of

Priority based resource allocation techniques is shown in **Table 9**.

Table 9. Priority based resource allocation techniques parameters

Author(s) name	Cost	Response time	Resource Utilization	Workload	Execution time	SLA	User Satisfaction
X.Liu[50]	–	↓	–	–	↓	↓	↑
C.S. Pawar[51]	↑	–	↑	↓	↓	↑	↓
Z. Lee[52]	↓	–	–	↓	↓	–	↑
X. Wu[53]	–	–	–	↑	↓	↑	↑

5.4.2. Round Robin

The Round Robin(RR) algorithm is developed for the tasks that arranged in the queue list where processing time is distributed among them. Due to this characteristic authors have used RR algorithm in the development of resource allocation technique. Such as, Abdulrazaq et.al [54] proposed a RR algorithm based resource allocation technique that allocates the processing time quantum to the tasks in the ready queue. The proposed technique calculates the dynamic time quantum of tasks based on the average burst time in a queue list. Extra time is allocated to complete the task in case the quantum time is not enough. This mechanism improved the system performance by minimizing the tasks context switches but resource utilization was not considered in the technique. Also, Mishra et.al [55] compares the static time quantum with the remaining burst time of a task after the first allocation in their proposed technique. Time quantum is reallocated in case of remaining time is less than one quantum else task is sent back to the waiting queue. The proposed technique has managed to decrease the response and execution time but cost was high. Furthermore, Siva et.al [56] proposed RR based scheduling technique that used the average burst time of task to calculate the time quantum. So, it can adjust to the task that needs extra time than the allocated quantum time.

Table 10. Round robin based resource allocation techniques parameters

Author(s) name	Cost	Energy	Resource Utilization	Response time	Execution time	User Satisfaction
A. Abdulrazaq[54]	↑	–	–	↓	↓	↑
M.K. Mishra[55]	↑	↓	↑	↓	↓	↑
G.S. N. Rao [56]	–	–	↓	↓	↓	↑
A. Noon[57]	↓	↓	–	↓	↓	↑

Execution time and response time is reduced with the help of the technique but the cost of resources was not mentioned. Subsequently, Noon[57] proposed an updated RR algorithm that calculates the dynamic time quantum depending on arrival time of tasks without any task arrangements. The technique was proved to be efficient as it lowers the cost, response and execution time but factor of maximum resource utilization was not discussed. The evaluation of round robin based resource allocation techniques is shown in **Table 10**.

5.5. Power

Data centers are consuming significant amout of power that lead the authors to develop resource allocation techniques to allocate the resources properly. An efficient resource allocation technique can not only minimize the power consumption, but also minimize the

operational cost. This paper categorized the power based resource allocation technique into 1) energy aware: energy efficient resource allocation is projected to produce financial advantages as well as the environmental harmony and 2) thermal-aware allocation: predicts the thermal impacts of a task placement and the resource allocation depend on the anticipated thermal impact.

5.5.1 Energy Aware

Energy aware resource allocation expects to expand benefit level execution measures under power dissemination and power utilization requirements. Energy aware resource allocation techniques have become critical in proper power utilization in datacenters. Few authors have proposed techniques to reduce the energy consumption by suitable placement of VMs. Such as, Dashti and Rahmani [58] introduced a novel practical swarm optimization based technique to improve energy efficient resource allocation by migrating virtual machines dynamically. One of the advantage of this technique was the minimum response time with balancing the load of virtual machines but execution time was remain unaddressed. Also, Gao.et.al.[59] solve the virtual machine placement by introducing Ant Colony based resource allocation technique. The aim of the authors is to reduce the resource wastage and decrease power consumption but the response time was not considered in the technique.
Few authors focused on green computing while developing the resource allocation techniques. For instance, Kansal and Chana **[60]** introduced an Artificial Bee Colony meta-heuristic technique to find the most suitable work hub. The proposed technique has decreased the energy utilization with contention among processor utilization and memory. Two workload types are considered in this technique 1) memory intensive and 2) CPU. This technique makes contributions to the green computing and helpful in increasing the satisfaction of cloud user but the response time was not considered as focus in the proposed technique by authors. Also, Yanggratoke[61] proposed a technique to reduce the energy consumption known as GRMP-Q protocol. The main objective of this technique was to maximize the workload on the server allocate the time duration of CPU to the tasks. The evaluation of energy aware resource allocation techniques is shown in **Table 11**.

