Cluster Comput
DOI 10.1007/s10586-016-0684-4

CrossMark

Recent advancements in resource allocation techniques for cloud computing environment: a systematic review

Syed Hamid Hussain Madni[1] · Muhammad Shafie Abd Latiff[1] · Yahaya Coulibaly[1] · Shafi'i Muhammad Abdulhamid[1]

Received: 21 January 2016 / Revised: 28 September 2016 / Accepted: 16 November 2016

Abstract There are two actors in cloud computing environment cloud providers and cloud users. On one hand cloud providers hold enormous computing resources in the cloud large data centers that rent the resources out to the cloud users on a pay-per-use basis to maximize the profit by achieving high resource utilization. On the other hand cloud users who have applications with loads variation and lease the resources from the providers they run their applications within minimum expenses. One of the most critical issues of cloud computing is resource management in infrastructure as a service (IaaS). Resource management related problems include resource allocation, resource adaptation, resource brokering, resource discovery, resource mapping, resource modeling, resource provisioning and resource scheduling. In this review we investigated resource allocation schemes and algorithms used by different researchers and categorized these approaches according to the problems addressed schemes and the parameters used in evaluating different approaches. Based on different studies considered, it is observed that different schemes did not consider some important parameters and enhancement is required to improve the performance of the existing schemes. This review contributes to the existing body of research and will help the researchers to gain more insight into resource allocation techniques for IaaS in cloud computing in the future.

✉ Syed Hamid Hussain Madni
madni4all@yahoo.com

Muhammad Shafie Abd Latiff
shafie@utm.my

Yahaya Coulibaly
coulibaly@utm.my

Shafi'i Muhammad Abdulhamid
shafii.abdulhamid@futminna.edu.ng

[1] Faculty of Computing, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia

Keywords Resource management · Resource allocation · Resource selection · Resource scheduling · Resource utilization · IaaS cloud

1 Introduction

Resource management is the procedure of assigning virtual machines, computing processes, networks, nodes and storage resources on-demands to a set of applications in cloud computing environment. Through this way, the whole resources are equally assigned between the infrastructure providers and users of cloud. Cloud providers provide resources efficiently within the limits of the service level agreements (SLAs) [1] to the cloud users. These resources are accomplished with the support of virtualization technologies, which assist them in statistical multiplexing of resources for the clients and applications.

Further, resource management helps in synchronization of resources which is emphasized by the management actions and accomplished by the both cloud providers and users. It is the process of resource allocation from resource providers to the resource users on the basis of pay-per-use. It also allows to assign and re-assign resources from the cloud providers to the cloud users where the cloud user can efficiently use the available resources of IaaS [2,3].

In a cloud computing environment there are two actors playing an important role these are cloud providers and cloud users From the perspective of a cloud provider, the providers have a large number of computing resources in their large data centers and they rent out these resources to the users on a pay-per-use basis to maximize the revenue by attaining

Journal: 10586 Article No.: 0684 MS Code: CLUS-D-16-00024.1 ☐ TYPESET ☐ DISK ☐ LE ☐ CP Disp.:2016/11/30 Pages: 45 Layout: Large

Fig. 1 Basic cloud environment

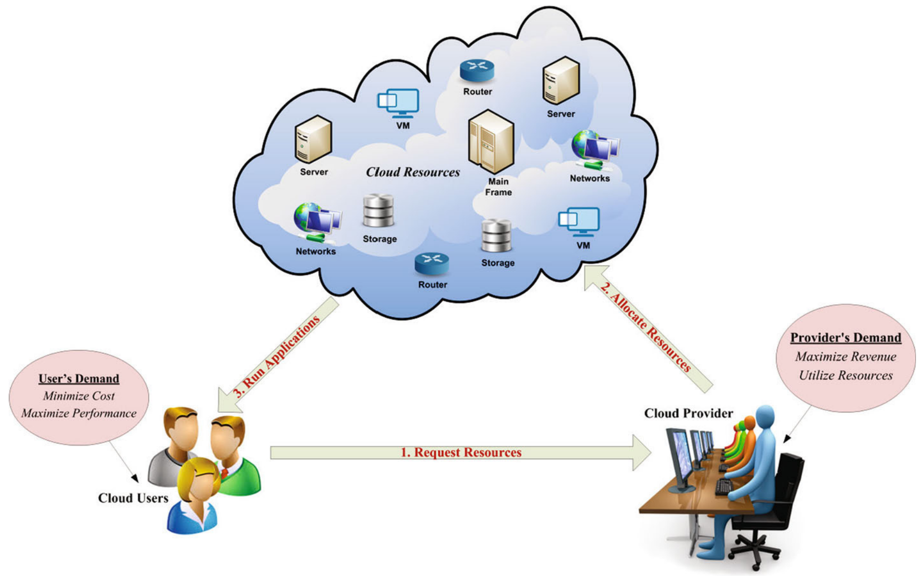
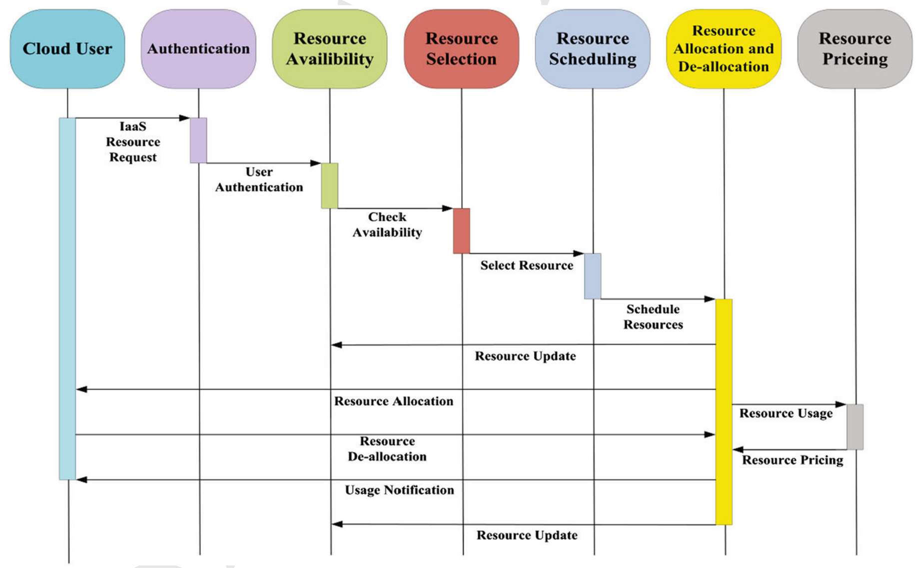


Fig. 2 Process of resource allocation



high resource utilization Resources are also in demand for the cloud users and applications with dynamic nature which are predicted by them. For the cloud users, who have applications with fluctuating loads lease the resources from the cloud providers and run their applications within minimum expenses Every cloud user wants a number of resources for a particular task or cloudlet that can maximize the performance and have to be finished on time as shown in Fig. 1.

In cloud computing, resource management is totally based on resource allocation. Resource allocation is the procedure which is based on the distribution of accessible resources to the required cloud application on the Internet in a systematic way [4,5] as depicted in Fig. 2. Moreover, IaaS plays an important role in the allocation of resources on-demands by supporting the pre-defined resource allocation policies to the cloud users. However, if the resources are not allocated on-demands to the cloud users, their services will not be long lasting. The solution of this problem is to permit the cloud providers to organize the resources of each module separately. Therefore, resource allocation is considered as a portion of resource management, and it shows a remarkable character in the allocation of resources economically and effectively.

Resource allocation in IaaS is a challenging issue due to management and provision of resources in cloud computing. Numerous research contributions have been made, which are focused on limited resources, resource heterogeneity, environmental requirements, locality limitations and on-demand resources allocation [6–13]. Moreover, the research requires an efficient and effective resource allocation process that is optimum to cloud computing environment.

Journal: 10586 Article No.: 0684 MS Code: CLUS-D-16-00024.1 TYPESET DISK LE CP Disp.:2016/11/30 Pages: 45 Layout: Large

1.1 Significance of resource allocation

Resource allocation policies, strategies and algorithms help assign or transfer resources that support the both cloud providers and users. The following are the resource allocation conditions, which should not be adopted during allocation to the users [14,15]:

1. Allocation of on-demand (extra) resources to the user that violates the policies of resource allocation.
2. It is under provisioning situation when the cloud provider assigns fewer resources to the user.
3. Resource congestion occurs, when two or more cloud users try to acquire the same resource in a specific instance.
4. Resource destruction occurs, when a countable number of resources are available in cloud, but the cloud provider does not fulfill the demands of the cloud users.
5. Resource deficiency occurs when there are limited numbers of resource in cloud.

Previous survey and review articles in this research field investigate resource management, resource allocation, resource scheduling, energy efficiency, load balancing, resource provisioning, VM allocation, QoS, and security in cloud computing. Hence, our major contributions in this review paper are as follows:

- We put forward a systematic literature review of resource allocation techniques for cloud computing system.
- We present taxonomy of current advances in resource allocation techniques, while emphasizing on their strengths and weaknesses.
- We chronicle the performance metrics employed for evaluating the prevailing approaches.
- We describe the previously mentioned future research works that guide in shaping the direction for present and future research.

The aim of this categorization is for building the foundations for future scholars in cloud computing system. The purpose of this review is to analyze the prevailing techniques and for understanding their focus of work. This is essential to develop additional suitable techniques which could be an enrichment of the existing techniques or taking benefits from earlier studies. The brief prefatory part of the review is followed by a structured argument spanning over the sections as follows: Sect. 2 discusses the related work of resource allocation in cloud computing. Section 3 elaborates the research methodology implemented in the paper, whereas Sect. 4 analyses and categorizes the existing studies of resource allocation for IaaS cloud computing. The resources and parameters used to evaluate existing literatures are presented and analyzed in Sect. 5. In Sect. 6, we present the future research areas in cloud computing environment, while last Sect. 7 summarizes the conclusion and provide recommendations for further research in this direction.

2 Related works

There is an increasing interest being shown by the global research community on resource allocation in cloud computing. The current researches and reviews are drawing the attention of researchers and practitioners towards resource allocation attainment. Therefore, this review presents existing contributions which have been made in resource management for IaaS cloud computing. Manvi and Krishna Shyam [6] focus on resource adaptation, allocation, provisioning and mapping. It is perceived that there are many issues to be addressed in cloud resource management with respect to flexibility, scalability, adaptability, customization and reusability. Moreover, Bi et al. [16] also investigate various parameters such as delay, bandwidth overhead, computation overhead, reliability, security and Quality of Experience.

Similarly, Chana and Singh [7] state that major problem concerned with resource allocation is assigning and scheduling of the resources in an efficient way to achieve the QoS performance goals as identified by SLA. Moreover, instead of cloud computing infrastructure, it is mandatory for the cloud providers to observe and examine the modifications in resource demand. Consequently, a cloud provider helps in the allocation and transfer of resources in CPUs and takes a decision regarding the acceptance of upcoming request while keeping in view the available resources [17]. However, elements which monitor the accessibility of system resources plays a significant role in observing the QoS requirements and user request, resources usage pricing, follow up and improvements via determining the real usage of resources and ends up by making the resources allocation a complex task.

Resource allocation has gained more relevance in cloud computing as its policies and algorithms affect the cloud performance and cost. Ma et al. [8] present five key issues in cloud computing based on energy aware provisioning, locality aware task scheduling, reliability aware scheduling, Software as a Service provisioning and workflow scheduling. However, these are further sub-divided as cost provisioning, performance provisioning and cost performance provisioning. Cloud resource policies regarding allocation and scheduling are described while keeping in view the concerned parameters. Therefore, a detailed analysis of five specified problems along with descriptive algorithm has been done. Regardless of this, future research in resource computing should further address the challenges of allocation and scheduling of resources regarding data locality in task scheduling and load balancing in cloud computing [18].

Journal: 10586 Article No.: 0684 MS Code: CLUS-D-16-00024.1 TYPESET DISK LE CP Disp.:2016/11/30 Pages: 45 Layout: Large

The guidelines and directions related to energy aware resource allocation of information communication technology (ICT) in cloud computing data center is identified [19], a modern and well-equipped research organization concerned about policies regarding resource adaption, objectivity, methodologies concerning allocation and operation. However, it plays an important role in the classification of current literature and application of procedures for analytical surveys as the current literature debates regarding its advantages and disadvantages. However, resource allocation is considered as an interesting issue from the cloud provider's point of view [20]. Keeping in view, the various QoS levels cloud providers normally deal with virtualized resources. Cloud computing shares the physical resources in the form of virtual resources among the cloud users. In view of this, allocation policies and strategies need to allocate the resources in a way to overcome the demand of users in an economical and cost-effective way, side by side fulfilling the QoS prior requirements [21].

The research work carried out in Huang et al. [22] present the current resource allocation policy, job scheduling algorithms along the concern issues of cloud environment and propose a methodology based on the solution. However, performance improvements concern with detail resource allocation strategy consists of failure law, vibrant resources for various assignments concerned with integrity ant colony optimization algorithm for resource allocation. Moreover, dynamic scheduling algorithm stands on the threshold, optimize genetic algorithm with multifaceted and enhance ant colony algorithm for job scheduling. Because of the convenience of predictable resources, it is necessary for the cloud providers to organize and distribute the resources to the cloud users on fluctuating demands [9]. An efficient resource allocation procedure always fulfills the standards that are QoS aware resource utilization, less expense and energy consumption. The main motive of resource allocation is to increase the revenue for the cloud providers and to reduce the charges for the cloud users in cloud computing.

The specifications of SLA as it demonstrate a suitable level of granularity named as tradeoffs between the clarity and intricacy. In this regard, to overcome the expectations of consumers it aims at simplified verification and evaluation procedure which is forced by resource allocation mechanism on cloud [23]. However, few researchers show the survey results of various methodologies for the solution of resource allocation problem [10]. Moreover, resource allocation methodology consists of dynamic self-directed resource management to provide the scalable, flexible and reduced allocation cost and size. It is multi-agent system consist of compound judgment analysis criteria, graph methods, optimization, simulation prediction, service oriented architecture and theoretical formulation.

Mohan and Raj [11] explain that capability of allocation, its resources management and energy utilization ends up with an exigent strategic goal. Moreover, some strategies have been specified for future researchers. Thus, a suitable methodology regarding the distribution of virtual machines plays a vital role in the maximization of energy conservation as it can be further extended to high level of competencies [24]. Moreover, SLA parameters have the capacity to be improved in various ways in order to enhance the efficiency level. Regardless of this, scheduling and application exploitation has attained remarkable attention in cloud, for realizing the objectives of efficient energy transmission in resource allocation.

A survey of the state of the art in the VM allocation problem relating to problem models and algorithms is presented by Mann [25]. Further, survey used the problem formulations, optimization algorithms, highlights the strengths and weaknesses, and point out areas that need to be further researched. Hameed et al. [12] and Akhter and Othman [26] classify the open challenges related to energy efficient resource allocation. Firstly summarize the problem and existing methods available for this purpose. In addition, available methods previously proposed in the literature are precised, with the benefits and drawbacks of the existing techniques. Besides numerous resource allocation approaches in literature emphasizes on open concern issues and future guidelines. Mustafa et al. [13] present a comprehensive review of resource management techniques that is based on the major metrics and illustrates their comprehensive taxonomy based on the distinct features. It points out the evaluation parameters and steps that are used to analyze the resource management methods.

3 Research methodology

This section presents the research steps followed to perform this review. It highlights the motivating factors for conducting this systematic review according to Moher et al. [27] and elaborates the review methodology in detail by SLR guidelines of Kitchenham et al. [28]. According to these authors, the research methodology for systematic review should contain the research questions which the current study attempts to answer. Various strategies are employed for searching the most significant research works like search strings and the chosen digital libraries. Finally, the selection of the existing studies is done through a set criteria.

3.1 Data sources

The review procedure involves the formulation of research questions, a search of different databases, analysis and identification of the different techniques. The research methodol-

Journal: 10586 Article No.: 0684 MS Code: CLUS-D-16-00024.1 TYPESET DISK LE CP Disp.:2016/11/30 Pages: 45 Layout: Large

Table 1 Databases sources

Source	URL
ACM Digital Library	URL:http://dl.acm.org/
IEEE Explore	URL:http://ieeexplore.ieee.org/
DBLP	URL:http://dblp.uni-trier.de/
Google Scholar	URL:https://scholar.google.com/
Science Direct	URL:http://www.sciencedirect.com/
Scopus	URL:https://www.scopus.com/
Springer	URL:http://www.springer.com/
Taylor & Francis	URL:http://taylorandfrancis.com/
Web of Science	URL:https://apps.webofknowledge.com/
Wiley Online Library	URL:http://onlinelibrary.wiley.com/

ogy adopted in this paper also requires finding of relevant papers from a variety of databases (such as ACM Digital Library, IEEE Explore, DBLP, Google Scholar, Science Direct, Scopus, Springer, Taylor & Francis, Web of Science and Wiley Online Library) as shown in Table 1 and a list of different questions that are to be addressed in Table 2. It is further refined by the identification of primary studies, then applying certain inclusion criteria and after that evaluating the results.

3.2 Search strategy

This study started in Jan 2015 and decision for searching for the required research works from Jan 2008 to Dec 2015. In generally, cloud computing publications started around 2008, so we decided to search for researches on resource allocation in cloud computing in the period from 2008 to 2016.

On the basis of the topic and the proposed research questions, we define the searching keywords as a first step to formulate the search string. We are also considered the search terms "resource allocation", "Infrastructure as a Service", "IaaS", "cloud" and "cloud computing" as the main keywords. We use the logical operators AND and OR for connecting the main keywords. Eventually, after several tests, we choose the following search string that gives us the sufficient amount of related research studies: ("resource allocation" * "Infrastructure as a Service" + "IaaS" * cloud + "cloud computing").

Quick search strategy is used to make this research up-to-date and well-intentioned in the area of cloud computing. For this purpose, we have used the quick search strategy to add recent 2015–2016 publications for this research by using the filtering tools in the databases. After using the quick search strategy, we considered the publication from 2008 to 2016 overall.

3.3 Research questions

Table 2 lists the different research questions and their corresponding motivations.

Table 2 Research questions and motivations

Questions	Motivation
Why resource allocation is necessary for cloud computing?	It helps to understand the implications of resource allocation in IaaS cloud computing
How resource allocation is beneficial for IaaS cloud computing?	It helps to enhance the benefits and achievement for the both cloud users and providers in IaaS cloud computing
What are the existing strategies, policies, and algorithms for realizing resource allocation in IaaS cloud computing?	Many techniques are discussed to ensure resource allocation in IaaS cloud computing with a thorough review, categorization and comparison of existing techniques
Which resources and parameters are more considered during resource allocation?	It helps to analyze the recourses and parameters that are more important for the cloud users and providers in resource allocation for cloud computing
How optimum resource allocation is achieved through existing strategies, policies and algorithms?	It helps in locating the ambiguities which are responsible for resource allocation in cloud computing
Which research gap remains unaddressed in the field of resource allocation in IaaS cloud computing?	This review article will help future researchers to understand clearly the current status, need and future requirements for resource allocation in IaaS cloud computing

3.4 Study selection procedure

The methodology used in this review starts with the definition of the research questions listed in Sect. 3.2. The search keywords help in refining the selection and search process. Only studies written in the English language are considered. After finding appropriate literatures, an analysis of resource allocation in IaaS is conducted for this systematic review.