Table 11. Energy aware resource allocation techniques parameters

Author(s) name	Cost	Energy	Workload	Resource Utilization	Response time	Execution time	User Satisfaction
S.E. Dashti[58]	↓	↓	–	↓	↓	–	↑
Y. Gao[59]	↓	↓	–	↓	–	–	↓
N.J. Kansal[60]	↑	↓	↑	↑		↓	↑
R. Yanggratoke[61]	↓	↓	↓	↑		–	↓

5.5.2 Thermal Aware

The temperature of physical machines is effected by the power consumption and that is why the performance and reliability of the system is affected[62]. Thermal aware techniques predicts the effect of increase in temperature on task employment and allocation of resources. Few authors have proposed techniques to control the thermal effect by decreasing the workload. Such as Anton Beloglazon et.al [63] proposed a technique that reallocates the virtual machines to the physical machines by keeping important QoS parameters of both service provider and consumer in prospect. Network bandwidth, CPU and RAM optimization were the main factors for allocation of virtual machines in their thermal aware

technique. While the main objective of the authors was the cooling system load and minimizing the workload of the excited nodes to avoid the hot spots, but the resource utilization and execution time were not addressed. Also, R .Ayoub [64] proposed a thermal aware resource allocation technique to transfer the workload from hot edges to cold edges. 1) socket level and 2) core level were the two levels of the proposed technique. The authors employed a schedular that takes performance, fan speed and temperature as an input and arranged the tasks at the socket level. This determines the temperature of every node and then based on the prediction it transfer the tasks from hot to cold node. The highlight of the work was energy and resource effeciency but the cost was high. Furthermore, Yuestu et.al.[65] proposed a technique to show that there is substantial irregularity in the processor's temperatures. That is main reason for difference in the power utilization of fans. The authors have successfully minimized the temperature of data center by arranging the scheduling of nodes properly but the workload and execution time was not addressed in technique.
Few techniques focused on the controlling the temperature of servers. For instance, Liu et.al [66] proposed a technique to reduce the power consumption by turning the servers on and off. To do this, the authors proposed a model which consists of the components like monitoring services, managed environment, a migration manager and front end that gives detail to the consumers. In the proposed model the authors particularly focused on discussing the live migration technique which limits the aggregate cost of optimum allocation of VMs. Similarly, authors in [67] proposed a thermal aware strategy based on RC-Thermal model [68] to decrease the maximum temperature of the HPC servers under stochastic workload. The evaluation of thermal aware resource allocation techniques is shown in **Table 12**.

Table 12. Thermal aware resource allocation technique parameters

Author(s) name	Cost	Energy	Workload	Resource Utilization	Execution time	User Satisfaction
A. Beloglazov[63]	↓	↓	↓	–	–	–
R .Ayoub [64]	↑	↓	↓	↑	–	↓
Yuestu[65]	↓	↓	–	–	↓	↑
S. Liu[66]	↓	↓	↑	–	–	↓
S. Liu[67]	↓	↓	↓	↑	↓	↑

6. Discussion

This section will discuss the evolution of the resource allocation techniques. Future directions are also available in this section.

6.1. Evolution

The design of a distributed system in general, and cloud infrastructure in particular, involves a variety of hardware and transmission media. Constantly evolving technology keeps on bringing more efficient and robust hardware solutions that perform better and transfer the data faster than previous systems. Similarly, evolution in the abstraction of distributed systems is helping the system designers and application developers in designing and developing large scale distributed systems efficiently and easily. This section discusses the evolution of cloud computing systems while focusing the resource allocation techniques.
In 2009, some important parameters of resource allocation techniques have been covered as Machina et.al [30] a proposed technique helps to reduce the execution time. Furthermore,

Özer et.al [73] proposed the resource allocation technique to cover the cost factor and Sotomayor et.al [20] proposed technique helps to reduce the response time. The focus of the study was multiple distributed resources and virtual infrastructure. The existing resource allocation techniques in that year could not cover all the basic parameters of resource allocation and resource utilization was not up to the mark for cloud service provider which leads the changes in techniques in coming years.