The study selection process is shown in Fig. 3. The search process ends very comprehensively to ensure the completeness of this review. Most of the studies were screened out because their titles were not relevant to the selection criteria or abstracts were not related to be incorporated in this review. As shown in Fig. 3, the initial search resulted in a total of 1332 studies, which were condensed to 426 studies on the basis of their titles, and 229 studies on the basis of their abstracts. After that, 229 selected studies were reviewed thoroughly for obtaining a final list of 159 studies on the basis of their content.

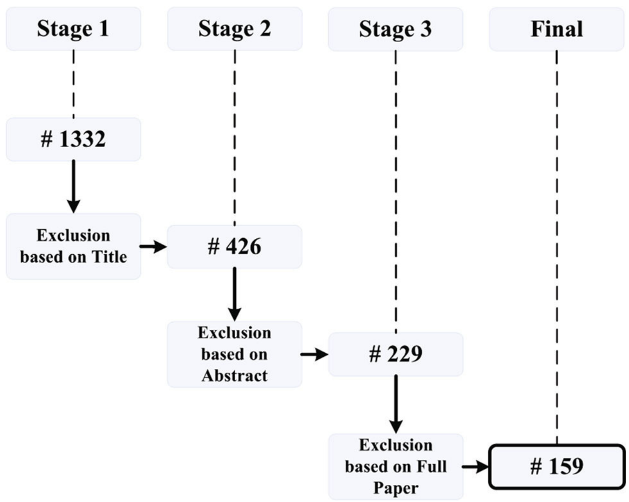


Fig. 3 Study selection process

Table 3 Studies inclusion/exclusion criteria

Inclusion criteria	Exclusion criteria
The study focuses on resource allocation in cloud computing	The study does not focus on other resource management issues in cloud computing
The study considers the Infrastructure as a Service (IaaS) for resource allocation only	The study does not consider the Software as a Service (SaaS) or Platform as a Service (PaaS)
The study is written in English only	The study is not written in the English language
The study is peer reviewed and published in scholarly society	The study is not peer reviewed such as workshop, descriptions and technical reports
The study is published in well-reputed Journals or Conferences	The study is not published in the form of books, abstracts, editorials or keynotes

3.5 Studies inclusion/exclusion criteria

For selecting the related important studies, the inclusion and exclusion criteria are applied. On the basis of the set criteria, the primary research studies are selected after going through the title, abstract and full content of the studies for ensuring that the results are related to the research area of this current research work. The inclusion and exclusion criteria, which used in this current systematic review is defined in Table 3.

4 Analysis of the studies

In this section, the review findings are explained. The key characteristics of existing resource allocation techniques for IaaS cloud computing are listed. The techniques are grouped into two main groups including strategic based and parametric based resources allocation. Furthermore, classify these groups into different subcategories and detailed classifications are presented as shown in Fig. 4. The objective of this categorization is to build the base of the resource allocation for future research in cloud computing.

4.1 Strategic based resource allocation

Strategic based resource allocation are further categorized into three groups including the artificial intelligence resource allocation, dynamic resource allocation and predicted resource allocation on the basis of techniques' behaviour and environment. The details of the categorization listed above are as follow.

4.1.1 Artificial intelligent resources allocation

Artificial intelligence is an area of cloud computing that emphasizes the creation of intelligent methodology that work and react like humans for resource allocation. This encompasses the application and development of artificial intelligent techniques, including resource allocation into aspects of autonomous and intelligent systems, nature-inspired intelligent systems, aspects of operational research, machine learning, neural networks, agent based system and expert systems [18]. With artificial intelligence, the chances of error and failure rate are almost zero, greater precision and accuracy are achieved for resource allocation in IaaS cloud computing.

Infrastructure as a service (IaaS) is responsible for the right to use to computing resources by establishing a virtualized cloud environment. Resources are easily leased to the cloud users. Still, due to a finite amount of resources, cloud provider cannot fulfill all the leases. Panda and Jana [29] recommend an algorithm for resource allocation in IaaS cloud, which is designed by using the innovative method of the alert time. Firstly, this one deals with the alert time to distribute the leases and then services transaction to reorganize the previously existing leases in case a lease is not scheduled through the alert time. By this tactic, resource allocation advance to provision the sensitive deadline leases by decreasing the denial of the lease, in discrepancy to dual current algorithms via Haizea. Correspondingly, Shyam and Manvi [30] propose an efficient resource allocation scheme using cloud provider's resource agent and cloud user's task agent in IaaS Cloud. With maximizing the resource utility, reducing the total cost, and preserving the QoS, the minimum usage of the amount of VMs is ensured. The Best Fit method increases the ratio of VM placement, which provides benefits to the both cloud providers and users. As well, the allocation of VMs with numerous resources determines a vital portion in enhancing the energy efficiency and performance in cloud data center. It helps in minimizing the usage of energy in

Journal: 10586 Article No.: 0684 MS Code: CLUS-D-16-00024.1 TYPESET DISK LE CP Disp.:2016/11/30 Pages: 45 Layout: Large

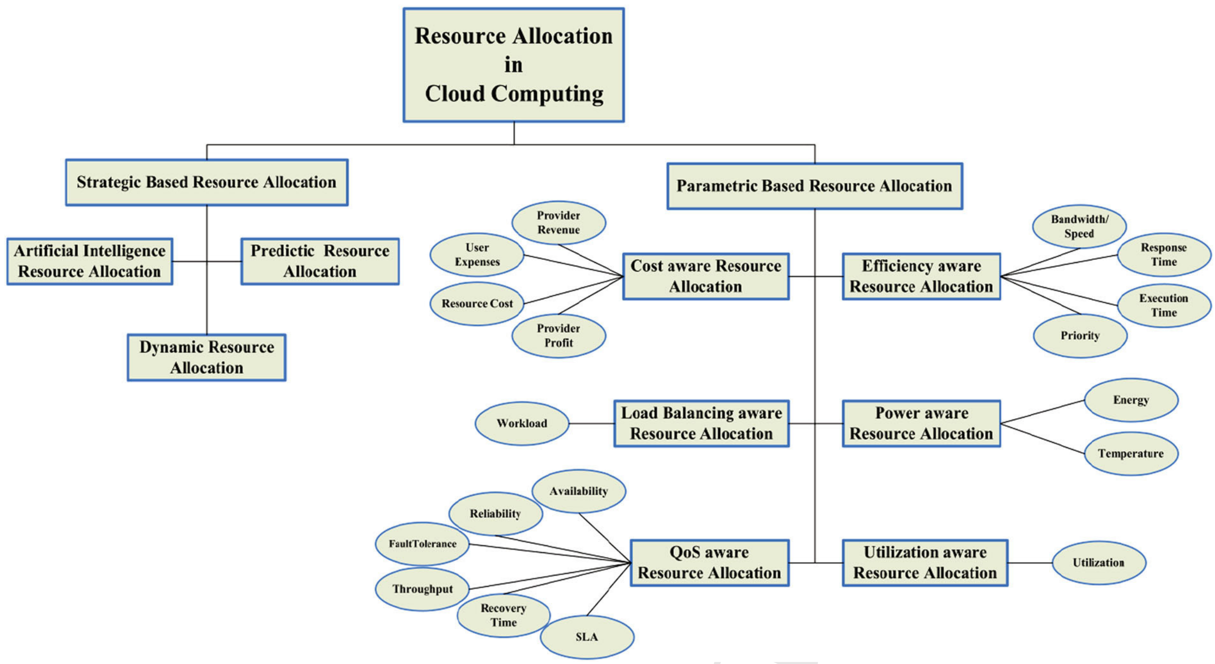


Fig. 4 Categorization of resource allocation in cloud computing

the data center. The Particle Swarm Optimization algorithm proficiently enhances the energy efficiency for VM allocation with numerous resources. But the techniques consider only the resource including processing and storage [31].

In [32], a innovative architecture for IaaS cloud computing system where the VM allocation of VMs are performed by genetical weight maximized the neural network. In such condition, the load of each PM in the data center is based on the information of resources. The neural networking forecast the load of PM in data center in future depends on past loads. It helps in the allocation of VM for choosing the right PM. The evolution is performed on the basis of the performance of genetical weight maximized Back Propagation Neural Network (BPNN), Elman Neural Network (ELNN) and Jordan Neural Network (JNN) for accurate forecasting. Meanwhile in [33], the resource optimization and management in the existing state of the art is used by Ant Colony Optimization (ACO), which fulfills the requirement of cloud computing infrastructure. The proposed algorithm predicts in advance the available resources and makes estimation of the required bandwidth. Moreover, it also guesses network quality and response time. However, Li and Li [34] present the combined optimization of efficient resource allocation for Software as a service (SaaS) and Infrastructure as a service (IaaS), accomplished with an iterative algorithm in cloud computing. Suggested joint optimization algorithm for proficient resource allocation is compared with additional existing algorithms, experimental results show a better performance.

The resource allocation and its management in cloud computing are the major challenging tasks in the current research. The numerous contributions have been done to address the problems of cloud computing environment. Therefore, Vernekar and Game [35] presents a Component Based Resource Allocation Model which uses the concept of Hierarchical P2P scheme. The Hierarchical P2P scheme is based on Metascheduler and Superschedule. The various virtual organizations (VOs) work as grid backbone for resource distribution in cloud computing among the users. The VOs are comprised of various nodes with the highest confirmation such as Metascheduler and Superscheduler. The Metascheduler node maintains the information about the nodes in a table known as Available Node LIST (ANL). The selection of the Metascheduler and Superscheduler nodes in the cloud nodes are based on the capacity degree. Vernekar and Game [35] model is suitable for resource allocation and can add more nodes in cloud without interruption of the underlying processes.

Wang et al. [36] address the cloud providers' issue of VM allocation to PM efficiently by reducing the energy consumption. Existing approaches are applied for VM allocation without considering the migration cost. A decentralized multi-agent based VM allocation method is presented, which is based on an auction-based and negotiation-based VM allocation method. It is designed for the decision of VMs allocation to PMs and exchanges the allocated VMs for saving the energy. Proposed approach is evaluated in both static and dynamic simulations. For migration cost, the approach show the outperformed than comparison techniques in both environment, but in term of energy cost results are same to comparison technique in dynamic environ-

Journal: 10586 Article No.: 0684 MS Code: CLUS-D-16-00024.1 TYPESET DISK LE CP Disp.:2016/11/30 Pages: 45 Layout: Large

Table 4 Artificial intelligent resource allocation

Reference	Algorithm, policy or strategy	Problem addressed	Improvement/ achievements	Weakness/ limitations
Panda and Jana [29]	Alert time based resource allocation (ALT RA)	VM allocation and placement	Better performance	Considered only four nodes
Shyam and Manvi [30]	User cloudlet agent Provider resource agent Best fit approach	VM allocation	Improved performance	Need more agent for searching
An-ping and Chun-xiang [31]	Particle swarm optimization algorithm	VM allocation	Minimize energy	Compare with traditional algorithms
Radhakrishnan and Kavitha [32]	Genetic algorithm (GA)	VM allocation	To select a right system for launching VM	CPU time is not satisfied
Li and Li [34]	An iterative algorithm Resource allocation algorithm for cloud users, IaaS provider and SaaS provider (RASP)	Efficient resource allocation	Increase resource utilization	The execution the success ratio is not better than other compared algorithm
Liang et al. [33]	Ant colony optimization algorithm	Resource allocation for computing	Improve performance	Depend on the grid system
Vernekar and Game [35]	A component based resource allocation model	Resource allocation for future	Helpful in the future resource allocation	Not implemented in practically yet
Wang et al. [36]	Algorithm 1: auction-based VM allocation Algorithm 2: compute profitable swap Algorithm 3: swap contract Algorithm 4: cluster contract	VM allocation	minimize energy and migration time	In the dynamic environment results are same to the comparison technique in term of energy

ment. Hence it shows better results in a static environment. All artificial intelligent resource allocation techniques are shown in Table 4. A comparison is mentioned of the existing techniques as per the operating environment, allocation algorithms, policies and strategies for using with elementary advantages and disadvantages. Further resources and parameter used for artificial intelligent resource allocation are presented in Table 13.

4.1.2 Dynamic resource allocation

To handle the fluctuating demands of the cloud users are considered a problematic issue in cloud computing. Dynamic resource allocation techniques are used to manage and fulfill these unstable demands according to the requirement of users' need in different scenarios and workloads [37]. Also provide guarantee the QoS for avoiding the SLA violence [38].

Saraswathi et al. [39] recommend an innovative method for implementation of high priority tasks. This method ignores formation of the latest VMs for the implementation of the newly arrived task. The proposed algorithm does a high priority task in the VM that leads to the suspension of low priority task. Again, begin the suspended task if any of the VM where task is fully completed. This method has little overhead to execute all tasks comparing with creating a new VM.Moreover, to resolve the problematic issue of enormous amount of messages produced during resources allocation, a dynamically hierarchical resource allocation algorithm (DHRA) is suggested. The suggested algorithm

Journal: 10586 Article No.: 0684 MS Code: CLUS-D-16-00024.1 TYPESET DISK LE CP Disp.:2016/11/30 Pages: 45 Layout: Large

meets large-scale application service demand with increasing system reliability in cloud computing. With evaluating and testing, the DHRA's effectiveness and feasibility is shown, and communication traffic and messages are condensed [40]. Also, Wolke and Ziegler [41] evaluate the applicability of Dynamic Server Allocation Problem (a linear Program) in a deterministic environment. DSAP calculates VM allocations and live migrations on workload designs identified a priori. Simulations calculate both test bed structure of experiments and efficiency. Experimental consequences show that models are fairly precise using the live migration and demand of the servers, but deliver individual estimates the QoS roughly.

An effective dynamic resource allocation based on learning model is proposed to obtain accounting management system through quality of service standards framework (QSSF). Also, the dynamic bilateral game and resources auction strategies are also assumed to influence the interesting relationship between cloud providers and users effectively so as to allocate resources to these cloud users with a higher request [42]. To reduce the energy consumption and efficient allocation of resources with achieving optimal system efficiency by using the cloud-based learning model. Results of simulation express that the resources and energy of cloud data centers are efficiently utilized more through the reasonable distribution of resources and energy usage or storage. Further, Zhang et al. [43] suggest a framework for the dynamically allocation of the resources to see the demands of the cloud users. In the meantime, the response time of each user's request has been made assured and the service providing rate is also reserved for the users in the locality of fixed value. Similarly, IaaS performance management architecture is presented and it describes the primary application, which depends upon OpenStack by Ali et al. [44]. The fundamental structures are a group of managers that distribute resources to user requests and collaborate to complete an initiated objective of management. The manager intentions hold typical components that substantiate for a precise objective of management. Then for the two specific objectives efficiency and cost estimate a prototype implementation.

Likewise, to assign/transfer the resource of IaaS, a novel resource allocation algorithm dependent on ant colony optimization (ACO) is developed by [45], in cloud computing. Firstly, the new ACO algorithm foreseen the ability of the possibly existing resource nodes then, it examined some aspects of instance network qualities and response times to accomplish a set of optimal compute nodes. In conclusion, the jobs are dispersed to the appropriate nodes. In the same way, an innovative multi cloud resource allocation algorithm, depend upon Markov decision process (MDP), proficient of dynamic allocation the resources including the computing and storage, with the intention of increasing the estimated profits of cloud management broker (CMB). While respecting the user requirements, since minor costs for the broker suggests an improved profitable contract to the cloud user [46].

A resource allocation method can prevent the overloading problem in the system effectually while reducing the quantity of servers load. In fact, the term of skewness is used to calculate the irregular utilization of the servers familiarized by Xiao et al. [47], also develop a load prediction algorithm that is used for sensed load the upcoming resource usages of applications precisely, deprived of VMs consideration. Further, Dai et al. [48] offer an inventive dynamic resource allocation algorithm for the VM, with assistance policy. Firstly, plan the model that is used for estimated the resource allocation problem hypothetically and further presented a heuristic information based algorithm with the collaboration of all the processing nodes. Simulation based experiments conduct for the determination of evaluating and appraising novel algorithm based on collaboration policy, to estimate the algorithm's performance. The outcomes realize that the proposed algorithm could be used for fast and effectively resource allocation as well as achieving higher performance. Also, On-demand resources allocations to multiple users in various timing and distribute the workload in a dynamic environment is one of the challenging jobs of the data centers and cloud infrastructures. Therefore, the time-series model based minimum cost maximum flow (MCMF) algorithm is proposed in a study [49]. The proposed algorithm predicates multiple users' requirements in advance and outperforms the modified Bin-Packing algorithm in terms of scalability.

Various research contributions are focused on the resource allocation problems. The problems include resource optimization, simulation, distributed multi-agent systems, and SOA. These problems are solved with the assistance of multi-agent system and criteria decision analysis; prediction, graph and theoretical formulation, and service-oriented architecture [50]. Moreover, the dynamic and autonomous resource management help in assigning of resource allocation to users that assist in scalability, and flexibility. This dynamic resource management reduces the cost of resources allocation. A more related issue in clouds is to connect various clouds to distribute the workload. In a study, Wuhib et al. [51] propose an architecture for IaaS performance management and describe a preliminary execution, which is done by OpenStack. The basic building blocks are a set of controllers that allocate resources to applications and collaborate to accomplish the management objective. The controller designs comprise generic mechanisms that instantiate for specific management objectives, including the efficiency and cost estimated a prototype implementation for computing resources only. In this content, a system-orient and focus on how to achieve system-level management objectives and implement a system of collaborating controllers in a dynamic environment. On the contrary, resource allo-

Journal: 10586 Article No.: 0684 MS Code: CLUS-D-16-00024.1 TYPESET DISK LE CP Disp.:2016/11/30 Pages: 45 Layout: Large

cation on-demands among the cloud users virtually helps in reducing the processing cost and engages minimum nodes for application processing. This approach is adopted in multi-dimensional resource allocation [52]. Moreover, the two-stage algorithm follows for a multi-constraint programming problem.

On-demand resource allocation to the users from the single cloud provider is a challenging job due to high energy consumption. Besides this, to generate enough revenue and satisfy the user's needs. Zhang et al. [53] use the model predictive control (MPC) on the basis of discrete-time optimal control which helps to find the solutions. Additionally, the development of perfect information model produces on the use of strict conditions. However, the development of the model fails due to the lack of the limited knowledge which is distributed on a large scale in the cloud. Various bid proportion models and game theories are used which help in the development of information model. The Bayesian nash equilibrium allocation (BNEA) algorithm is proposed by Teng and Magoulès [54], which satisfy the heterogeneous demands of the cloud users. The proposed algorithm outperforms regarding resource allocation to the cloud users which helps in the development of perfect information system. Further, The issue of optimal resource allocation in virtual data centers (VDCs) for four illustrious management objectives are fair allocation, load balancing, service difference and energy consumption [55]. For a key organizer, the Dynamic Placement Controller, a comprehensive disperse design based on a gossip protocol that shift among management objectives. Wuhib et al. [55] test the dynamic placement of VDCs for a large cloud beneath fluctuating load and VDC churn over and done with simulation. Simulation outcomes show that this controller is highly scalable and effective for the management objective measured. Table 5 compares the techniques according to the dynamic demand of resource in cloud computing, while parameter used for dynamic resource allocation are presented in Table 14.

4.1.3 Predicted resource allocation

Sometimes predicting the users' demand for the future, influential resource requirements using automatically assigning of resources are considered substantial for resource allocation in cloud computing. For these purposes, predicted resource allocation is applied to allocate or reserve the resources for the future before they are needed [56] . It is significant and essential for effective resource allocation in IaaS cloud computing [57].

An adaptive, effective and simple framework is recommended for precise workloads prediction and saves energy in cloud centers. It is a combination of machine learning clustering and stochastic theory, which predicts VMs' demands and cloud resources related to every demand. It helps to increase the accuracy over time and neglects the requirement for frequents model that suffers the other approaches. It is also appropriate for energy aware resource management decisions in cloud data centers. Google data traces are used to calculate the efficiency of proposed framework [58]. Moreover, in cloud computing, Vasu et al. [59] focus to design, evaluate and implement a neural load predicted method for optimum resource allocation. The main objective is to minimize the energy consumption for virtualized networks. The proposed method indicates a relatively precise prediction methodology that predicts the load for future, by using the previous history of the servers. It makes sure that the demand is assigned to an optimum server, which is deserved to finish the job with less usage of energy and resource wastage. Further, Wang et al. [60] design an energy conserving resource allocation scheme with prediction (ECRASP) for VM allocation to PM in cloud computing. It predicts the trends of arriving job and related features for the future demand, which helps the system to take sufficient decisions. Numerical results show that the proposed scheme outperformed as compared to conventional algorithms for resource allocation to enhance the energy consumption.

An auction based online (AO) mechanism is designed for VM allocation and pricing issue that considers various kind of resources including the VM, CPU and Storage in cloud. The proposed online mechanism is invoked the resource availability, selection and updating status with the demand of the cloud user. It also estimates the price for the cloud users against the usage of required resource of their demand. The simulation results show that proposed mechanism achieves the faster quick response, maximum revenue and incentive compatibility, which are critical in case of online services providing in cloud [61]. In addition, Goutam and Yadav [62] present an effective algorithm for fault tolerance, which is used for advanced reservation of resources by considering the deploying of service for multiple SLA. Firstly, it checks the availability of resources locally, if resource is available or free then it is allocated to users. In case, if resource is not available or free then check the preempt-able resource and moves towards allocation, otherwise request put in waiting list as an advanced reservation. It is simulated by local simulation for fault tolerance, deployment of service and utilization of resources.

An online greedy allocation with reservation (OGAWRR) mechanism is proposed by Wu et al. [63] for IaaS private clouds. This mechanism provides the service guarantees for job completion according to the cloud users' demand. It adopts separate VM reservation method for flexible jobs and inflexible jobs. To enhance the allocation of efficiency, continuous and discontinuous reserving method are used. Finally, it is evaluated using data from RIKEN integrated cluster of clusters (RICC) and shows the better result for VM allocation and user satisfaction [64]. Similarly, Gu et

Journal: 10586 Article No.: 0684 MS Code: CLUS-D-16-00024.1 ☐ TYPESET ☐ DISK ☐ LE ☐ CP Disp.:2016/11/30 Pages: 45 Layout: Large

Table 5 Dynamic resource allocation

Reference	Algorithm, policy or strategy	Problem addressed	Improvement/achievements	Weakness/limitations
Ali et al. [44]	Cartesian Genetic Programming evolved Artificial Neural Network (CGPAN N)	Exactly predicting the client request in the Data Centers	Better Performance	Focus only the computing resources
Dai et al. [48]	Improved MapReduce model	Dynamically VM Allocation	Better performance	Depend upon the master node and Historical Information
	Heuristic information-based algorithm with cooperation Strategy			
Hadji and Zeghlache [49]	Bin-Packing algorithm	Dynamically VM allocation	Better performance and scalability	Numerical implemented
	Minimum cost maximum flow algorithm (MCMF)			
	Directed graph			
Hu et al. [45]	An allocation algorithm based on Ant colony optimization (ACO)	Computing resource allocation	Reduce response time and high performance	Compare the algorithm that bases on grid environment
Oddi et al. [46]	Novel multi-cloud resource allocation algorithm, based on a Markov decision process (MDP)	Multi-cloud resources management	High performances, better exploited and increase revenue	Not implemented in Practically
Saraswathi et al. [39]	Priority-based preemption policy			
	Procedure 1: selection of job for execution of high priority job	Resource utilization	Improve the performance	Suspend the low priority jobs
	Algorithm 1: execution of high priority job when all existing resources are allocated			
Teng and Magoulès [54]	A new Bayesian nash equilibrium allocation algorithm (BNEA)	Dynamically resource allocation	The proposed algorithm is effectively and easily implemented	Use the auction and bidding for the resource allocation
Wang and Liu [50]	Multi-agent system			
	Topology aware resource allocation (TARA)	VM allocation	Provide scalability, flexibility and reduce the size and cost of allocation	Simulation results and comparison are not shown in a study
Wang and Su [40]	Dynamically hierarchical resource-allocation algorithm	Efficient resource allocation	Enhance the performance	Compare with traditional algorithm
Wolke and Ziegler [41]	Dynamic Server allocation problem (DSAP) linear program	Dynamically VM allocation	Enhance the energy efficiency and decrease the server demand	Prediction the migration overhead is hard in simulations
Wuhib et al. [51]	Design of the two controllers that implement the placement scheduler	Monitor resource utilization and dynamically resource allocations	Increase the efficiency of completing a management objective	The cost of effectiveness increases may become prohibitive in a highly dynamic system with the level of VM churn
	Initial placement controller			
	Dynamic placement controller			

Journal: 10586 Article No.: 0684 MS Code: CLUS-D-16-00024.1 ☐ TYPESET ☐ DISK ☐ LE ☐ CP Disp.:2016/11/30 Pages: 45 Layout: Large

Table 5 continued

Reference	Algorithm, policy or strategy	Problem addressed	Improvement/achievements	Weakness/limitations
Wuhib et al. [55]	A gossip protocol	Allocation of computing and network resources jointly in a large IaaS cloud	Improved the load balancing, fair allocation, energy efficiency, and service differentiation	The resource allocation system is accessible to at least 100,000 machines and VDCs
Xiao et al. [47]	Load prediction algorithm The skewness algorithm	Overload avoidance and green computing	Accomplishes overload avoidance and green computing	Do not compare with other algorithms
Xie and Liu [42]	Quality of service standards framework Resource auctions strategy Dynamic bilateral game strategy	Dynamic effective resource allocation	Enhance the performance	Compare with non-familiar algorithms in cloud computing
Yin et al. [52]	Multi-dimensional resource allocation scheme (MDRA)	Dynamically resource allocation and job scheduling,	Enhance the resource utilization and minimize the costs	Only focus on the economical point of view
Zhang et al. [43]	A dynamic resource allocation framework Queue algorithm Priority-balance (PB)	Satisfy the QoS requirements including the service rate and response time	The proposed framework adapted to the dynamic cloud and shows better performance	Not implemented in practically
Zhang et al. [53]	Model predictive control (MPC)	Dynamically resource allocation	Improving the revenue, energy cost, and response time	Compare with simple strategy

Journal: **10586** Article No.: **0684** MS Code: **CLUS-D-16-00024.1** ☐ TYPESET ☐ DISK ☐ LE ☐ CP Disp.:**2016/11/30** Pages: **45** Layout: **Large**

Table 6 Predicted resource allocation

Reference	Algorithm, policy or strategy	Problem addressed	Improvement/ achievements	Weakness/ limitations
Dabbagh et al. [58]	Energy aware resource provisioning framework	VM allocation	Save energy for the data centers	Focus only on cloud providers
Goutam and Yadav [62]	Algorithm 1: forming a task list based on priorities Algorithm 2: priority based scheduling algorithm Algorithm 3: advanced reservations and preemption based cloud min–min algorithm Algorithm 4: algorithm for fault tolerance	Preemptive resource allocation	Improve the deployment of service fault tolerance and utilization of resources	Allocation is based on high priority
Gu et al. [65]	Latest reservation online (LRO) mechanism	Online VM allocation	Improve the performance	Mechanism focus only one VM per time unit
Mashayekhy et al. [61]	Auction based online (AO) Mechanism	Online VM allocation	Improve the performance	Mechanism does not forecast future demand
Vasu et al. [59]	Fast up slow down (FUSD) algorithm Back propagation	Load prediction and energy consumption	Maximize the utilization	Only focus on server and do not consider CPU, Storage and VMs
Wang et al. [60]	Energy conserving resource allocation scheme with prediction (ECRASP)	Energy consumption	Improve the performance	Not implemented in practically yet
Wu et al. [63]	Online greedy allocation with reservation (OGAWR) mechanism	Online VM allocation	Improve the performance	Do not cloud computing system for the simulation

al. [65] use the latest-reservation online (LRO) Mechanism for enhancing the social welfare in resource allocation in IaaS private clouds. Various predicted resource allocation techniques are compared according to various metrics and primary differences are listed in Table 6 and used resources and parameters in these techniques are presented in Table 15.

4.2 Parametric based resource allocation

Parametric based resource allocation is classified into further six diverse groups containing the cost aware resource allocation, efficiency aware resource allocation, load balancing aware resource allocation, power aware resource allocation, QoS aware resource allocation and utilization aware resource allocation.

In briefly, cost aware resource allocation focuses on the overall cost, which includes the cloud providers' profit and revenue, users' expenses and prices of resource. Efficiency aware resource allocation attentions on the efficiency to enhance the performance by minimizing the execution and response time, maximize the bandwidth or speed and priority. Load balancing aware resource allocation emphases on workload to the distribution of resources to the several users in various data centers. Power aware resource allocation concentrates on the green computing to reduce the energy and heat consumption in the data centers. QoS ware resource allocation deliberates on the improvement of services for the cloud user in term of availability, fault tolerance, reliability, recovery time, throughput and SLA violation. Utilization aware resource allocation emphases on utilization to increase the usage of cloud resources, professionally. The details of the categorization listed above are as follow.

4.2.1 Cost aware resource allocation

Cost aware resource allocation is a crucial issue in cloud computing, it is responsible for the services in economical way according to the definition of cloud [66]. Cloud providers are responsible for distributing the services to fulfill user's need in efficient way. In return, they want the growth of profit and revenue with extreme resource utilization, while cloud users' want to receive the services within minimum amount to pay with high performance [4]. In this case, efficient resource allocation mechanisms or techniques play a significant role in cloud computing.

A demand based preferential resource allocation method is proposed in [67], that proposes for resource allocation a market driven auction mechanism based on their capacities.

Journal: 10586 Article No.: 0684 MS Code: CLUS-D-16-00024.1 TYPESET DISK LE CP Disp.:2016/11/30 Pages: 45 Layout: Large

In term of payment and it implements a payment strategy based on the service preferences of the buyer. There are two steps in resource allocation technique, first, a driven payment process which ensures that a lesser amount is paid by the winner than the bid value provided that the bidding reflects the best paying capacity. Second, a market driven auction process which guarantees profit and reliability to the service provider. Additionally, a comparison between the famous offline VCG auction mechanism and the proposed allocation technique is presented, and results predict a performance advantage in revenues to the service provider, payments of the cloud users besides ensuring an optimum resources usage. A new technique position balanced parallel particle swarm optimization (PBPPSO) algorithm is proposed for allocation of resources in IaaS cloud [68]. The main objective of PBPPSO is to find out the optimization of resources for the group of jobs with minimum makespan and cost.

In a study, Nezarat and Dastghaibifard [69] propose a method based on an auction, which applies game theory mechanism to determines the auction winner and holding a repetitive game with inadequate information. At the last point of the game theory approach is the Nash equilibrium. Where user no longer need to change the bid for the required resource, in the final stage the user bid satisfied the auctioneer's utility function in game theory approach. In the end, simulation results conclude that this method comes together with shorter response time, lowest SLA violations and the higher resource utilization to the provider. Moreover, the combinatorial double auction resource allocation (CDARA) model is recommended by Samimi et al. [70] for the both user and cloud provider's perception inefficient and intensive from. The proposed model is confirmed through simulation and estimated based on two evaluation standards: the involved economic efficiency and the incentive compatibility. The experimental results obviously demonstrate that the proposed method is cost effective, efficient and intensive for the both user and cloud provider while producing higher revenues for providers and reduce the cost for users.

The resource swarm algorithm employs to adjust the cost and price of the resources in cloud computing. The swarm algorithm uses dual models in which they adjust the price of the resources that are: initial price model (IPM) and resource swarm algorithm price adjustment model (RSAPAM) suggested by Li et al. [71]. The IPM presumes the initial prices of the cloud resources. This information with on-demand changes to the RSAPAM and this algorithm computes and adjusts the required resource price according to the users. Therefore, these resources with on-demand will be handed over to each user in the most appropriate time. Similarly, Chintapalli [72] proposes an algorithm for assigning resources to the cloud user's demand with lower cost and a specified constraints budget and deadline. At this point, the study considers several cloud providers for assigning these cloud user's requirements. In the end, based on the results and proposed algorithm implementation, it is concluded that it will run on linear time. Furthermore in [73], resource allocation for cloud customers are assigned according to their needs, and on-demand where all types of details are kept hidden from the customers through virtualization. Moreover, it has been noticed that services are similar regarding functionalities and interfaces, but this is not justified financially to pay more for on-demand service and provides the regular services. However, the study shows that resources are allocated in cloud to the users by their needs and biding.

In the research, Kumar et al. [74] develop a VMs allocation algorithm to the user's application with the help of real-time task. The VMs allocation is expressed as a resource optimization problem and solved this problem with the help of a polynomial-time heuristic. In the end, the cost attained is associated by the proposed heuristic with the optimal solution, and an earliest deadline first (EDF-greedy) strategy, complex analysis of parameter of the concerned problem. Furthermore, Yi et al. [75] consider the budget optimization allocation for IaaS model in distributed grid or clouds of joint resources including the network, processor, and storage from the consumer's viewpoint. And recommend a Best Fit heuristic algorithm with several job scheduling policies and with a new resource model, design a mixed integer linear programming (MILP) formulation. To reduce the expenses for every single user to attain sufficient resources to implement their submitted jobs while supporting the grid or cloud provider to receive several job requests from the cloud users while considering the basic objectives.

Casalicchio et al. [76] explain that to enhance the revenue, cloud provider subject towards capacity, availability of SLA and VM migration constraints. However, to solve this, NOPT Near Optimal also known as a NP-hard problem as it argues about the results along with a relevant allocation strategy. However, while the allocation of combined resource allocation framework for network cloud is based on the formulation of optimal network cloud mapping problem as an assorted integer programming. Nevertheless, it identifies the objective concerning the cost effectiveness of resource mapping procedure as enduring the user requests regarding QoS aware virtual resources. Additionally, a mechanism needs to design for exposes the accurate values for random task arrival and maximize the cost. In a study, Gu et al. [77] anticipate a mechanism for online truthful VMs allocation. It is compared with offline mechanism through the simulation and show the more efficient competitive ratio. Also, mechanism is used to analysis the performance and capacity. Table 7 compares the previously mentioned techniques that are applied for cost aware resource allocation while resources and parameters used for cost aware resource allocation are shown in Table 16.

Journal: 10586 Article No.: 0684 MS Code: CLUS-D-16-00024.1 TYPESET DISK LE CP Disp.:2016/11/30 Pages: 45 Layout: Large

Table 7 Cost aware resource allocation

Reference	Algorithm, policy or strategy	Problem addressed	Improvement/achievements	Weakness/limitations
Casalicchio et al. [76]	Near optimal (NOPT) Algorithm 1: near optimal (NOPT) hill climbing local search method Algorithm 2: neighbours function	NP-hard problem and optimal allocation of VMs requested	Improvement in average revenue and maintained the availability	Compared with only with best fit strategy
Chintapalli [72]	Cost and time optimization algorithm	Deadline and budget aware resource allocation	Improve the performance	Do compare with existing algorithms
Gu et al. [77]	Preemptive VMs allocation online mechanism	Online VM allocation	Improve the performance	Do compare with existing algorithms
Kumar et al. [74]	EDF (earliest deadline first)-greedy scheme	Allocation of VMs to applications with real-time tasks	Allocate resources efficiently	Focus only the cost
Kumar and Saxena [67]	Demand-based preferential resource allocation technique	Resource allocation based on payment	Better performance	Allocation is based on the priority
Li et al. [71]	Initial price model (IPM) Resource swarm algorithm price adjustment model	Resource pricing in cloud bank model	To spread the best realistic price with time	Only proposed the model and focused on the economical allocation of the resources
Mohana [68]	Position balanced parallel particle swarm optimization (PB-PPSO)	Optimal resource allocation	Improve performance	The resources are allocated to learning the rules for new user request
Nezarat and Dastghaibifard [69]	Game theory Algorithm 1: user i bidding algorithm Algorithm 2: auctioneer allocation algorithm	Multi-user allocation	Enhance performance and increase profit rate	Do not compare with other algorithms
Samimi et al. [70]	The combinatorial double auction resource allocation (CDARA)	Market-based resource allocation	Economic efficiency & Incentive compatibility	Only focus on economical point of view
Teng and Magoules [73]	Game theory	Resource pricing	Support financially	Policy is based on prediction
Yi et al. [75]	Mixed integer linear programming (MILP) Best-fit heuristic algorithm Resource co-allocation Algorithm 1: best-fit algorithm Algorithm 2: bandwidth resource allocation	Over-provisioning of resources and costly energy consumption	Reduce cost and better performance	Focus only on cloud user

Journal: 10586 Article No.: 0684 MS Code: CLUS-D-16-00024.1 ☐ TYPESET ☐ DISK ☐ LE ☐ CP Disp.:2016/11/30 Pages: 45 Layout: Large

4.2.2 Efficiency aware resource allocation

Efficiency aware resource allocation directly affects the performance, which specifies the satisfaction of the cloud users in cloud computing. It helps enhance and improve the bandwidth or speed, execution time, priority and response time for allocation of resources to the cloud users in more proficient economically and efficiently way [78].