In 2010, the researchers felt the need to improve the utilization of resources as Ming et.al [74] and Mohsen et.al [75] proposed resource allocation techniques to increase profit of cloud provider by reducing the cost of resources. The Focus of study was process sharing, clustering and map reduce. The demand of cloud computing was increasing gradually so the need of new resource allocation techniques had also increased to keep the system up to date.

In 2011, with the introduction of new technology the researchers improve the resource allocation technique according to the demands of the consumers as Wu et.al[47] and Emeakaroha et.al [48] proposed the techniques to focus on QoS of cloud service provider and cloud service consumer. On the other hand, Zaman et.al [76] introduced the improved cost aware resource allocation techniques to reduce the cost of resource utilization. Subsequently, The researchers observed that the energy consumption has to be controlled in order to allocate the resources efficiently and support green computing so R. Yanggratoke[61] proposed the allocation technique to control the energy consumption in data centers. Furthermore, N. Abbas et.al [57] M. Rakesh et.al[71] proposed the resource allocation scheduling techniques to minimize the execution time and response time. Focus of study of these researches was response time, user satisfaction, VM heterogeneity and consolidation of workloads. User satisfaction was always one of the important objectives for resource allocation technique which comes to consideration strongly in 2011. Almost all the basic parameters of resource allocation techniques have been covered but there were still need of improvements.

In 2012, the researchers improved the resource utilization techniques considering the basic parameters used in previous years. For instance, Xin LU [39] proposed a technique to maximize the resource utilization. Subsequently, E. Maghawry [72] and Mishra et.al [55] proposed resource allocation techniques to reduce the execution time and energy consumption respectively. The Focus of these studies was Network Performance and virtual infrastructure. The workload on VMs was increasing as the more people started to use the services of cloud computing

In 2013, with the most number of people using cloud computing the researchers turned their focus to reduce the workload. for instance, Chandrashekhar et.al [51] proposed a technique to reduce the workload by live migration of VMs. Similarly, Zhen et.al [53] proposed a technique with the aim of minimize the response time. Parallel Workloads were the focus of study of these studies. As the new hardware to support new technology becomes expensive the need of cost effective resource allocation techniques increases.

In 2014, with new technology was emerging the researcher reconsider few basic parameters of resource allocation techniques. For instance, Wang et.al [27] proposed a technique to reduce the resource cost of resource utilization.similarly, A. Abdulrazaq et.al [54] proposed a technique to reduce the execution time of the tasks. Service level agreement and divisible task scheduling were main focus of study. The deficiency of power aware resource allocation techniques was the issue till this year.

In 2015, the issue of energy consumption was raised and Kansal and Chana [60] proposed the resource allocation technique to consume less energy, Saraswathi et.al [69] proposed a dynamic resource allocation technique to reduce the cost. The focus of study was heterogeneous Workloads and Clustering of Workloads. One of the important observation has

been made that the energy consumption can also control the cost factor of resources allocation techniques.

During 2016-2017 the improvements in the techniques continued to keep up with the new technology. Dashti[58] proposed the resource allocation techniques to control the cost by reducing the energy consumption. Similarly, Gupta[77] proposed a technique in which better resource utilization results in better power consumption. The focus of study was networks and clustering.

In 2018-2020 devices become more smart by connecting to the internet, which causes the more data to be saved on the cloud and task to allocate the resources become more complex. Fog and edge computing is utilized to allocate the resources more efficiently such as Qinglin et.al [4] proposed a hierarchy reference architecture for smart devices.

It has been observed that the parameter cost is influenced by the behavior of energy. The studied literature reveals that most of the times during the resource allocation lower energy consumption results into lower cost. But this is not the case in every technique and the inverse behavior between cost and energy has also been observed since the parameter cost also depends on other factors including resource utilization and workload etc. For instance, in [28] energy consumption is low but cost is high because of the high workload.

Cloud computing has emerged very quickly and become the very important part of modern era but there are still few applications and services which are do not get benefit from this technology due to the unacceptable latency, lack of location awareness and mobility support. Consequently the Edge and Fog computing systems were introduced in 2012 by Cisco to address the challenges of Internet of Things applications and to provide the reliable infrastructure to provide flexible resources at the edge of network[70]. A summarized view of evolution of resource allocation techniques is shown in **Fig. 4**.

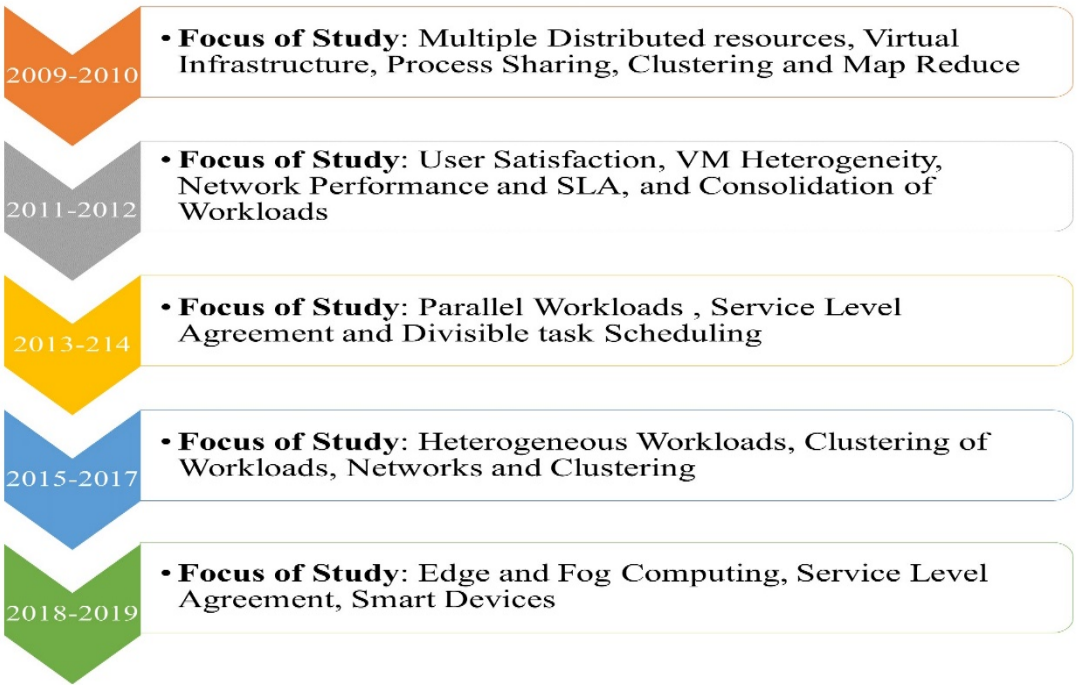

Fig. 4. Evolution of resource allocation techniques

6.2. Future Direction

Cloud computing enables the cloud service provider to allocate to the tasks according the consumer demands. Based on the literature reviewed in this study there are still gaps that should be covered in this domain.

Strategic: One emerging area for the authors to research is to find the details for detection of resource and workload for improved mappings for the scheduling of jobs and execution. 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, But at the same time artificial based resource allocation should also focus on cost factor and also improve the technique to make them suitable for larger systems as well.

Target Resource: The communication among the VMs should be minimum which belong to various servers in network aware resource allocation. The authors should develop the techniques to minimize the communication cost by finding the shortest path among the VMs.

Optimization: Profit of cloud service provider and consumer satisfaction should be the main focus of optimization based resource allocation techniques keeping the SLA negotiation between service provider and service consumer in view. Also, authors should also consider the penalty limitation in case system failures.

Scheduling: Based on the existing research there is need to assess the resource scheduling algorithms on real environment. According to the literature in this study dynamic resource scheduling is an open issue.

Power: According to this study green optimization resource allocation techniques in data center is an open issue in power aware resource allocation and comprehensive research is required in this domain. Authors should work on the relationship between varying workloads, while an attempt should be made to build frameworks that can minimize the trade-offs between SLA and provide energy efficiency in techniques.

7. Conclusion

This study has provided readers with an intellectual understanding of the crucial concepts of resource allocation in cloud computing. This study will also help the readers to find the gap between the existing resource allocation techniques and what is required to identify important issues for further investigation. Based on the discussion a high-level categorization of 77 research papers from 2007 to 2020 in resource allocation techniques domain has been presented in the form of a taxonomy. Apart from presenting a summary of the selected articles under proper heads, this article also presents a discussion involving an evolution in the resource allocation techniques during these years. It also presents promising future directions in the field of resource allocation in cloud computing. However, in order to develop more cost-effective allocation schemes there are other opportunities that require further study. The focus of resource allocation techniques in cloud computing in future should be to increase security, performance isolation, smooth virtual machine migration, interoperability, resilience to failure, graceful recovery and reducing the operational cost of data centers Lastly, it is envisaged that the services of cloud computing will become an integral part of almost all types and scales of information systems.