Mashayekhy et al. [79] identify the issues of online allocation and scheduling of virtual machines in the presence of numerous categories of resources in cloud, then design an offline and online incentive-compatible procedures. The recommended offline procedure is perfectly assumed that the info on all the upcoming demands is identified a priori. On the other hand, proposed online procedures make no presumption for future request of VMs. Planned online procedures are raised quickly as the user places a demand. Otherwise, particular assigned resources are free and become accessible. The procedures not only dynamically allocate and schedule the resources but also conclude the user's expenses such that the incentive-compatibility is assured. Further, Nejad et al. [80] repeat using the approximation proportion of the recommended greedy approach and examine their results by executing in-depth experiments. The outcomes show that the suggested greedy approach conclude near-optimal results with minimizing the execution time while allocating and scheduling computing resources to match the user's request, and creating high expenses for the cloud providers.

Cloud providers are controlled and allocated all computational resources in a flexible manner according to the cloud users' demand. Hence, still there is difficulty to face the optimal resource allocation in cloud computing. Pradhan et al. [81] propose a modified round robin algorithm to fulfill the cloud users demands by decreasing the response time. Time quantum is considered to be basic elementary of RR algorithm, whereas the difference of dynamic and fixed time quantum is also found to further enhancement of resource allocation in cloud computing. In addition, User's demands for realtime dynamic alteration are very hard to realize precisely. The meta-heuristic ant colony algorithm is considered to resolve these types of problematic issues, but the algorithm has slow convergence speed and parameter selection problems. To resolve this problematic issue, Yang et al. [82] propose an optimize ant colony algorithm based on particle swarm algorithm for resolving resources allocation problem in IaaS cloud. Hence, Xu and Yu [83] investigate the issue of resource allocation in cloud computing. Several forms of resources like CPU, network, and storage on VM level are considered. A recommended allocation FUGA algorithm not only supports the optimal resource allocation for the cloud users but also helps in the efficient utilization of resources for each physical server. The issue of resource allocation is demonstrated as an extensive finite game with accurate info and the FUGA algorithm consequences in a Nash equilibrium decision. Table 8 comprehensively compares various efficiency aware resource allocation techniques, while resources and parameters used in these techniques are presented in Table 17.

4.2.3 Load balancing aware resource allocation

Balancing a load of data centers or VMs is a feasible procedure with the help of allocation of resources through sharing loads in a systematic way to attain high performance and utilization of resources [84,85]. Optimal resource allocation must confirm that resources are certainly accessible on users' demand and competently operate under condition of high/low load [86].

Allocation of virtual machines, utilization of the cloud resources and appropriate load balancing policies show a critical part in IaaS cloud computing. VM allocation algorithm assigns the VMs to the data center hosts whereas the cloudlet allocation algorithm performs as a load balancing procedure. It defines a way to bind cloudlets to VM, so each cloudlet has less execution time and high speed to complete the job. A fair allocation of cloudlets between the VMs is provided by proposed algorithms. Both algorithms are designed, simulated and analyzed in CloudSim simulator [87]. In the research, Bhise and Mali [88] discourse a problem of resource provisioning in IaaS clouds on user sides. Specifically, the user adopts the virtual machine for the implementation and quantity of virtual machines desire to satisfy the QoS requirements (e.g. deadlines) before performing a workload. The workload constitutes a cloudlet or a group of independent cloudlets. The similar workloads have different price and performances regarding the allocation and scheduling approach with concerns to the two pricing choices. The aim is to minimize the inclusive cost of virtual machine provisioning, with reservation and on demand possibility. Amazon EC2 selects a pricing option to make it extra convincing. Experimental enhancement is verified with Boinc Project workload and proposed method improves the cost performance.

Ray and Sarkar [89] present a new algorithm for distribution of the jobs to control workload balancing. Allocation is completed depending upon the requirement presented by the cloud users, and at the end, a service level agreement is made between cloud providers and cloud users. In the discussed algorithm requisite or features of the job is presented by the cloud users to the cloud provider. Cloud providers store the request in the source in XML design. The ultimate selection of the resource depends on the resource use matrix, execution time of jobs and expenses. Further, Villegas et al. [90] propose an extensive and empirical performance of cost analysis for resource allocation and scheduling policies for IaaS Cloud. Firstly, this study presents the taxonomy of mutual types of policies, based on the information type used

Journal: 10586 Article No.: 0684 MS Code: CLUS-D-16-00024.1 TYPESET DISK LE CP Disp.:2016/11/30 Pages: 45 Layout: Large

Table 8 Efficiency aware resource allocation

Reference	Algorithm, policy or strategy	Problems addressed	Improvement/ achievements	weakness/ limitations
Mashayekhy et al. [79]	Algorithm 1: VCG-VMPAC mechanism (C) Algorithm 2: OVMPAC-X mechanisms (event, A,P) Algorithm 3: OVMPAC-X-ALLOC(t,Qt, Ct) Algorithm 4: OVMPAC-X-PAY(t,Qt,At, Ct)	NP-hard problem	Fast resource allocation	Focus on only online mechanism and do not compare with existing techniques
Nejad et al. [80]	Algorithm 1: VCG-VMPAC mechanism Algorithm 2: OVMPAC-X mechanisms Algorithm 3: OVMPAC-X-ALLOC allocation algorithm Algorithm 4: PAY(payment function)	VM allocation	Decrease the execution time and fulfill the user demand and generating revenue	Focus on only online mechanism and do not compare with existing technique
Pradhan et al. [81]	Modified round Robin algorithm	Optimal resource allocation	Improve the performance	Focus only on cloud user
Xu and Yu [83]	Game theory FUGA algorithm	Multi-resource allocation	Improve the performance of fair allocation	Compare with traditional algorithms
Yang et al. [82]	Ant colony optimization algorithm based on particle swarm algorithm	Efficient resources allocation	Efficient task allocation	Not implemented

in the decision process and categorized into eight provisioning and four allocation policies. Furthermore, these policies are examined for cost and performance through using Amazon EC2 as a cloud. Moreover, Zhang et al. [91] present an approach that is a combination of mutually resource prediction and resources allocation in cloud, which is used for the virtual machines allocation to the cloud users. The resources are allocated by the statistic based load balance (SLB) while the virtual machine is used for load balancing. SLB contains two portions, one deals with the online statistical analysis of virtual machine's performance and predict the demand for the resources and another one is used as algorithm for load balancing by selecting the accurate host in the resource pool based on the prediction and the past load data of hosts.

Effective resource allocation algorithm may enhance the bandwidth, load balancing, delay and reliability for cloud computing. Liu et al. [92] propose the multi-QoS load balance resource allocation method (MQLB-RAM) strategy based on resource allocation. It combines the users' demands and providers' services while allocates the VMs to PMs and binds the task by specific sensor correspondingly. It also compares weight of each index value to fulfill the demand with resources, to succeed the good load balancing, resource utilization and reduce the cost. Simulation results show the proposed algorithm outperformed than the Round Robbin (RR) and throtted load balance (TLB) algorithms. Table 9 compares the miscellaneous techniques for resource allocation to balance the workload of cloud, while resources and parameters used for load balancing are presented in Table 18.

4.2.4 Power aware resource allocation

Power aware resource allocation mechanisms are succeed in dealing with the problems arising due to the heat generation and energy consumption in data centers. It is essential for the cloud providers and data centers to generate less heat, reducing the energy consumption and saving the cost [93]. Due to the rapid growth of data center, increasing the amount of servers, huge load, highly demands and loss or wastage of idle power are major causes of energy and heat ineffectiveness [94]. Green computing is anticipated for optimal resource allocation and utilization, by reducing heat and energy consumption in data centers [95,96].

In a study, Ali et al. [97] focus on VM allocation problem considering a bin packing in IaaS cloud computing. The main intention is to reduce the consumption of energy in the data centers. An energy efficient (EE) algorithm is proposed to place VMs demands on most energy efficient PMs first. For

Journal: 10586 Article No.: 0684 MS Code: CLUS-D-16-00024.1 TYPESET DISK LE CP Disp.:2016/11/30 Pages: 45 Layout: Large

Table 9 Load balancing aware resource allocation

Reference	Algorithm, policy or strategy	Problem addressed	Improvement/ achievements	Weakness/ limitations
Bhise and Mali [88]	Heuristic algorithm Pseudo code 1: provisioning/scheduling(W(t[n], ct[n], d[n]) Pseudo code 2: PlanSubscription (VM, utilization)	Optimal resource allocation	Improve the cost performance when increase deadline of a workload	Do not compare with existing algorithm
Liu et al. [92]	Multi-QoS load balance resource allocation method (MQLB-RAM)	Optimal resource allocation	Improve the performance with minimizing cost	Do not show proper results for balancing the load
Parikh et al. [87]	Task binding policy Hungarian algorithm	VM allocation	Better performance	Do not compare with existing algorithm
Ray and Sarkar [89]	Novel load balancing algorithm	Load balancing and Job scheduling to utilize the resources	Reduce the workload	Fail-safe state of the resource is not considered
Villegas et al. [90]	Provisioning polices Startup On-demand, single VM (OD-S) OD-geometric (OD-G) OD-ExecTime OD-ExecAvg OD-ExecKN OD-Wait OD-2Q Allocation polices First-come, first-served (FCFS) FCFS-nowait (FCFS-NW) Shortest-Job First (SJF) FCFS-MultiQueue	To manage workload	Better performance and cost	None of them work combined to show be performance
Zhang et al. [91]	Statistic based load balance approach, (SLB)	To manage workload	Load balancing in time	Focus on the time and other resources are not mentioned

this purpose, dynamic voltage frequency scheduling (DVFS), power aware (PA) and non-power aware (NPA) techniques are adopted in proposed algorithm. In simulation environment, EE algorithm achieves more high energy efficiency than the comparison algorithms. However, Energy-efficient based policies and algorithm are proposed in [98]. Before developing the principles for cloud computing architecture, the designer must describe energy efficient management scheme that helps to design an algorithm and allocates resource to the users on the basis of demand. Further, Dashti and Rahmani [99] use PSO algorithm to dynamically VMs migration for improving resource allocation and gain more benefit in the data center. To assure a less response time and QoS (SLA) by presenting an innovative heuristic method for dynamic resource re-allocation, with balancing the cloud provider's overloaded. Similarly, associated cloud provider's under load and power, to get more energy efficiency and power saving.

Conversely, a multi-purpose ant colony system algorithm is suggested for the VM placement problem by Gao et al. [100]. The objective is to gain an efficiently appropriate solution that concurrently minimizes overall power consumption and resource wastage. Particular instances verify the proposed algorithm from the literature. After comparing the performance with an existing multi-purpose grouping genetic algorithm, the outcomes show that the suggested algorithm can compete professionally with other favorable algorithms. In addition, Kansal and Chana [101] present a model for resource utilization to organize the resources of cloud and increase their usage proficiently. The objective is to decline energy consumption of clouds without affecting user application performance. Based on ABC meta-heuristic technique

Journal: **10586** Article No.: **0684** MS Code: **CLUS-D-16-00024.1** ☐ TYPESET ☐ DISK ☐ LE ☐ CP Disp.:**2016/11/30** Pages: **45** Layout: **Large**

a resource utilization technique is proposed to find the fittest job-node pair, it tries to enhance the energy efficiency through the finest use of resources. The consumption of energy is reduced with the conflict among memory and processor utilizations. Two types of workloads are considered including CPU and memory intensive. In order to avoid contention and conflict among the resources, these workloads are carefully associated. Therefore, this model helps in increasing the satisfaction of cloud users and directly contributes to the green computing by minimizing energy consumption and carbon emission also. However, Yanggratoke et al. [102] propose a protocol in order to minimize the energy consumption of the consumer computers and servers known as GRMP-Q protocol. They focus on migrating most of the load towards servers and allocate the CPU slots to the consumers. Their findings show that they do not change the structure and size of the system and supports 100,000 servers regarding resource allocations.

The energy consumption of servers is increasing due to a linear way of resource utilization. In this case, share the load or load balancing techniques are not effective and help to reduce the energy consumption. Jha and Gupta [103] propose a policy to minimize the energy consumption and expenses of the cloud providers. Proposed policy is performed better to reduce the energy consumption and maximum utilization or resources via testing in CloudSim simulator. Similarly, Gupta and Ghrera [104] propose a power and failure aware resource allocation (PFARA) algorithm to minimize the energy consumption and expenses of the cloud providers. Also, proposed algorithm is outperformed to minimize the energy consumption and enhance the resource utilization within simulation experiments. Furthermore, Dynamic VM placement emphasis on the mapping of VMs to PMs, with maximum utilization and no disturbance occur at the time of execution. Pavithra and Ranjana [105] present a weighted first-cum-first-served (WFCFS) algorithm for developing an energy efficient resource provisioning framework with dynamic VM placement. The simulations results based on CloudSim show that the better performance by reducing the energy consumption, cost and execution time as compared to static environment.

Cloud data center heterogeneous and homogenous architecture require different usage of energy to utilize the workload. Green cloud data centers and QoS assurance are considered to be main issues in cloud computing. Peng et al. [106] recommend an evolutionary energy efficient virtual machine allocation (EEE-VMA) method for minimizing the energy consumption of the data centers. To fulfill the VM allocation request, GA algorithm is used for saving energy, cost and utility in the method. The approach shows the better results in simulation and Openstack for reducing energy, cost and workload. Hence, Singh and Kaushal [107] focus on improvement of the VM allocation procedure in IaaS cloud computing by reducing the energy consumption. An algorithm is proposed to reduce the energy consumption, maximum utilization of resources to PMs, maintain proper schedule of VM and compare the difference of energy consumption before and after the VM allocation. The simulation results show the decreasing amount of VM migration by affecting the energy consumption in the data centers. Table 10 comprehensively compares previous various techniques that are applied in energy aware resource allocation, while resources and parameters used in power aware resource allocation are presented in Table 19.

4.2.5 QoS aware resource allocation

QoS aware resource allocation plays an important role in cloud computing. It implies to distribution of resources according to the cloud user's demand regarding to the QoS, which emphases on the availability, fault tolerance, recovery time, reliability, throughput and SLA for the both cloud providers and users [108]. At the time of resource allocation, the QoS must considers to avoid the increasing the failure rates, non-availability of resources, poor resource utilization and SLA violence [109].

Resource management module (ReMM), is a self-managed and dynamic module that is proposed for efficient resource utilization, QoS and workload balancing, computing resources and quantity of those resources are assigned to the cloud users with dissimilar workloads and are precised during the performance analysis. In this way, it is possible to calculate the guidance of configurations of fluctuating demand of users. The simulation results show that the proposed module is able to fulfill the altering demand of resources by confirming the QoS with comparative variations in the cost [110]. Additionally, Li et al. [111] advise a layered progressive resource allocation algorithm based on the multiple knapsack problem called LPMKP. The LPMKP algorithm considers the VM requirements of different tenants and their relationship. It introduces the allocation goal of minimizing the sum of the VM's network diameters of all tenants. A reduction in resource fragmentation in cloud data centers is achieved by decreasing the differences in QoS among tenants, and improving the overall QoS across all tenants for cloud data centers. The experimental results show that LPMKP efficiently deals with the VM resource allocation problem for multi-tenant in cloud data centers.

A novel QoS aware VMs consolidation approach is presented by Horri et al. [112] that adopts a method based on resource utilization by using distant past of virtual machines. Using resource utilization history of VMs minimize the energy consumption and SLAV as follows: the energy consumption reduces because with a high probability, the peak load of VMs do not occur at the same time and reducing the number of times a host reaches it's their peak (100 %

Journal: 10586 Article No.: 0684 MS Code: CLUS-D-16-00024.1 TYPESET DISK LE CP Disp.:2016/11/30 Pages: 45 Layout: Large

Table 10 Power aware resource allocation

Reference	Algorithm, policy or strategy	Problem addressed	Improvement/ achievement	Weakness/ limitations
Ali et al. [97]	Energy efficient (EE) algorithm	VM allocation	Improve the performance	Compared with only basic algorithms
Beloglazov et al. [98]	Algorithm 1: modified best fit decreasing (MBFD) Algorithm 2: minimization of migrations (MM)	Reducing power consumption of a data center	Reduce the energy consumption in data center	Reduce the intention on QoS and SLA violation
Dashti and Rahmani [99]	Particle swarm optimization	VM placement	Improve the performance	Compare with traditional algorithm
Gao et al. [100]	Multi-objective ant colony system algorithm	VM placement	Better performance	Focus only on cloud providers parameters
Gupta and Ghrera [104]	Power and failures aware resource allocation (PFARA) algorithm	Energy consumption	Improve the performance	Do not compare with existing polices and also not focus on the failure request (reliability)
Jha and Gupta [103]	Power and load aware VM allocation policy	VM allocation	Improve the performance	Do not compare with existing polices and also not focus on the load balancing
Kansal and Chana [101]	Artificial bee colony	Energy consumption	Minimize execution time and energy efficiency	Workload of nodes does not consider
Pavithra and Ranjana [105]	Weighted FCFS	VM placement	Improve the resource utilization	Do not compare with existing algorithms
Peng et al. [106]	Evolutionary energy efficient virtual machine allocation (EEE-VMA)	VM allocation	Minimize the energy, cost and utility	Do not compare with existing algorithms
Singh and Kaushal [107]	Power stability algorithm (PSA)	VM allocation	Improve the performance	Compare with only based algorithm whereas do not compare with existing algorithms
Yanggratoke et al. [102]	GRMP, a generic gossip protocol for resource management GRMP-Q, under overload the protocol gives a fair allocation of CPU resources to clients	Reducing power consumption of a data center	Reduce the energy consumption in data center	Do not compare with other protocols

utilization) reduce SLAV. The main focus is to familiarize an efficient SLA aware algorithm, to avoid SLA violation as much as possible and dramatically condense the operation cost. The suggested algorithms reflect the trade-off between performance and energy consumption.

On demand resources allocation to the end users in cloud is obtained with proposed algorithm known as selective algorithm [113]. The proposed algorithm uses the concept of min–min and max–min algorithms in order to allocate the resources to users on the scheduling basis which is considered in the conventional scheduling algorithm. The selection of the Min–min algorithm or max–min algorithm is based on the heuristic techniques that consume fewer resources of the machines. The machine based resources sharing can be spaced or time matter. The proposed algorithm uses CloudSim simulator and allocation of resources is performed

Journal: 10586 Article No.: 0684 MS Code: CLUS-D-16-00024.1 ☐ TYPESET ☐ DISK ☐ LE ☐ CP Disp.:2016/11/30 Pages: 45 Layout: Large

on First Come First Serve (FCFS). However, the finding of the proposed algorithm is quite satisfactory and reduced the cost of machine resources. Likewise, Lee et al. [114] propose a competent algorithm that goes along with a strategy best-fit for virtual machine allocation to the physical machines. To realize the VM migration, a performance analysis scheme is designed for each host node in observation of processing and storage specification. Proposed resource allocation system provides for allocating virtual machine on the optimal node to supply the service considering user needs and to use effectively the high and low performance of node considering each performance. Experiment results show that the proposed framework enhance the resource utilization without exchanging the allocation time, for supporting user's demand at a time. Also, Li [115] emphases on the rental problem of a virtual machine for the long/short term. A learning algorithm based on statistical learning techniques and dynamic virtual machine rental algorithm is anticipated for resource requirement. These algorithms reduced the operational cost even though stabilizing determined quality of service (QoS) requirement.

On demand resources allocation and task scheduling investigate in this study [116], which is the core module of cloud computing. The proposed scheduling algorithm uses the vector of resource and task matching which differentiates between on-demand and ordinary requirements of the users. The allocation of resources to the users is based on the availability of the QoS service. During allocation of resources to the users, it also investigates and uses the batch and online modes for load balancing. The outcomes of the scheduling algorithm are satisfactory to allocate resources on run-time to the users. In a study, Kang and Wang [117] familiarize an innovative auction approach, to allocate the resources to the suitable cloud facilities in cloud computing. Although, facilities are capable of finding their appropriate services and resources, to discover the high worth of resources with a high level of service easily. This approach structures the perception of fitness and the re-design bargaining function and procedure to calculate the last trade price. The overall market competence is completely enhanced in this way. Experimental results certify the algorithm and express that efficient resource allocation is easily achieved by lacking the fitness function.

The Scheduling and leasing based on a dynamic scheduling algorithm is proposed in [118], in order to permit new leases on-demand. The proposed algorithm determines multiple slots and uses swapping and backfilling to accommodate the leases which are deadline sensitive. The swapping and preemption techniques are used to reschedule the slots leases when they require for deadline sensitive and on-demand services. If both techniques fail to reschedule the slots for leases then the proposed solution uses backfilling which can assign the idle slots (resources) to the leases. The objective of the study [119], is to suggest a resource allocation structural design for cloud computing that offers the dimension of value indicators recognize amongst the key performance indicators (KPI) explain with cloud services measurement initiative consortium (CSMIC). Proposed structural design recommends various resource allocation policies including both reactive and predictive. In this architecture, according to the SLA the provision decisions are taken. In conclusion, the initial investigational outcomes show that the suggested structural design improves quality in cloud. Besides, a quality of service constrained resource allocation issue is addressed by Wei et al. [120], where cloud users expect to clarify sophisticated computing issue through requesting the resources utilization across a cloud based network. A price of each network node is based on the quantity of processing. A performance based QoS and computation concentrated cloudlets in a cloud environment are discussed. Wei et al. [120] focus on the parallel tasks allocation problematic issues on distinct networks connected to the Internet.

Nguyen et al. [121] precede transition diagram that clarifies all possible situations in a data center. With the help of this diagram, the probability of rejection and the response time based on the probabilities of every step of description is formulated. Also, the effective number of slots for reservation of the migration process is decided. As a result, the cloud providers can increase revenue by reducing energy consumption and costs used for the redundant slots. Besides, Papagianni et al. [122] explain a methodology regarding effective and efficient mapping of resource request on to a substrate interconnection of numerous computing resources, as it follows a heuristic methodology while taking into account a problem.

The Machine learning method is defined to form a distribution method for resource mapping and prediction. With the simulation, the resources are distributed to the fresh cloud users by learning the instructions of the preparation method. Similarly, a resource allocation and adaptive job scheduling (RAAJS) algorithm is designed for cloud computing formed with the help of grid computing [123]. In this case, the grid is in accordance to the resources as the circulation of resources are both locally and worldwide as the sharing of resources are among cloud computing and grid computing environment. Moreover, Kumar et al. [123] suggest the use of new weight matrics (WM) to carry out various task and selection of resources. Thus, WM re-arranged task and it enhances the competence of the proposed algorithm. However, the algorithm is calculated in accordance with various metrics, and its competence shows the reduction in job completion time and the various attempts required to get accessibility of specific service as it enhances the percentage of resource allocation. Also, a cooperative game theoretic framework is used to solve network resource allocation problem in view of both efficiency and fairness. Fair Allocation policy with both

Journal: 10586 Article No.: 0684 MS Code: CLUS-D-16-00024.1 TYPESET DISK LE CP Disp.:2016/11/30 Pages: 45 Layout: Large

online and offline algorithm is designed to achieve fairness in terms of guarantee the bandwidth and share it according to the weights in the network. Experimental results show that proposed policy provides flexible reliability and balances the load for better utilization of network in the data centers [124]. Table 11 shows the comparison of QoS aware resource allocation techniques, while resources and parameters used for these techniques are presented in Table 20.

4.2.6 Utilization aware resource allocation

Generally, efficient utilization of resources directly influences the success of cloud computing. Although, the cloud providers always have limited amount resources in their data centers and efforts to organize them in extreme utilization through optimal resource allocation [125]. To achieve the several requirements of the cloud users with maximum utilization of all resources efficiently, when several cloud users demand various resources at the same time is challenging issue [126].

Lin et al. [127] focus on the cloud providers to efficiently utilize resources by fixing VM arrangement to the cloud users for IaaS by historical empiric service data traces. The foremost influences are to describe a problem of VM allocation and define the appropriate beta distribution of the CPU component by the use of empirical data collection to resolve the issue. With the help of simulations, the CPU module is useful for IaaS administrators to correct usage of VM and proficiently notice the resources with reservation parameters and SLA. To avoid underutilization of resources, Pillai and Rao [128] expose the usage of the uncertainty standards of game theory to model association development between machines in cloud. The benefit of the proposed method avoids the complexities of integer programming by explaining the optimization issue of coalition formation. Beside, resource allocation mechanism aims to achieve less resource wastage, minor task allocation time, and higher user's satisfaction. Firstly defines the problem that is to the placement of particular VMs on the presented physical machines, especially for the advanced reservation request model. Then suggest an algorithm depend upon integer linear programming (ILP) to resolve certain communal situations of the issue. Lastly, the algorithm is executed with the help of Haizea simulator, and the simulation values are associated with the Haizea greedy algorithm and several heuristics techniques [129].

In addition, Srinivasa et al. [130] suggest a utilization maximization (UM) model for resource allocation issues in IaaS cloud. Initially, by using Cloudsim simulator to simulate various entities included for resource allocation in IaaS cloud and the interactions and procedures concerning included entities. Also, the resource algorithms for the broker and cloud users are recommended. Further, Tyagi and Manoria [131] identify the data security issue and enhance the resource utilization for a storage system in cloud computing. Cuckoo search algorithm is applied for the selection of server and user authentication. It helps to improve the reliability and efficient utilization of resources. Proposed algorithm is compared with GA and SLPSO and showed the outperformed performance by using the Matlab. Table 12 compares the techniques according to utilization aware resource allocation and further detail for resources and parameters used in these techniques are presented in Table 21.

5 Analysis of resources and parameters used in current studies

In this section, resources and parameters used in assessing the existing research works are given in Table 13, 14 15 16 17 18 19 20, and 21 below. The tables show that the IaaS cloud resources [132] used by the existing researchers are CPU, Network, Node, Storage and VM.

- **CPU**: In cloud computing, cloud providers deliver shared resources and data for computing and processing on demand of the cloud users. A CPU also known as a virtual processor, is a physical central processing unit that is allocated to a VMs. It depends on the cloud users demand, either demand required single, dual or multiple CPU cores.
- **Network**: includes the hardware and software resources (Routers, Switches, LAN cards, Wireless routers, Cables, Firewall and Network security applications) of the entire network that enables network connectivity, communication, operations and management of an initiative network. In simple words, it provides the communication path and services between users, processes, applications, services and external networks/the Internet.
- **Node**: is a connection point, either a redistribution point or an end point for data transmissions in general. In cloud computing, Nodes are known as servers or end nodes. It may sometimes actually be a virtual node for avoiding heterogeneity of the nodes but usually, it is considered to be a physical server or host machines.
- **Storage**: is a cloud resource in which data and applications are stored on remote servers retrieved from cloud. It is maintained, operated and managed by cloud providers on storage servers that are built on virtualization techniques.
- **VM**: is becoming more common with the evolution of virtualization technology. It is frequently generated to execute certain tasks by software competition ways or hardware virtualization techniques that are different than tasks are executed in a host environment.

Journal: 10586 Article No.: 0684 MS Code: CLUS-D-16-00024.1 TYPESET DISK LE CP Disp.:2016/11/30 Pages: 45 Layout: Large

Table 11 QoS aware resource allocation

Reference	Algorithm, policy or strategy	Problem addressed	Improvement/achievements	Weakness/limitations
Batista et al. [110]	Resource management module (ReMM)	Performance based resource allocation	Improve the performance	Dynamic demands have not significant impact on response
Guo et al. [124]	Fair allocation policy	Fair network bandwidth allocation	Improve the performance while reducing the overall load of network	Focus only on cloud providers
	Algorithm 1: update weight and base bandwidth			
	Algorithm 2: bandwidth allocation on server			
	Algorithm 3: online bandwidth allocation			
Horri et al. [112]	SLA-aware algorithm	To avoid SLA violation or reducing energy cost, live migration of VM	Reduce the number of VM migration, SLAV and total transmitted data	The energy consumption reduces because with a high probability, peak load of VMs not possible together
	Algorithm 1: finding new placement of VMs			
Kang and Wang [117]	Cloud resource allocating algorithm via fitness-enabled auction (CRAA/FA)	QoS constrained resource allocation	Improve the overall market efficiency	They study of algorithm in term of economic efficiency and system performance
Katyal and Mishra [113]	Selective algorithm is based on min–min and max–min algorithms	Resource provision (allocation and scheduling)	Increase throughput through reducing makespan	Compared only with the FCFS algorithm
Kumar et al. [123]	Meta scheduler	To decrease the resources consumption to avoid resource starvation	Resource availability reduces completion time and enhances the QoS for cloud users	Depend on the grid computing
	Resource allocation and adaptive job scheduling (RAAJS) algorithm			
	Weight matrics			
Lee et al. [114]	Performance analysis based resource allocation scheme	Efficient allocation of VM	Increase the resource utilization	Compared with basic algorithms
	Virtual machine scheduling algorithm			
Li [115]	Algorithm 1: resource requirements learning algorithm	VM placement	Enhance the performance	Do not compare with other algorithms

Journal: 10586 Article No.: 0684 MS Code: CLUS-D-16-00024.1 ☐ TYPESET ☐ DISK ☐ LE ☐ CP Disp.:2016/11/30 Pages: 45 Layout: Large

Table 11 continued

Reference	Algorithm, policy or strategy	Problem addressed	Improvement/achievements	Weakness/limitations
	Algorithm 2: dynamic server hosting algorithm			
	Markov modulated poisson process (MMPP)			
Li et al. [111]	Layered progressive resource allocation algorithm based on the multiple Knapsack problem (LP-MKP)	Multi-tenant VM allocation	Enhance the performance	Fix the range of leaf nodes
Nathani et al. [118]	Swapping of consecutive leases	VM allocation and placement	Reduce the request rejection rate and satisfying resource allocation strategies	Increase the overall overhead of the system
	Backfilling of leases			
Nguyen et al. [121]	Transaction diagram	VM migration	Improved performance	Numerically tested and not compared with other algorithms
	Algorithm 1: determine K			
	Algorithm 2: determine _N, K,optimal cost			
Pan et al. [116]	Management system of task scheduling and resource allocation of cloud computing	Assigning user jobs to the suitable resources	Reduce the number of task and completion time, with increasing the utilization	Do not perform the experiment or simulation for testing purpose
Papagianni et al. [122]	Node mapping phase	Optimal NCM problem (network cloud mapping)	Structured, flexible, and fair performance evaluation	Focus on fixed and wired networks and infrastructures
	Link mapping phase			
	NCM approach			
Sagbo and Houngue [119]	Resource allocation architecture for quality in cloud computing	Efficient resource allocation	Minimize SLAs violation and improve QoS	Focus only on cloud provider
Wei et al. [120]	Game theory	QoS constrained resource allocation	Divide the multiple cooperative subtasks in many cloud based computing and data store services	Do not compare with other algorithms
	A binary integer programming			
	Nash equilibrium			
	Algorithm 1: SPELR minimization			
	Algorithm 2: GELR minimization			
	Algorithm 3: evolutionary optimization			

Journal: 10586 Article No.: 0684 MS Code: CLUS-D-16-00024.1 ☐ TYPESET ☐ DISK ☐ LE ☐ CP Disp.:2016/11/30 Pages: 45 Layout: Large

Table 12 Utilization aware resource allocation

Reference	Algorithm, policy or strategy	Problem addressed	Improvement/achievements	Weakness/limitations
Lin et al. [127]	Beta distribution model Service data traces	VM allocation	Improved performance	Cause overloading
Pillai and Rao [128]	Game theory Algorithm 1: open coalition formation algorithm Algorithm 2: coalition dissolving algorithm Algorithm 3: task allocation algorithm	Underutilization of resources	Avoid the complexity of integer programming & Enhance the performance	Each task has only types of request
Rezvani et al. [129]	Integer linear programming	VM allocation and migration	Improve the performance	Compare with traditional algorithms
Srinivasa et al. [130]	Min–max game approach, Cloud resource allocation games (CRAGs) Utilization maximization [UM] model	VM allocation and migration	Qualitative and economic improvement	Consider only a static scenario
Tyagi and Manoria [131]	Cuckoo search algorithm	Utilization of servers	Improve the reliability and efficiency	Data security and storage is not considered in the experiments

Journal: 10586 Article No.: 0684 MS Code: CLUS-D-16-00024.1 ☐ TYPESET ☐ DISK ☐ LE ☐ CP Disp.:2016/11/30 Pages: 45 Layout: Large

Table 13 Matrix of resources and parameters for artificial intelligent resource allocation in IaaS cloud

Reference	Resources						Parameters				
	CPU	Network	Node	Storage	VM	Task/cloudlet	Availability	Bandwidth/speed	Cost	Energy	Execution time
An-ping and Chun-xiang [31]	✓		✓	✓	✓					✓	
Li and Li [34]	✓			✓				✓	✓	✓	
Liang et al. [33]		✓						✓			
Panda and Jana [29]			✓								
Radhakrishnan and Kavitha [32]	✓		✓		✓						
Shyam and Manvi [30]					✓	✓			✓		
Vernekar and Game [35]	✓		✓		✓						
Wang et al. [36]		✓			✓					✓	✓

Reference	Parameters										
	Memory	Performance	Priority	Reliability	Response time	SLA	Time	Throughput	Temperature	Utilization	Workload
An-ping and Chun-xiang [31]										✓	
Li and Li [34]							✓				
Liang et al. [33]	✓	✓			✓						
Panda and Jana [29]							✓				
Radhakrishnan and Kavitha [32]				✓			✓				
Shyam and Manvi [30]							✓				
Vernekar and Game [35]					✓	✓		✓			
Wang et al. [36]											

Journal: 10586 Article No.: 0684 MS Code: CLUS-D-16-00024.1 ☐ TYPESET ☐ DISK ☐ LE ☐ CP Disp.:2016/11/30 Pages: 45 Layout: Large

Table 14 Matrix of Resources and Parameters for Dynamic Resource Allocation in IaaS Cloud

Reference	Resources						Parameters				
	CPU	Network	Node	Storage	VM	Task/cloudlet	Availability	Bandwidth/speed	Cost	Energy	Execution time
Ali et al. [44]			✓	✓							
Dai et al. [48]			✓			✓					
Hu et al. [45]			✓					✓			
Hadji and Zeghlache [49]									✓		
Oddi et al. [46]									✓		
Saraswathi et al. [39]			✓			✓					
Teng and MagoulFs [54]	✓					✓			✓		
Wang and Liu [50]					✓						
Wang and Su [40]			✓			✓					✓
Wolke and Ziegler [41]			✓		✓				✓	✓	
Wuhib et al. [51]					✓			✓	✓	✓	
Wuhib et al. [55]	✓	✓			✓					✓	
Xie and Liu [42]		✓		✓		✓				✓	
Xiao et al. [47]	✓				✓						
Yin et al. [52]					✓	✓					
Zhang et al. [43]		✓							✓		
Zhang et al. [53]	✓	✓			✓	✓			✓		

Reference	Parameters										
	Memory	Performance	Priority	Reliability	Response time	SLA	Time	Throughput	Temperature	Utilization	Workload
Ali et al. [44]							✓				
Dai et al. [48]							✓			✓	
Hu et al. [45]		✓		✓	✓						
Hadji and Zeghlache [49]							✓				
Oddi et al. [46]							✓				
Saraswathi et al. [39]							✓				
Teng and MagoulFs [54]							✓				
Wang and Liu [50]						✓					
Wang and Su [40]											

Journal: 10586 Article No.: 0684 MS Code: CLUS-D-16-00024.1 ☐ TYPESET ☐ DISK ☐ LE ☐ CP Disp.:2016/11/30 Pages: 45 Layout: Large

Table 14 continued

Reference	Parameters										
	Memory	Performance	Priority	Reliability	Response time	SLA	Time	Throughput	Temperature	Utilization	Workload
Wolke and Ziegler [41]											
Wuhib et al. [51]											✓
Wuhib et al. [55]						✓				✓	✓
Xie and Liu [42]										✓	
Xiao et al. [47]							✓			✓	✓
Yin et al. [52]							✓			✓	✓
Zhang et al. [43]		✓			✓						
Zhang et al. [53]							✓				✓

Journal: 10586 Article No.: 0684 MS Code: CLUS-D-16-00024.1 TYPESET DISK LE CP Disp.:2016/11/30 Pages: 45 Layout: Large

Table 15 Matrix of resources and parameters for predicted resource allocation in IaaS cloud

Reference	Resources						Parameters				
	CPU	Network	Node	Storage	VM	Task/cloudlet	Availability	Bandwidth/speed	Cost	Energy	Execution time
Dabbagh et al. [58]	✓			✓	✓					✓	✓
Goutam and Yadav [62]				✓	✓						✓
Gu et al. [65]					✓	✓			✓		
Mashayekhy et al. [61]				✓	✓				✓		✓
Vasu et al. [59]		✓								✓	
Wang et al. [60]					✓	✓				✓	
Wu et al. [63]					✓	✓			✓		✓
Wu et al. [64]					✓	✓			✓		✓

Reference	Parameters										
	Memory	Performance	Priority	Reliability	Response time	SLA	Time	Throughput	Temperature	Utilization	Workload
Dabbagh et al. [58]										✓	
Goutam and Yadav [62]				✓	✓						
Gu et al. [65]					✓						
Mashayekhy et al. [61]	✓									✓	✓
Vasu et al. [59]				✓						✓	
Wang et al. [60]					✓						✓
Wu et al. [63]					✓						
Wu et al. [64]					✓						

Journal: 10586 Article No.: 0684 MS Code: CLUS-D-16-00024.1 ☐ TYPESET ☐ DISK ☐ LE ☐ CP Disp.:2016/11/30 Pages: 45 Layout: Large

Table 16 Matrix of resources and parameters for cost aware resource allocation in IaaS cloud

Reference	Resources						Parameters				
	CPU	Network	Node	Storage	VM	Task/cloudlet	Availability	Bandwidth/speed	Cost	Energy	Execution time
Casalicchio et al. [77]					✓		✓		✓		
Chintapalli [73]	✓	✓	✓						✓		
Gu et al. [78]					✓	✓			✓		
Kumar and Saxena [68]									✓		
Kumar et al. [124]	✓		✓	✓					✓		
Kumar et al. [75]					✓	✓		✓	✓		
Li et al. [72]									✓		
Mohana [69]					✓				✓		
Nezarat and Dastghaibifard [70]					✓				✓		
Samimi et al. [71]	✓			✓	✓			✓	✓		
Teng and Magoules [74]	✓					✓			✓		
Yi et al. [76]	✓		✓		✓	✓		✓	✓		

Reference	Parameters										
	Memory	Performance	Priority	Reliability	Response time	SLA	Time	Throughput	Temperature	Utilization	Workload
Casalicchio et al. [77]						✓	✓				
Chintapalli [73]							✓				
Gu et al. [78]		✓									
Kumar and Saxena [68]			✓							✓	
Kumar et al. [124]		✓		✓				✓			
Kumar et al. [75]		✓					✓				✓
Li et al. [72]											
Mohana [69]					✓						
Nezarat and Dastghaibifard [70]										✓	✓
Samimi et al. [71]	✓										
Teng and Magoules [74]							✓				
Yi et al. [76]										✓	

Journal: 10586 Article No.: 0684 MS Code: CLUS-D-16-00024.1 ☐ TYPESET ☐ DISK ☐ LE ☐ CP Disp.:2016/11/30 Pages: 45 Layout: Large

- **Task/cloudlet**: In cloud computing, cloudlet is a mini cloud set to serve a specific purpose in a given environment on the demand of the cloud users. However, in the simulation tools, it is known as a task to perform certain operation.