References

[1] R.Buyya, C.S.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.25, no.6,pp. 599-616, june 2009. Article (CrossRef Link)

[2] Gong S, Yin B, Zheng Z, Cai KY, "Adaptive Multivariable Control for Multiple Resource Allocation of Service-BasedSystems in Cloud Computing," *IEEE*, 13817–13831, 2019. Article (CrossRef Link)

[3] S.H.H. Madni, M.S. A. Latiff, Y.Coulibaly and S. 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. Article (CrossRef Link)

[4] Qi Q, Tao F. A, "Smart Manufacturing Service System Based on Edge Computing, Fog Computing, and Cloud Computing," *IEEE*, 7, 86769–86777, 2019. Article (CrossRef Link)

[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. Article (CrossRef Link)

[6] S. Singh and I. Chana, "A survey on resource scheduling in cloud computing: Issues and challenges," *Journal of grid computing*, vol.14, no.2, pp.217-264, 2016. Article (CrossRef Link)

[7] Lavanya, B. M., & Bindu, C. S., "Systematic literature review on resource allocation and resource scheduling in cloud computing," *international Journal of Advanced Information Technology (IJAIT)*, vol 6, no.4, pp. 1-15, 2016. Article (CrossRef Link)

[8] Hameed, A., Khoshkbarforoushha, A., Ranjan, R., Jayaraman, P. P., Kolodziej, J., Balaji, P., & Khan, S. U., "A survey and taxonomy on energy efficient resource allocation techniques for cloud computing systems," *Computing*, vol.98, no.7, pp.751-774, 2016. Article (CrossRef Link)

[9] Mohamaddiah, M. H., Abdullah, A., Subramaniam, S., & Hussin, M., "A survey on resource allocation and monitoring in cloud computing," *International Journal of Machine Learning and Computing*, vol. 4, no. 1, pp. 31-38, 2014. Article (CrossRef Link)

[10] Bhavani, B. H., & Guruprasad, H. S., "A comparative study on resource allocation policies in cloud computing environment," *Compusoft*, vol. 3, no.6, pp. 893, 2014.

[11] P. Mell and, T.Grance, "The NIST definition of cloud computing," *NIST Special Publication*, 2011. Article (CrossRef Link)

[12] S.T. Selvi, C. Valliyammai and V.N.Dhatchayani, "Resource allocation issues and challenges in cloud computing," in *Proc. of IEEE International Conference on Recent Trends in Information Technology*, pp 1-6, 2014. Article (CrossRef Link)

[13] ARM—the architecture for the digital world. http://www.arm.com/. Accessed 18 January 2020

[14] intel®AtomTMProcessor.http://www.intel.com/content/www/us/en/processors/atom/atom-processor.html. Accessed 18 January 2020.

[15] Memristor. http://www.memristor.org/. Accessed 18 January 2020

[16] R. E. Simpson, P. Fons, A. V. Kolobov, T. Fukaya1 M. Krbal, T. Yagi and J. Tominaga, "interfacial phase-change memory," *Nature nanotechnology*, vol. 6 , no. 8, pp. 501-505, 2011. Article (CrossRef Link)

[17] N. Ekker, T. Coughlin and J. Handy, "Solid State Storage 101 An introduction to Solid State Storage," *Storage Network Industry Association*, 2009.

[18] C. Guo, G. Lu, D. Li, H. Wu, X. Zhang, Y. Shi, C. Tian, Y. Zhang and S. Lu, "BCube: a high performance, server-centric network architecture for modular data centers," *ACM SIGCOMM Computer Communication Review*, vol. 39, no. 4, pp.63-74, 2009. Article (CrossRef Link)

[19] C. 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. Article (CrossRef Link)