It also shows that a number of parameters have been presented for the purpose of comparison and these includes the availability, bandwidth/ speed, cost, energy, execution time, memory, performance, QoS, priority, reliability, response time, SLA, temperature, throughput, time, utilization and workload.

- **Availability**: is committable, operable, or usable of resources, depend upon the cloud users' request to implement its designated or required operation. It is the combination of resource's accessibility, maintainability, reliability, securability and serviceability in cloud computing [133,134].

$$Availability = \sum_{resource^i} \left(\frac{MTBM}{MTBM + MTTR}\right) \tag{1}$$

where $MTBM$ represents the Mean Time Between Maintain and $MTTR$ represents the Mean Time to Repair of $resource^i$

- **Bandwidth/speed**: is the maximum data transfer rate of a network. It measures how much data or resources can be sent over a specific connection in a given amount of time in cloud computing [135].

$$BW = \sum_{resource^i} \left(\frac{Size}{Capicity}\right) \tag{2}$$

- **Cost**: is an amount that has to be paid against the usage of resource in cloud computing. It is profit and revenue for the cloud providers and expense for the cloud users besides the utilization of resources in cloud computing [5,136].

$$Cost_{Total} = \sum_{resource^i} (C_i * T_i) \tag{3}$$

where C_i represents the cost of resource i per unit time and T_i represents the time of utilization of $resource$

- **Energy**: is a strength or vitality required for execution of cloudlets or tasks for certain resources of the cloud users demand in cloud computing. Simply, it is a form an electricity to run the PMs in data centers. The energy consumption of given resource i at a time T with placement F [98,137]

$$Energy_{Total} = \sum_{resource^i} \int_{Str_{Time}}^{Fnh_{Time}} E_i(F, T). \tag{4}$$

where E_i represents the energy is consumed by the resource i from its starting time to finishing time of utilization.

- **Execution Time**: is a time in which cloudlets or tasks are running or computing as the demand of the cloud users. It is also known as completion time, which is required for the specific cloudlets or tasks to complete the job [138].

$$Exe_{Time} = task_i(Fnh_{Time} - Str_{Time}) \tag{5}$$

where Fnh_{Time} denotes the finishing time and Str_{Time} represen t starting time of $task_i$

- **Memory**: is a process in which the cloudlet or tasks are encoded, retrieved or stored as the requirement of the cloud users in cloud computing. Therefore, all the data is loaded from the cloud storage into the memory to match the processing speed before it is executed by cloud processor [139].
- **Performance**: is an amount of cloudlet or task accomplished on the demand of the cloud users [140].

$$Performance = task_i\left(\frac{I * CPI}{R}\right). \tag{6}$$

where, I denotes the instruction and CPI re presents the computing performance improvement, which depend of many factors like memory, execution te etc. and R shows the reciprocal of time.

- **Priority**: is a cloudlet or task that has more importance than other or has right to execute or proceed before others. It is necessary due to the cloud user pay more than for its urgent requirement or beneficial for cloud provider in cloud computing [78].

$$Priortity = \sum_{task^i} \quad (Exe_{Time} + Capicity * Number\ of\ Requests) \tag{7}$$

- **Reliability**: is the ability of cloudlet or task to execute its required function within specific time successfully. It provides the assurance of completion and avoid or reduce the failure rate in cloud computing [139,141].

$$Reliability = \frac{\sum_{task^i}(Exe_{Time})}{Total_{Time}} \tag{8}$$

- **Response time**: is a time, takes to respond to the request for service or when cloudlet or task starts the execution and comes out from the waiting queue [142].

$$Res_{Time} = \sum_{task^i}(Sub_{Time} + Str_{Time}). \tag{9}$$

where Sub_{Time} denotes the submission time and Str_{Time} represents the starting time of the $task_i$

Journal: 10586 Article No.: 0684 MS Code: CLUS-D-16-00024.1 ☐ TYPESET ☐ DISK ☐ LE ☐ CP Disp.:2016/11/30 Pages: 45 Layout: Large

- **SLA**: is an agreement between the cloud providers and cloud users against the utilization of resources. Every cloud provider wants to deliver their best services to fulfill the requirement of the cloud user and avoid the SLA violence [143].

$$\text{SLA} = \frac{Number\ of\ executed\ tasks\ sucessfully\ or\ \sum_{task^i}(Exe_{Time})}{Number\ of\ servicies\ or\ resources\ offered} * 100. \tag{10}$$

- **Temperature**: is a degree or strength of heat present or generate in cloud computing environment. In this environment, it refers to the heat generation in data center when cloudlets or task are executing on the PMs [144,145].

$$Specific_{Heat} = resource_i\left(\frac{heat}{m * \triangle T}\right) \tag{11}$$

where m denotes the mass and $\triangle T$ represents the time of the $resource_i$

- **Throughput**: is a total amount of cloudlets or tasks that are executed successfully within given time period in cloud computing [13].

$$Throughput = \sum_{task^i}(Exe_{Time}) \tag{12}$$

- **Time**: is a plan or schedule, when tasks or resources should be executed or allocated to the cloud users. It is a measured or measurable period during which an action, process or condition exists or continues in cloud computing.

$$Time = \sum_{task^i}\left(\frac{Distance}{Speed}\right) \tag{13}$$

- **Utilization**: is the total amount of resources actually consumed in the data centers. The objective is to utilize the resources effectively is to maximize the cloud providers' revenue and profit with the cloud users' satisfaction [4,5].

$$Utilization = \frac{\sum_{resource^i}(Exe_{Time})}{Makespen\ or\ max_{task^i}(Exe_{Time})} \tag{14}$$

- **Workload**: is the amount of processing to be done or handled within given time period. In simple, it is the ability to handle or process work in cloud computing. Degree of imbalance is used for calculating the load of work in data centers [146].

$$Degree\ of\ Imbalance = \frac{max_{task}(Exe_{Time}) - min_{task}(Exe_{Time})}{Avg_{task^i}(Exe_{Time})} \tag{15}$$

Artificial intelligence is a branch of cloud computing that intentions to generate intelligent techniques for IaaS resource allocation. It has become an essential part of the modern technology. Resource allocation associated with artificial intelligent is highly technical and specialized. The resources and parameters used for artificial intelligent resource allocation in existing techniques are shown in Table 13.

The dynamic resource allocation studies focus on various fluctuating on-demand resource allocations to the cloud users. The resources and parameters used for dynamic resource allocation in current techniques are mentioned in Table 14.

Prediction considers various metrics and behaviour of methods during the allocation of resources. Therefore, resource allocation must satisfy all the the requirements of the cloud users to meet the SLA. These metrics and prediction can be used for optimum resource allocation for IaaS cloud computing. The resources and parameters used for predicted resource allocation in previous techniques are stated in Table 15.

In cloud system, cloud providers' main target is to achieve high profit and revenue with maximum utilization of all cloud resources. For this motive, resources are assigned to the cloud users in that way so that it reduces the energy consumption, workload, SLA violations and enhance resource utilization with users' satisfaction. While cloud users always want to get these cloud services and resources with high performance within minimum expenses. The resources and parameters used for cost aware resource allocation in previous researches are shown in Table 16.

Every cloud provider and user want high performance with the extreme utilization of cloud resources in cloud computing. It is realized by reducing the execution and response time while enhancing the bandwidth or speed. The resources and parameters used for efficiency aware resource allocation in recent techniques are presented in Table 17.

Overloaded and unbalanced resources are the source of failure of a system and SLA violence. For these purposes, load balancing techniques are implemented for resource allocation in cloud computing. The resources and parameters used for load balancing aware resource allocation in previous techniques are displayed in Table 18.

The growth of cloud data centers is increased day by day due to the rising demand and popularity of cloud computing. Heat generation and energy consumption are a major problems in data centers so that these issues can be controlled by power aware resource allocation in cloud computing. The resources and parameters used for power aware resource allocation in existing techniques are shown in Table 19.

QoS is considered to be the main feature of cloud computing to deliver cloud resources and services. It can be

Journal: 10586 Article No.: 0684 MS Code: CLUS-D-16-00024.1 ☐ TYPESET ☐ DISK ☐ LE ☐ CP Disp.:2016/11/30 Pages: 45 Layout: Large

Table 17 Matrix of resources and parameters for efficiency aware resource allocation in IaaS cloud

Reference	Resources						Parameters				
	CPU	Network	Node	Storage	VM	Task/cloudlet	Availability	Bandwidth/speed	Cost	Energy	Execution time
Mashayekhy et al. [79]					✓	✓					✓
Nejad et al. [80]	✓		✓	✓	✓				✓		✓
Pradhan et al. [81]	✓										
Xu and Yu [83]	✓		✓	✓	✓						
Yang et al. [82]			✓			✓					✓

Reference	Parameters										
	Memory	Performance	Priority	Reliability	Response time	SLA	Time	Throughput	Temperature	Utilization	Workload
Mashayekhy et al. [79]											✓
Nejad et al. [80]										✓	✓
Pradhan et al. [81]					✓						
Xu and Yu [83]							✓			✓	
Yang et al. [82]											

Journal: 10586 Article No.: 0684 MS Code: CLUS-D-16-00024.1 TYPESET DISK LE CP Disp.:2016/11/30 Pages: 45 Layout: Large

Table 18 Matrix of resources and parameters for load balancing aware resource allocation in IaaS cloud

Reference	Resources					Parameters					
	CPU	Network	Node	Storage	VM	Task/cloudlet	Availability	Bandwidth/speed	Cost	Energy	Execution time
Bhise and Mali [88]					✓				✓		
Liu et al. [92]					✓				✓		✓
Parikh et al. [87]	✓			✓	✓	✓		✓			
Ray and Sarkar [89]	✓			✓					✓		
Villegas et al. [90]									✓		
Zhang et al. [91]	✓	✓			✓						

Reference	Parameters										
	Memory	Performance	Priority	Reliability	Response time	SLA	Time	Throughput	Temperature	Utilization	Workload
Bhise and Mali [88]							✓				✓
Liu et al. [92]					✓						
Parikh et al. [87]										✓	✓
Ray and Sarkar [89]	✓						✓				✓
Villegas et al. [90]		✓									✓
Zhang et al. [91]							✓			✓	✓

Journal: **10586** Article No.: **0684** MS Code: **CLUS-D-16-00024.1** ☐ TYPESET ☐ DISK ☐ LE ☐ CP Disp.:**2016/11/30** Pages: **45** Layout: **Large**

Table 19 Matrix of Resources and Parameters for Power aware Resource Allocation in IaaS Cloud

Reference	Resources						Parameters			
	CPU	Network	Node	Storage	VM	Task/cloudlet	Availability	Bandwidth/speed	Cost	Energy
Ali et al. [98]					✓					✓
Beloglazov et al. [99]		✓			✓					✓
Dashti and Rahmani [100]					✓	✓				✓
Gao et al. [101]					✓					✓
Gupta and Ghrera [105]		✓			✓					✓
Jha and Gupta [104]					✓					✓
Kansal and Chana [102]			✓		✓					✓
Pavithra and Ranjana [106]					✓				✓	✓
Peng et al. [107]	✓	✓			✓				✓	✓
Singh and Kaushal [108]					✓	✓				✓
Yanggratoke et al. [103]	✓									✓

Reference	Parameters											
	Execution time	Memory	Performance	Priority	Reliability	Response time	SLA	Time	Throughput	Temperature	Utilization	Workload
Ali et al. [98]							✓			✓		
Beloglazov et al. [99]								✓			✓	
Dashti and Rahmani [100]					✓			✓				
Gao et al. [101]					✓						✓	
Gupta and Ghrera [105]												
Jha and Gupta [104]												
Kansal and Chana [102]											✓	
Pavithra and Ranjana [106]	✓										✓	
Peng et al. [107]												✓
Singh and Kaushal [108]												
Yanggratoke et al. [103]												

Journal: 10586 Article No.: 0684 MS Code: CLUS-D-16-00024.1 TYPESET DISK LE CP Disp.:2016/11/30 Pages: 45 Layout: Large

Table 20 Matrix of resources and parameters for QoS aware resource allocation in IaaS cloud

Reference	Resources						Parameters															
	CPU	Network	Node	Storage	VM	Task/ cloudlet	Availability	Bandwidth speed	Cost	Energy	Execution time	Memory	Performance	Priority	Reliability	Response time	SLA	Time	Throughput	Temperature	Utilization	Work load
Batista et al. [110]	✓			✓	✓		✓		✓		✓		✓			✓						✓
Guo et al. [124]		✓					✓	✓			✓			✓					✓		✓	
Horri et al. [112]	✓				✓					✓							✓				✓	
Katyal and Mishra [113]					✓	✓						✓						✓	✓			
Kang and Wang [117]								✓	✓				✓				✓					
Lee et al. [114]	✓			✓	✓		✓						✓			✓					✓	
Li [115]					✓				✓				✓				✓					
Li et al. [111]			✓		✓								✓									
Nathani et al. [118]			✓															✓			✓	
Nguyen et al. [121]			✓												✓	✓						
Pan et al. [116]	✓		✓	✓			✓		✓												✓	✓
Papagianni et al. [122]	✓	✓		✓					✓						✓						✓	
Sagbo and Houngue [119]	✓				✓												✓				✓	✓
Wei et al. [120]						✓			✓									✓				

Journal: 10586 Article No.: 0684 MS Code: CLUS-D-16-00024.1 ☐ TYPESET ☐ DISK ☐ LE ☐ CP Disp.:2016/11/30 Pages: 45 Layout: Large

Table 21 Matrix of resources and parameters for utilization aware resource allocation in IaaS cloud

Reference	Resources						Parameters			
	CPU	Network	Node	Storage	VM	Task/cloudlet	Availability	Bandwidth/speed	Cost	Energy
Lin et al. [127]	✓				✓					
Pillai and Rao [128]		✓		✓	✓	✓	✓			
Rezvani et al. [129]	✓				✓					
Srinivasa et al. [130]						✓	✓		✓	✓
Tyagi and Manoria [131]	✓			✓	✓					

Reference	Parameters											
	Execution time	Memory	Performance	Priority	Reliability	Response time	SLA	Time	Throughput	Temperature	Utilization	Workload
Lin et al. [127]							✓				✓	
Pillai and Rao [128]					✓			✓			✓	
Rezvani et al. [129]								✓			✓	✓
Srinivasa et al. [130]	✓					✓					✓	
Tyagi and Manoria [131]		✓									✓	

Journal: 10586 Article No.: 0684 MS Code: CLUS-D-16-00024.1 ☐ TYPESET ☐ DISK ☐ LE ☐ CP Disp.:2016/11/30 Pages: 45 Layout: Large

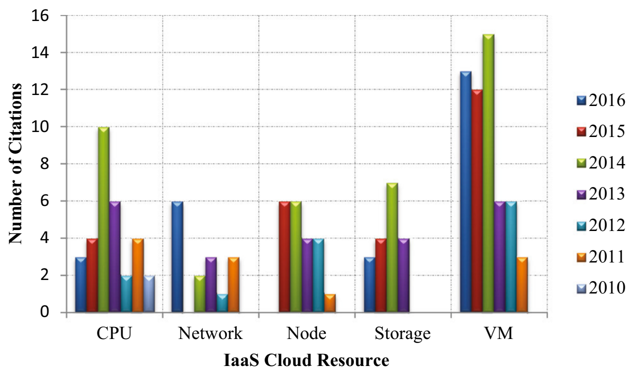


Fig. 5 Analysis of IaaS cloud resources from 2010 to 2016

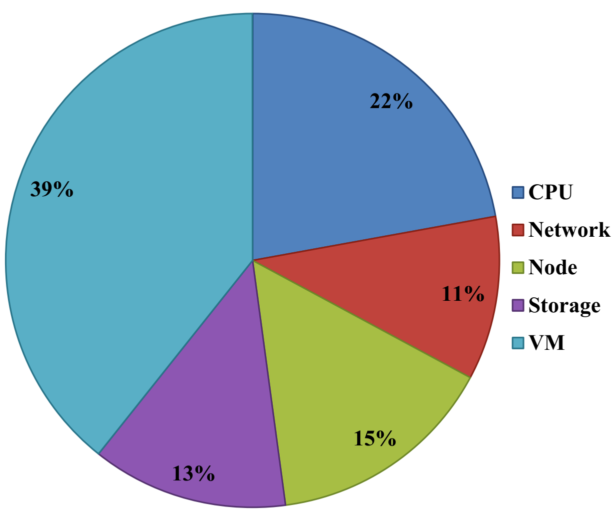


Fig. 6 Analysis of IaaS cloud resources from 2010 to 2016

achieved by the assurance of availability, reliability, reducing the failure rate and SLA violence in cloud. The resources and parameters used for QoS aware resource allocation in previous techniques are presented in Table 20.

Optimal resource utilization directly affects the cloud providers' profit and revenue. For this purpose, utilization aware resource allocation techniques are played a signification role to fair distribution of resources, reducing energy consumption and resources usage. The resources and parameters used for utilization aware resource allocation in existing techniques are displayed in Table 21.

Figures 5 and 6 explain that majority of the scholars are concentrated on the VMs and computation resources in the research area of cloud computing for resource allocation in IaaS, while some of them are focused on the other resources. As we understand that storage and network resources are the fundamental necessities of cloud computing that fully depend on these resources.