[20] 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*, pp. 59-68, 2009. Article (CrossRef Link)
[21] C. Li and L.Y. Li, "Optimal resource provisioning for cloud computing environment," *The Journal of Supercomputing*, vol. 62, no. 2, pp. 989-1022, 2012. Article (CrossRef Link)
[22] 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*, vol. 26, no. 2, pp. 412-433, 2014. Article (CrossRef Link)
[23] 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. Article (CrossRef Link)
[24] 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. 1-6, 2012. Article (CrossRef Link)
[25] G. Wei, A.V. Vasilakos, Y. Zheng and N. Xiong, "A game-theoretic method of fair resource allocation for cloud computing services," *The journal of supercomputing*, vol. 54, no.2, pp.252-269, 2010. Article (CrossRef Link).
[26] Y. Liang, Q.P. Rui and J. Xu, "Computing resource allocation for enterprise information management based on cloud platform ant colony optimization algorithm," *advanced Materials Research*, vol. 791-793, pp. 1232-1237, 2013. Article (CrossRef Link)
[27] C.F. Wang, W.Y. Hung and C.S. Yang, "A prediction based energy conserving resources allocation scheme for cloud computing," in *Proc. of the IEEE International Conference on Granular Computing*, pp. 320–324. 2014. Article (CrossRef Link)
[28] S. Goutam and A.K. Yadav, "Preemptable priority based dynamic resource allocation in cloud computing with fault tolerance," in *Proc. of the IEEE International Conference on communication networks*, pp. 278-285, 2015. Article (CrossRef Link)
[29] 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. Article (CrossRef Link)
[30] 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*, pp. 1-8, 2009. Article (CrossRef Link)
[31] 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*, vo. 47, pp 3–14, 2012. Article (CrossRef Link)
[32] 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*, vol.7, pp. 1113-1118, 2007.
[33] J. Archer, A. Boehme, D. Cullinane, P. Kurtz, N. Puhlmann, J. Reavis, "Top Threats to Cloud Computing V 1.0," *Cloud Security Alliance*, 2010.
[34] J. Frey, "Network Management and the Responsible, Virtualized Cloud," *research rep*, 2011.
[35] G. Sun, V. Anand, H.F. Yu , D. Liao and L. Li, "Optimal Provisioning for Elastic Service Oriented Virtual Network Request in Cloud Computing," *IEEE GLOBECOM*, pp. 2517-2522. 2012. Article (CrossRef Link)
[36] X. Meng, V. Pappas and L. Zhang, "Improving the scalability of data center networks with traffic-aware virtual machine placement," *IEEE INFOCOM*, pp 1-9, 2010. Article (CrossRef Link)
[37] M. Alicherry and T,V. Lakshman, "Network aware resource allocation in distributed clouds," *IEEE INFOCOM*, pp.963-971, 2012. Article (CrossRef Link)

[38] A. Aldhalaan and D.A. Menascé, "Autonomic allocation of communicating virtual machines in hierarchical cloud data centers," in *Proc. of the IEEE International Conference on Cloud and Autonomic Computing*, pp. 161-171,2014. Article (CrossRef Link)

[39] 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.

[40] 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. Article (CrossRef Link)

[41] 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*, pp. 447-454, 2011. Article (CrossRef Link)

[42] Z. Abbasi, G. Varsamopoulos and S.K.S. Gupta, "Thermal aware server provisioning and workload distribution for internet data centers," in *Proc. of the nineteenth ACM international symposium on high performance distributed computing*, pp.130-141, 2010. Article (CrossRef Link)

[43] S, Son, G. Jung and S.C. 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. Article (CrossRef Link)

[44] S. Singh and I. Chana, "QoS-aware autonomic resource management in cloud computing: a systematic review," *ACM Computing Surveys (CSUR)*, vol. 48, no. 3, pp. 42, 2015. Article (CrossRef Link)

[45] Iqbal, S., Kiah, M. L. M., Dhaghighi, B., Hussain, M., Khan, S., Khan, M. K., & Choo, K. K. R., "On cloud security attacks: A taxonomy and intrusion detection and prevention as a service," *Journal of Network and Computer Applications*, 74, 98-120, 2016. Article (CrossRef Link)

[46] F.I. Popovici and J.Wilkes, "Profitable services in an uncertain world," in *Proc. of the 18th IEEE/ACM Conference On Supercomputing*, pp. 36, 2005. Article (CrossRef Link)

[47] 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*, pp. 195-204, 2011. Article (CrossRef Link)

[48] V.C. Emeakaroha, I. Brandic, M. Maurer and I. Breskovic, "SLA-aware application deployment and resource allocation in clouds," in *Proc. of the 35th annual IEEE computer software and applications conference workshops*, pp. 298-303. 2011. Article (CrossRef Link)

[49] D.Ergu, G. Kou, Y. Peng, Y. Shi and Y. Shi, "The analytic hierarchy process: task scheduling and resource allocation in cloud computing environment," *The Journal of Supercomputing*, vol. 64, no. 3, pp. 835–848, 2013. Article (CrossRef Link)

[50] 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, pp. 14-22, 2009. Article (CrossRef Link)

[51] C.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*, pp. 311-316, 2013. Article (CrossRef Link)

[52] Z. 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*, vol. 9, pp. 4665-4669, 2011. Article (CrossRef Link)

[53] X. Wu, M. Deng, R. Zhang, B. Zeng and S. Zhou, "A task scheduling algorithm based on QoS-driven in Cloud Computing," *Procedia Computer Science*, vol. 17, pp. 1162–1169, 2013. Article (CrossRef Link)
[54] A. Abdulrazaq, E. A. Saleh, Junaidu and B. Sahalu, "A new improved round robin (NIRR) CPU scheduling algorithm," *international Journal of Computer Applications*, vol. 90, no. 4, pp. 27-33, 2014. Article (CrossRef Link)
[55] M.K. Mishra, "AN IMPROVED ROUND ROBIN CPU SCHEDULING ALGORITHM," *Journal of Global Research in Computer Science*, Vol.3, No. 6, pp. 64-69, 2012.
[56] 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, pp. 1480-1485, 2015. Article (CrossRef Link)
[57] 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.
[58] 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. Article (CrossRef Link)
[59] Y. Gao, H. Guan, Z. 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. Article (CrossRef Link)
[60] 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. 5, pp. 1207-1225, 2015. Article (CrossRef Link)
[61] 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*, pp. 1-9, 2011.
[62] A. Beloglazov and R. Buyya, Y.C. Lee and A. Zomaya, "Chapter 3 - A taxonomy and survey of energy-efficient data centers and cloud computing systems," *Advances in computers*, vol. 82, no. 47-111, 2011. Article (CrossRef Link)
[63] 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*, pp. 826-831, 2010. Article (CrossRef Link)
[64] 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. 1359-1372, 2011. Article (CrossRef Link)
[65] 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. Article (CrossRef Link)
[66] L. Liu, H. Wang, X. Liu, X. Jin, W.B. He, Q.B. Wang and Y. Chen, "GreenCloud: a new architecture for green data center," in *Proc. of the 6th ACM international conference industry session on Autonomic computing and communications industry session*, pp. 29-38, 2009. Article (CrossRef Link)
[67] S. Liu and M. Qiu, "Thermal-aware scheduling for peak temperature reduction with stochastic workloads," in *Proc. of the 16th IEEE Real-Time and Embedded Technology and Applications Symposium*, pp. 59-62. 2010.
[68] R. Ranjan, A. Harwood and R. Buyya, "SLA-based coordinated super scheduling scheme for computational Grids," in *Proc. of the 8th IEEE International Conference on Cluster Computing*, pp. 1-8, 2006. Article (CrossRef Link)

[69] A.T. Saraswathi, Y.R.A. Kalaashri and S. Padmavathi, "Dynamic resource allocation scheme in cloud computing," *Procedia Computer Science*, vol. 47, pp. 30-36, 2015. Article (CrossRef Link)

[70] Y. Xiao, M. Krunz, "QoE and power efficiency tradeoff for fog computing networks with fog node cooperation," in *Proc. of the IEEE International Conference on INFOCOM*, pp. 1-7, 2017. Article (CrossRef Link)

[71] 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. Article (CrossRef Link)

[72] E. Maghawry, R. Ismail, N. Badr and M. Tolba, "An enhanced resource allocation approach for optimizing sub query on cloud," in *Proc. of International Conference on Advanced Machine Learning Technologies and Applications*, pp. 413-422, 2012. Article (CrossRef Link)

[73] A.H. Özer and C. Özturan, "An auction based mathematical model and heuristics for resource co-allocation problem in grids and clouds," in *Proc. of the fifth International Conference on Soft Computing, Computing with Words and Perceptions in System Analysis, Decision and Control*, pp. 1-4, 2009. Article (CrossRef Link)

[74] M Mao, J Li and M. Humphrey," Cloud auto-scaling with deadline and budget constraints," in *Proc. of the 11th IEEE/ACM international conference on grid computing*, pp. 41–48, 2010. Article (CrossRef Link)

[75] M.A. Salehi and R. Buyya, "Adapting market-oriented scheduling policies for cloud computing," in *Proc. of the International Conference on Algorithms and Architectures for Parallel Processing*, pp. 351-362, 2010. Article (CrossRef Link)

[76] S. Zaman and D. Grosu, "Efficient bidding for virtual machine instances in clouds," in *Proc. of the 4th IEEE International Conference on Cloud Computing*, pp. 41-48, 2011. Article (CrossRef Link)

[77] R.S. Jha and P. Gupta, "Power & load aware resource allocation policy for hybrid cloud," *Procedia Computer Science*, vol. 78, pp. 350-357, 2016. Article (CrossRef Link)

[78] B.Fekade, T.Maksymyuk and M.Jo, "Clustering hypervisors to minimize failures in mobile cloud computing," *Wireless Communications and Mobile Computing*, vol. 16, no.18, pp.3455-3465, 2016. Article (CrossRef Link)

[79] D.Satria, D.Park and M.Jo, "Recovery for overloaded mobile edge computing," *Future Generation Computer Systems*, vol. 70, pp. 138-147, 2017. Article (CrossRef Link)

MUHAMMAD FARAZ MANZOOR received his BS(Hons.) degree from University of Punjab, Pakistan in 2016, while MS degree from University of Management and Technology, Pakistan in 2018. He is currently pursuing the Ph.D. degree with the University of Management and Technology Lahore, Pakistan, under the supervision of Dr. Adnan Abid and co-supervision of Dr. M. Shoaib Farooq. He is currently working as Head of Computer Science Department in Punjab Group of Colleges. His research interests include IOT, Internet of Vehicles, Distributed Systems and Computer Science Education.

MUZAMMIL HUSSAIN received the Bachelor of Science degree (Hons.) in computer science from COMSATS University Islamabad, in 2013, and the Ph.D. degree from the University of Malaya, Kuala Lumpur, Malaysia. From 2013 to 2017, he was a Research Associate with the Department of Computer Systems and Technology, Faculty of Computer Science and Information Technology, University of Malaya. He is currently an Assistant Professor with the Department of Computer Science, School of Systems and Technology, University of Management and Technology Lahore, Pakistan. He is also the Director of Graduate Studies and Research with the School of Systems and Technology, University of Management and Technology. He led or was a member for many funded research projects and he has published more than 15 research articles in prestigious international conferences and journals. His potential research areas include Web services, Web of Things, network security, FinTech security, blockchain, android security, bioinformatics, the IoT security, SDN security, and Internet security. He received the Bright Sparks Program Scholarship from the University of Malaya for his Ph.D. studies.

ADNAN ABID obtained his BS Degree from National University of Computer and Emerging Science, Pakistan in 2001. While, he obtained his MS degree in Information Technology from National University of Science and Technology, Pakistan in 2007. He spent one year in EPFL, Switzerland to complete his MS thesis. Later on, he did his PhD in Computer Science from Politecnico Di Milano, Italy in 2012. His research interests include Computer Science Education, Information Retrieval, and Data Management. Currently, he is working as Associate Professor in the Department of Computer Science in University of Management and Technology, Pakistan. He has several publications in different international journals and conferences. He has served as reviewer in many international conferences and journals, and is also Associate Editor of IEEE Access journal.

MUHAMMAD SHOAIB FAROOQ obtained his MS degree from Government College University, Pakistan while PhD degree from Abdul Wali Khan University, Pakistan. He is working as Associate Professor of Computer Science at University of Management and Technology, Lahore. He is also affiliate member of George Mason University, USA. He possesses more than 24 years of teaching experience in the field of Computer Science. He has published many peer-reviewed international journal and conference papers. His research interests include Theory of Programming Languages, Big Data, IOT, Internet of Vehicles, Machine Learning, Distributed Systems and Education.

.

UZMA FAROOQ received the B.S. degree in Computer Science from University of the Punjab in 2005. She obtained MS degree in Computer Science from National University Computer and Emerging Sciences, Pakistan in 2007. Currently, she is working as Assistant Professor in the Department of Computer Science at University of Management and Technology, Pakistan. She is teaching basic and advanced computer programming courses in the undergraduate programs in Computer Science and Software Engineering.