After reviewing of Figs. 7 and 8, it is observed that cost, energy, time and utilization are thought to be the most beneficial parameters described by scholars in the field of resource allocation. Although, the bandwidth or speed, execution time, performance, reliability, response time, SLA and workload are emphasized by some scholars. However, there is a strong necessity for concentrating on the parameters. Meanwhile, in IaaS cloud computing, the availability, memory, priority throughput and temperature are thought to be the primary parameters for resource allocation but a little number of scholars are applied these parameters in their studies. In fact, cloud is a business model, where every cloud provider wishes a reduction in the expenditure (energy, temperature, storage, etc.) for enhancing the revenues with maximum usage of resources competently. However, cloud users always look for higher performances of the services with least cost and time. Therefore, cost, energy, reliability, utilization and workload are thought to be most essential parameters in the field of cloud computing research for resource allocation. But there is need to be more focus on the temperature, priority and throughput in the future research in cloud computing for maintaining the heat generation in data centers, fair allocation and enhancing the resource utilization in cloud computing.

Cloud computing, green computing [147], and big data [148] are of critical concern. The aim of green computing is cleaning the cloud environment with a focus on the energy, temperature and storage. However big data attention is on the data management. The achievement and attraction behind cloud computing is due to the services provided by cloud. Because of having a countable number of resources, it is of eminence importance for providers to manage and allocate the cloud resources in time to the cloud users as per the dynamic nature of their demands. In this review, several resource allocation strategies, policies, and algorithms in IaaS cloud computing environments have been analyzed, with their important parameters.

6 Future works

The main issues commonly associated with IaaS in cloud computing are resource management, network infrastructure management, virtualization and multi-tenancy, data management (Big Data), energy, heat and storage management (Green Computing), application programming interfaces (APIs) and interoperability, etc. Resource management related problems include resource provisioning, resource allocation, resource adaptation, resource mapping, resource modeling, resource discovery, resource brokering and resource scheduling. Figure 9 is a bubble graph that chronicles the future directions in resource allocation as pointed out in previous research articles by other authors.

- **Green computing**: is going to be limitless with the rapid growth of business in the future. It is a procedure to use computing resources environmentally and

Journal: **10586** Article No.: **0684** MS Code: **CLUS-D-16-00024.1** ☐ TYPESET ☐ DISK ☐ LE ☐ CP Disp.:**2016/11/30** Pages: **45** Layout: **Large**

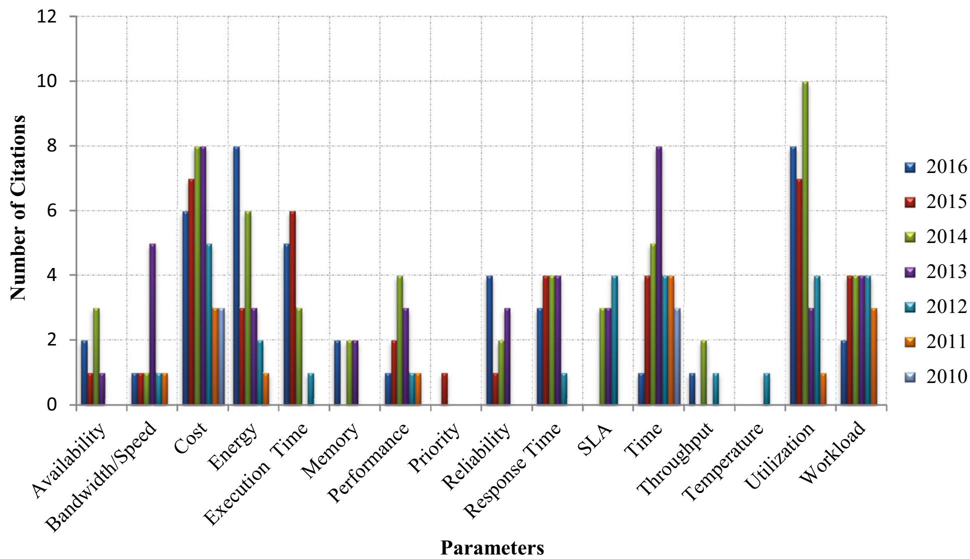


Fig. 7 Analysis of resource allocation parameters from 2010 to 2015

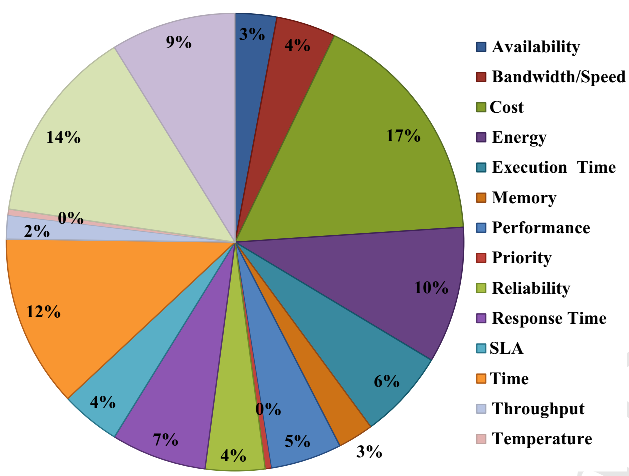


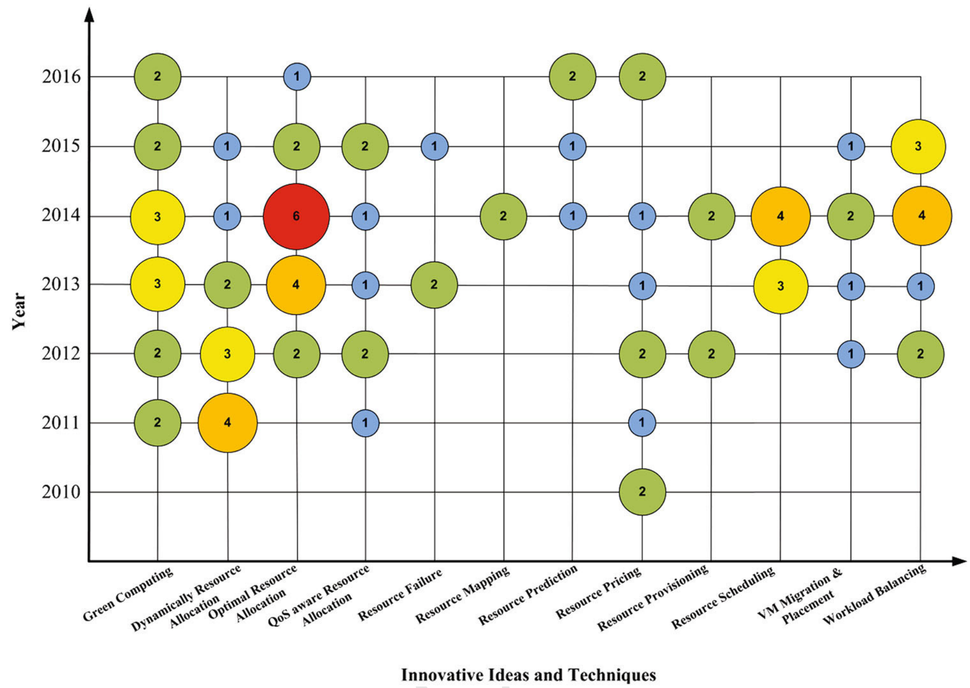
Fig. 8 Analysis of resource allocation parameters from 2010 to 2015

user friendly while maintain overall computing performance. To reduce the use of hazardous materials, minimize energy consumption, less heat generation and resource wastage are problematic issues in computing [12,36,43,44,55,62,67,75,91,98,102,105,112].

- **Dynamic resource allocation**: is applied for increase or decrease allocation of resources according to the fluctuating demands of the cloud users. It allows cloud users to scale up and down resources based on their needs [10,24,49,50,53,76,87,91,102,130].
- **Optimal resource allocation**: Due to the constantly increasing demands of the cloud users for services or resources. It is very challenging to distribute the resources precisely to the cloud users' demands in order to fulfill their requirements and also gives the guarantee of QoS to the cloud users regarding to the SLA by the cloud providers [5,6,19,21,22,35,51,55,76,79,83,112].
- **QoS aware Resource Allocation:** is required for high performance, availability of resources, handle of conflicts of resource demands, fault-tolerance and reliability [53,68,80,118,119,130].
- **Resource failure**: various types of resource failures are directly influenced by the failure or success of cloud services in cloud computing. These are including overflow, underflow timeout, resource missing, computing failure, software failure, storage failure, database failure, hardware failure, and network failure [24,89,92].
- **Resource mapping**: is a need of automating discovery, allocation processes and make the monitoring process to be more vigorous. It is able to allocate and re-allocate resources according to demand or the current status of resource utilization in the data centers of cloud. In this way, self-management of resources and self-adaption of configurations can be possible conferring to diverse situations [6,19].
- **Resource prediction**: is required for a given set of workloads running on a VMs or PMs predict the utilization of resources (such as CPU, storage, etc.) that are required for enhancing the performance. It also required for SLA to estimate the cost of resource utilization, to determine that which resource is suitable to meet SLA and to assessment the resources requirement for given workload in cloud computing [58,61,63,64].
- **Resource pricing**: computes the value of cloud resources that reflect the both economic and environment in cloud computing. It is required because how to resource pricing allocates limited resources among alternative cloud users for maximizing the usage of resources. It reduces the cost of resource for the cloud users and increases the profit and revenue for the cloud providers with maximum resource utilization [36,43,54,65,71,74,75,119].

Journal: 10586 Article No.: 0684 MS Code: CLUS-D-16-00024.1 TYPESET DISK LE CP Disp.:2016/11/30 Pages: 45 Layout: Large

Fig. 9 Innovative ideas for cloud computing



- **Resource provisioning**: is scheduling and allocation of resources to the cloud users from the cloud providers. This process is conducted in various ways to enhance the resource utilization in cloud computing such as adaptive resource provisioning, dynamic resource provisioning, user self-provisioning etc. [6,21,90].
- **Resource scheduling**: is a procedure or plan used to calculate the required resources deliver to the cloud users and when they will be required. It ensures that the efficient and effective utilization of resources, realistic confidence and early identification of resource capacity, restricted access and conflicts [6,8,22,75,89,113,130].
- **VM migration and placement**: is a procedure of transferring a running VM among various PMs in data centers without any interruption and disconnecting the cloud users. Processing, networking and storage connectivity is required during the VM migration from source to destination PMs [25,31,98,149,150].
- **Workload balancing**: is the procedure of allocating workloads and resources in a cloud computing systems. It requires initiatives to manage workload or users' demands by assigning resources among multiple computers, networks or servers. It also includes accommodating the distribution of workload and users' demands that exist in cloud computing [8,10,43,87,88,98,101,110,112,113].

To achieve the optimal solution for resource allocation, each algorithm, strategy or policy in cloud computing should be aware of the status of all resources in the infrastructure. Then, the technique should be applied to achieve a better allocation of physical or virtual resources to the cloud users, according to the requirements pre-established in SLA by the cloud providers.

Most of the research problems shown in the bubble graph are not addressed properly till date. Therefore, the authors recommend the application of recent meta-heuristic optimization techniques which have proven to be more effective than previous ones. These include league championship algorithm (LCA) [151] as detailed in [152], lion optimization algorithm (LOA) [153], optics inspired optimization (OIO) [154], sine cosine algorithm (SCA) [155], swallow swarm optimization (SSO) [156], teaching learning based optimization (TLBO) [157] and water wave optimization (WWO) [158] to mention but a few.

Further, meta-heuristics algorithm can be improved in term of quality of solutions or convergence speed by combining it with another population based, nature based, biology based or some local search based heuristic and meta-heuristic algorithms. One of the advantages of combining two population based meta-heuristic algorithms is that the shortcomings of one algorithm can be overcome by the strengths of another algorithm. Local based algorithms can be used to further improve the solution of population based algorithms. The best region in search problem is identified by population based meta-heuristic algorithms whereas the local search techniques help in finding the optimal solution. In addition, more research needs to consider other parameters aside from the dominant time and cost. Hence, the authors recommend that important parameters such as availability, priority, reliability and execution time should be considered.

7 Conclusion and recommendations

This paper presents a systematic review of resource allocation schemes and algorithms that are used by different researchers and categorized these approaches on the basis of problems addressed, schemes used and the performance of the approaches. Based on different studies considered in this review, we observed that different schemes did not consider some important parameters and enhancement is required to improve the performance of the existing schemes. This paper would help cloud administrators, users and researchers, who wish to carry out further research in resource allocation for cloud computing environment in future.

Cloud computing as a business model needs to consider user's priorities about resource availability and allocation. Therefore, IaaS cloud computing as an on-demand paradigm should improve on user's satisfaction through the priority based resource allocation. It is recommended for further research in the prioritization of resource allocation in relation to the finite available resources. Additionally, it is also recommended that an extensive research is needed on energy based resource allocation schemes especially with regard to the data center green optimization. This review is intended to serve as the basis for further research in resource allocation for IaaS cloud computing.

References

1. Jennings, B., Stadler, R.: Resource management in clouds: survey and research challenges. J. Netw. Syst. Manag. **23**, 567–619 (2015)
2. Whaiduzzaman, M., Haque, M.N., Chowdhury, M.R.K., Gani, A.: A study on strategic provisioning of cloud computing services. Sci. World J., 1–16 (2014)
3. Abdulhamid, S.M., Abd Latiff, M.S., Abdul-Salaam, G., Madni, S.H.H.: Secure scientific applications scheduling technique for cloud computing environment using global league championship algorithm. PLoS ONE **11**(7), e0158102 (2016)
4. Zhang, Q., Cheng, L., Boutaba, R.: Cloud computing: state-of-the-art and research challenges. J. Internet Serv. Appl. **1**(1), 7–18 (2010)
5. Madni, S.H.H., Latiff, M.S.A., Coulibaly, Y., Abdulhamid, S.I.M.: An appraisal of meta-heuristic resource allocation techniques for IaaS Cloud. Indian J. Sci. Technol. **9**(4), 1–14 (2016)
6. Manvi, S.S., Shyam, G.K.: Resource management for Infrastructure as a Service (IaaS) in cloud computing: a survey. J. Netw. Comput. Appl. **41**, 424–440 (2014)
7. Chana, I., Singh, S.: Quality of service and service level agreements for cloud environments: issues and challenges. In: Mahmood, Z. (ed.) Cloud Computing, pp. 51–72. Springer, New York (2014)
8. Ma, T., Chu, Y., Zhao, L., Ankhbayar, O.: Resource allocation and scheduling in cloud computing: policy and algorithm. IETE Tech. Rev. **31**(1), 4–16 (2014)
9. Parikh, S.M.: A survey on cloud computing resource allocation techniques. In: 2013 Nirma University International Conference on Engineering (NUiCONE), pp. 1–5. IEEE (2013)
10. Elghoneimy, E., Bouhali, O., Alnuweiri, H.: Resource allocation and scheduling in cloud computing. In: 2012 International Conference on Computing, Networking and Communications (ICNC), pp. 309–314. IEEE (2012)
11. Mohan, N., Raj, E.B.: Resource Allocation Techniques in Cloud Computing–Research Challenges for Applications. In: 2012 Fourth International Conference on Computational Intelligence and Communication Networks (CICN), pp. 556–560. IEEE (2012)
12. Hameed, A., Khoshkbarforoushha, A., Ranjan, R., Jayaraman, P.P., Kolodziej, J., Balaji, P., Zeadally, S., Malluhi, Q.M., Tziritas, N., Vishnu, A.: A survey and taxonomy on energy efficient resource allocation techniques for cloud computing systems. Computing **98**, 751–774 (2014)
13. Mustafa, S., Nazir, B., Hayat, A., Madani, S.A.: Resource management in cloud computing: taxonomy, prospects, and challenges. Comput. Electr. Eng. **47**, 186–203 (2015)
14. Pawar, C.S., Wagh, R.: A review of resource allocation policies in cloud computing. World J. Sci. Technol. **2**(3), 165–167 (2012)
15. Vinothina, V., Sridaran, R., Ganapathi, P.: A survey on resource allocation strategies in cloud computing. Int. J. Adv. Comput. Sci. Appl. **3**(6), 97–104 (2012)
16. Bi, J., Zhu, Z., Yuan, H.: SLA-aware dynamic resource provisioning for profit maximization in shared cloud data centers. In: Wu, Y. (ed.) High Performance Networking, Computing, and Communication Systems, pp. 366–372. Springer, Berlin (2011)
17. Abdulhamid, S.M., Latiff, M.S.A., Bashir, M.B.: Scheduling techniques in on-demand grid as a service cloud: a review. J. Theor. Appl. Inform. Technol. **63**, 10–19 (2014)
18. Endo, P.T., de Almeida Palhares, A.V., Pereira, N.N., Goncalves, G.E., Sadok, D., Kelner, J., Melander, B., Mångs, J.-E.: Resource allocation for distributed cloud: concepts and research challenges. IEEE Netw. **25**(4), 42–46 (2011)
19. Mohamaddiah, M.H., Abdullah, A., Subramaniam, S., Hussin, M.: A survey on resource allocation and monitoring in cloud computing. Int. J. Mach. Learn. Comput. **4**(1), 34 (2014)
20. Bashir, M.B., Abd Latiff, M.S., Abdulhamid, S.M., Loon, C.T.: Grid-based search technique for massive academic publications. Paper presented at the the 2014 third ICT international student project conference (ICT-ISPC2014), Thailand (2014)
21. Toosi, A.N., Calheiros, R.N., Buyya, R.: Interconnected cloud computing environments: challenges, taxonomy, and survey. ACM Comput. Surv. **47**(1), 7 (2014)
22. Huang, L., Chen, H.-S., Hu, T.-T.: Survey on resource allocation policy and job scheduling algorithms of cloud computing1. J. Softw. **8**(2), 480–487 (2013)
23. Gong, Y., Ying, Z., Lin, M.: A survey of cloud computing. In: Proceedings of the 2nd International Conference on Green Communications and Networks 2012 (GCN 2012), Vol. 3, pp. 79–84. Springer, New York (2013)
24. Ergu, D., Kou, G., Peng, Y., Shi, Y., Shi, Y.: The analytic hierarchy process: task scheduling and resource allocation in cloud computing environment. J. Supercomput. **64**(3), 835–848 (2013)
25. Mann, Z.Á.: Allocation of virtual machines in cloud data centers-a survey of problem models and optimization algorithms. ACM Comput. Serv. **48**, 11–34 (2015)
26. Akhter, N., Othman, M.: Energy aware resource allocation of cloud data center: review and open issues. Clust. Comput. **19**, 1163–1182 (2016)

27. Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L.A.: Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst. Rev. **4**(1), 1 (2015)
28. Kitchenham, B., Brereton, O.P., Budgen, D., Turner, M., Bailey, J., Linkman, S.: Systematic literature reviews in software engineering—a systematic literature review. Inform. Softw. Technol. **51**(1), 7–15 (2009)
29. Panda, S.K., Jana, P.K.: An efficient resource allocation algorithm for IaaS cloud. In: Distributed Computing and Internet Technology, pp. 351–355. Springer, New York (2015)
30. Shyam, G.K., Manvi, S.S.: Resource allocation in cloud computing using agents. In: 2015 IEEE International Advance Computing Conference (IACC), pp. 458–463. IEEE (2015)
31. An-ping, X., Chun-xiang, X.: Energy efficient multiresource allocation of virtual machine based on PSO in cloud data center. Mathematical Problems in Engineering (2014)
32. Radhakrishnan, A., Kavitha, V.: Trusted virtual machine allocation in cloud computing IaaS service. Res. J. Appl. Sci. Eng. Technol. **7**(14), 2921–2928 (2014)
33. Liang, Y., Rui, Q.P., Xu, J.: Computing resource allocation for enterprise information management based on cloud platform ant colony optimization algorithm. Adv. Mater. Res. **791**, 1232–1237 (2013)
34. Li, C., Li, L.: Efficient resource allocation for optimizing objectives of cloud users, IaaS provider and SaaS provider in cloud environment. J. Supercomput. **65**(2), 866–885 (2013)
35. Vernekar, S.S., Game, P.: Component based resource allocation in cloud computing. In: Proceedings of the International Conference on Information Systems Design and Intelligent Applications 2012 (INDIA 2012) held in Visakhapatnam, India, January 2012, pp. 907–914. Springer, New York (2012)
36. Wang, W., Jiang, Y., Wu, W.: Multiagent-based resource allocation for energy minimization in cloud computing systems. IEEE Transactions on Systems, Man and Cybernetics (2016)
37. Shelke, R., Rajani, R.: Dynamic resource allocation in cloud computing. Int. J. Eng. Res. Technol. **10** (2013)
38. Jayanthi, S.: Literature review: dynamic resource allocation mechanism in cloud computing environment. In: 2014 International Conference on Electronics, Communication and Computational Engineering (ICECCE), pp. 279–281. IEEE (2014)
39. Saraswathi, A., Kalaashri, Y., Padmavathi, S.: Dynamic resource allocation scheme in cloud computing. Proc. Comput. Sci. **47**, 30–36 (2015)
40. Wang, Z., Su, X.: Dynamically hierarchical resource-allocation algorithm in cloud computing environment. J. Supercomput. **71**, 2748–2766 (2015)
41. Wolke, A., Ziegler, L.: Evaluating dynamic resource allocation strategies in virtualized data centers. In: 2014 IEEE 7th International Conference on Cloud Computing (CLOUD), pp. 328–335. IEEE (2014)
42. Xie, F., Liu, F.: Dynamic effective resource allocation based on cloud computing learning model. J. Netw. **9**(11), 3092–3097 (2014)
43. Zhang, H.R., Yang, Y., Li, L., Cheng, W.Z., Ding, C.: A dynamic resource allocation framework in the cloud. Appl. Mech. Mater. **441**, 974–979 (2014)
44. Ali, J., Zafari, F., Khan, G.M., Mahmud, S.A.: Future clients' requests estimation for dynamic resource allocation in cloud data center using CGPANN. In: 2013 12th International Conference on Machine Learning and Applications (ICMLA), pp. 331–334. IEEE (2013)
45. Hu, W.X., Zheng, J., Hua, X.Y., Yang, Y.O.: A computing capability allocation algorithm for cloud computing environment. Appl. Mech. Mater. **347**, 2400–2406 (2013)
46. Oddi, G., Panfili, M., Pietrabissa, A., Zuccaro, L., Suraci, V.: A resource allocation algorithm of multi-cloud resources based on Markov decision process. In: 2013 IEEE 5th International Conference on Cloud Computing Technology and Science (CloudCom), pp. 130–135. IEEE (2013)
47. Xiao, Z., Song, W., Chen, Q.: Dynamic resource allocation using virtual machines for cloud computing environment. IEEE Trans. Parallel Distrib. Syst. **24**(6), 1107–1117 (2013)
48. Dai, J., Hu, B., Zhu, L., Han, H., Liu, J.: Research on dynamic resource allocation with cooperation strategy in cloud computing. In: 2012 3rd International Conference on System Science, Engineering Design and Manufacturing Information (ICSEM), pp. 193–196. IEEE (2012)
49. Hadji, M., Zeghlache, D.: Minimum cost maximum flow algorithm for dynamic resource allocation in clouds. In: 2012 IEEE 5th International Conference on Cloud Computing (CLOUD), pp. 876–882. IEEE (2012)
50. Wang, L.Y., Liu, A.M.: The study on cloud computing resource allocation method. Appl. Mech. Mater. **198**, 1506–1513 (2012)
51. Wuhib, F., Stadler, R., Lindgren, H.: Dynamic resource allocation with management objectives—implementation for an OpenStack cloud. In: 2012 8th International Conference and 2012 Workshop on Systems Virtualiztion Management (SVM) Network and Service Management (CNSM), pp. 309–315. IEEE (2012)
52. Yin, B., Wang, Y., Meng, L., Qiu, X.: A multi-dimensional resource allocation algorithm in cloud computing. J. Inform. Comput. Sci. **9**(11), 3021–3028 (2012)
53. Zhang, Q., Zhu, Q., Boutaba, R.: Dynamic resource allocation for spot markets in cloud computing environments. In: 2011 Fourth IEEE International Conference on Utility and Cloud Computing (UCC), pp. 178–185. IEEE (2011)
54. Teng, F., Magoulès, F.: A new game theoretical resource allocation algorithm for cloud computing. In: Advances in Grid and Pervasive Computing. Lecture Notes on Computer Science, vol. 6104, pp. 321–330. Springer, Berlin (2010)
55. Wuhib, F., Yanggratoke, R., Stadler, R.: Allocating compute and network resources under management objectives in large-scale clouds. J. Netw. Syst. Manag. **23**(1), 111–136 (2015)
56. Islam, S., Keung, J., Lee, K., Liu, A.: Empirical prediction models for adaptive resource provisioning in the cloud. Future Gener. Comput. Syst. **28**(1), 155–162 (2012)
57. Patel, R., Dahiya, D.: Aggregation of cloud providers: a review of opportunities and challenges. In: 2015 International Conference on Computing, Communication & Automation (ICCCA), pp. 620–626. IEEE (2015)
58. Dabbagh, M., Hamdaoui, B., Guizani, M., Rayes, A.: Energy-efficient resource allocation and provisioning framework for cloud data centers. IEEE Trans. Netw. Serv. Manage. **12**(3), 377–391 (2015)
59. Vasu, R., Nehru, E.I., Ramakrishnan, G.: Load forecasting for optimal resource allocation in cloud computing using neural method. Middle-East J. Sci. Res. **24**(6), 1995–2002 (2016)
60. Wang, C.-F., Hung, W.-Y., Yang, C.-S.: A prediction based energy conserving resources allocation scheme for cloud computing. In: 2014 IEEE International Conference on Granular Computing (GrC), pp. 320–324. IEEE (2014)
61. Mashayekhy, L., Nejad, M.M., Grosu, D., Vasilakos, A.V.: An online mechanism for resource allocation and pricing in clouds. IEEE Trans. Comput. **65**(4), 1172–1184 (2016)
62. Goutam, S., Yadav, A.K.: Preemptable priority based dynamic resource allocation in cloud computing with fault tolerance. In: 2015 International Conference on Communication Networks (ICCN), pp. 278–285. IEEE (2015)
63. Wu, X., Gu, Y., Tao, J., Li, G., Jayaraman, P.P., Sun, D., Ranjan, R., Zomaya, A., Han, J.: An online greedy allocation of VMs

Journal: **10586** Article No.: **0684** MS Code: **CLUS-D-16-00024.1** ☐ TYPESET ☐ DISK ☐ LE ☐ CP Disp.:**2016/11/30** Pages: **45** Layout: **Large**

with non-increasing reservations in clouds. J. Supercomput. **72**(2), 371–390 (2016)

64. Wu, X., Gu, Y., Li, G., Tao, J., Chen, J., Ma, X.: Online mechanism design for VMS allocation in private cloud. In: IFIP International Conference on Network and Parallel Computing, pp. 234–246. Springer, Berlin (2014)
65. Gu, Y., Tao, J., Wu, X., Ma, X.: Online mechanism with latest-reservation for dynamic VMs allocation in private cloud. Int. J. Syst. Assur. Eng. Manag. (2016). doi:10.1007/s13198-016-0422-6
66. Qian, L., Luo, Z., Du, Y., Guo, L.: Cloud computing: an overview. In: Jaatun, M.G., Zhao, G., Rong, C. (eds.) Cloud Computing, pp. 626–631. Springer, Berlin (2009)
67. Kumar, N., Saxena, S.: A preference-based resource allocation in cloud computing systems. Proc. Comput. Sci. **57**, 104–111 (2015)
68. Mohana, R.: A position balanced parallel particle swarm optimization method for resource allocation in cloud. Indian J. Sci. Technol. **8**(S3), 182–188 (2015)
69. Nezarat, A., Dastghaibifard, G.: Efficient nash equilibrium resource allocation based on game theory mechanism in cloud computing by using auction. PloS ONE **10**(10), e0138424 (2015)
70. Samimi, P., Teimouri, Y., Mukhtar, M.: A combinatorial double auction resource allocation model in cloud computing. Inform. Sci. **357**, 201–216 (2016)
71. Li, H., Pu, Y., Lu, J.: A cloud computing resource pricing strategy research-based on resource swarm algorithm. In: 2012 International Conference on Computer Science & Service System (CSSS), pp. 2217–2222. IEEE (2012)
72. Chintapalli, V.R.: A deadline and budget constrained cost and time optimization algorithm for cloud computing. In: International Conference on Advances in Computing and Communications, pp. 455–462. Springer, Berlin (2011)
73. Teng, F., Magoules, F.: Resource pricing and equilibrium allocation policy in cloud computing. In: 2010 IEEE 10th International Conference on Computer and Information Technology (CIT), pp. 95–202. IEEE (2010)
74. Kumar, K., Feng, J., Nimmagadda, Y., Lu, Y.-H.: Resource allocation for real-time tasks using cloud computing. In: 2011 Proceedings of 20th International Conference on Computer Communications and Networks (ICCCN), pp. 1–7. IEEE (2011)
75. Yi, P., Ding, H., Ramamurthy, B.: Budget-minimized resource allocation and task scheduling in distributed grid/clouds. In: 2013 22nd International Conference on Computer Communications and Networks (ICCCN), pp. 1–8. IEEE (2013)
76. Casalicchio, E., Menascé, D.A., Aldhalaan, A.: Autonomic resource provisioning in cloud systems with availability goals. In: Proceedings of the 2013 ACM Cloud and Autonomic Computing Conference. ACM, New York (2013)
77. Gu, Y., Tao, J., Li, G., Sun, D.W., Wu, X., Jayaraman, P.P., Ranjan, R.: A preemptive truthful VMs allocation online mechanism in private cloud. J. Comput. Sci. (2016). doi: 10.1016/j.jocs.2016.05.006
78. Younge, A.J., Von Laszewski, G., Wang, L., Lopez-Alarcon, S., Carithers, W.: Efficient resource management for cloud computing environments. In: 2010 International Green Computing Conference, pp. 357–364. IEEE (2010)
79. Mashayekhy, L., Nejad, M.M., Grosu, D., Vasilakos, A.V.: Incentive-compatible online mechanisms for resource provisioning and allocation in clouds. In: 2014 IEEE 7th International Conference on Cloud Computing (CLOUD), pp. 312–319. IEEE (2014)
80. Nejad, M.M., Mashayekhy, L., Grosu, D.: Truthful greedy mechanisms for dynamic virtual machine provisioning and allocation in clouds. IEEE Trans. Parallel Distrib. Syst. **26**(2), 594–603 (2015). doi:10.1109/tpds.2014.2308224
81. Pradhan, P., Behera, P.K., Ray, B.: Modified round robin algorithm for resource allocation in cloud computing. Proc. Comput. Sci. **85**, 878–890 (2016)
82. Yang, Z., Liu, M., Xiu, J., Liu, C.: Study on cloud resource allocation strategy based on particle swarm ant colony optimization algorithm. In: 2012 IEEE 2nd International Conference on Cloud Computing and Intelligent Systems (CCIS), pp. 488–491. IEEE (2012)
83. Xu, X., Yu, H.: A game theory approach to fair and efficient resource allocation in cloud computing. Mathematical Problems in Engineering (2014)
84. Kaur, R., Luthra, P.: Load balancing in cloud computing. In: Proceedings of International Conference on Recent Trends in Information, Telecommunication and Computing, ITC. (2012)
85. Aslam, S., Shah, M.A.: Load balancing algorithms in cloud computing: a survey of modern techniques. In: 2015 National Software Engineering Conference (NSEC), pp. 30–35. IEEE (2015)
86. Katyal, M., Mishra, A.: A comparative study of load balancing algorithms in cloud computing environment. (2014). arXiv:1403.6918
87. Parikh, K., Hawanna, N., Haleema, P.K., Iyengar, N.C.S.: Virtual machine allocation policy in cloud computing using CloudSim in Java. Int. J. Grid Distrib. Comput. **8**(1), 145–158 (2015)
88. Bhise, V.K., Mali, A.S.: Cloud resource provisioning for Amazon EC2. In: 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), pp. 1–7. IEEE (2013)
89. Ray, S., Sarkar, A.D.: Resource allocation scheme in cloud infrastructure. In: 2013 International Conference on Cloud & Ubiquitous Computing & Emerging Technologies (CUBE), pp. 30–35. IEEE (2013)
90. Villegas, D., Antoniou, A., Sadjadi, S.M., Iosup, A.: An analysis of provisioning and allocation policies for infrastructure-as-a-service clouds. In: 2012 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid), pp. 612–619. IEEE (2012)
91. Zhang, Z., Wang, H., Xiao, L., Ruan, L.: A statistical based resource allocation scheme in cloud. In: 2011 International Conference on Cloud and Service Computing (CSC), pp. 266–273. IEEE (2011)
92. Liu, L., Mei, H., Xie, B.: Towards a multi-QoS human-centric cloud computing load balance resource allocation method. J. Supercomput. **72**, 2488–2501 (2016)
93. Buyya, R., Beloglazov, A., Abawajy, J.: Energy-efficient management of data center resources for cloud computing: a vision, architectural elements, and open challenges. (2010). arXiv:1006.0308
94. Beloglazov, A., Buyya, R.: Energy efficient resource management in virtualized cloud data centers. In: Proceedings of the 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, pp. 826–831. IEEE Computer Society (2010)
95. Pandi, K.M., Somasundaram, K.: Energy efficient in virtual infrastructure and green cloud computing: a review. Indian J. Sci. Technol. (2016). doi:10.17485/ijst/2016/v9i11/89399
96. Singh, S.: Green computing strategies & challenges. In: 2015 International Conference on Green Computing and Internet of Things (ICGCIoT), pp. 758–760. IEEE (2015)
97. Ali, A., Lu, L., Zhu, Y., Yu, J.: An energy efficient algorithm for virtual machine allocation in cloud datacenters. In: Conference 2016, pp. 61–72. Springer, Berlin
98. Beloglazov, A., Abawajy, J., Buyya, R.: Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing. Future Gener. Comput. Syst. **28**(5), 755–768 (2012)
99. Dashti, S.E., Rahmani, A.M.: Dynamic VMs placement for energy efficiency by PSO in cloud computing. J. Exp. Theor. Artif. Intell. **28**, 351–367 (2016)

100. Gao, Y., Guan, H., Qi, Z., Hou, Y., Liu, L.: A multi-objective ant colony system algorithm for virtual machine placement in cloud computing. J. Comput. Syst. Sci. **79**(8), 1230–1242 (2013)
101. Kansal, N.J., Chana, I.: Artificial bee colony based energy-aware resource utilization technique for cloud computing. Concurr. Comput. **27**(5), 1207–1225 (2015)
102. Yanggratoke, R., Wuhib, F., Stadler, R.: Gossip-based resource allocation for green computing in large clouds. In: 2011 7th International Conference on Network and Service Management (CNSM), pp. 1–9. IEEE (2011)
103. Jha, R.S., Gupta, P.: Power & load aware resource allocation policy for hybrid cloud. Proc. Comput. Sci. **78**, 350–357 (2016)
104. Gupta, P., Ghrera, S.: Power and fault aware reliable resource allocation for cloud infrastructure. Proc. Comput. Sci. **78**, 457–463 (2016)
105. Pavithra, B., Ranjana, R.: Energy efficient resource provisioning with dynamic VM placement using energy aware load balancer in cloud. In: 2016 International Conference on Information Communication and Embedded Systems (ICICES), pp. 1–6. IEEE (2016)
106. Peng, Y., Kang, D.-K., Al-Hazemi, F., Youn, C.-H.: Energy and QoS aware resource allocation for heterogeneous sustainable cloud datacenters. Optical Switching and Networking (2016)
107. Singh, K., Kaushal, S.: Energy efficient resource provisioning through power stability algorithm in cloud computing. In: Proceedings of the International Congress on Information and Communication Technology, pp. 255–263. Springer, Berlin (2016)
108. Abdelmaboud, A., Jawawi, D.N., Ghani, I., Elsafi, A., Kitchenham, B.: Quality of service approaches in cloud computing: a systematic mapping study. J. Syst. Softw. **101**, 159–179 (2015)
109. Ardagna, D., Casale, G., Ciavotta, M., Pérez, J.F., Wang, W.: Quality-of-service in cloud computing: modeling techniques and their applications. J. Internet Serv. Appl. **5**(1), 1–17 (2014)
110. Batista, B.G., Estrella, J.C., Ferreira, C.H.G., Leite Filho, D.M., Nakamura, L.H.V., Reiff-Marganiec, S., Santana, M.J., Santana, R.H.C.: Performance evaluation of resource management in cloud computing environments. PloS ONE **10**(11), 1–21 (2015)
111. Li, J., Li, D., Ye, Y., Lu, X.: Efficient multi-tenant virtual machine allocation in cloud data centers. Tsinghua Sci. Technol. **20**(1), 81–89 (2015)
112. Horri, A., Mozafari, M.S., Dastghaibyfard, G.: Novel resource allocation algorithms to performance and energy efficiency in cloud computing. J. Supercomput. **69**(3), 1445–1461 (2014)
113. Katyal, M., Mishra, A.: Application of selective algorithm for effective resource provisioning in cloud computing environment. (2014). arXiv:1403.2914
114. Lee, H.M., Jeong, Y.-S., Jang, H.J.: Performance analysis based resource allocation for green cloud computing. J. Supercomput. **69**(3), 1013–1026 (2014)
115. Li, Y.K.: QoS-aware dynamic virtual resource management in the cloud. In: Applied Mechanics and Materials, pp. 5809–5812. Trans Tech Publ 1 (2014)
116. Pan, B.L., Wang, Y.P., Li, H.X., Qian, J.: Task scheduling and resource allocation of cloud computing based on QoS. Adv. Mater. Res. **915**, 1382–1385 (2014)
117. Kang, Z., Wang, H.: A novel approach to allocate cloud resource with different performance traits. In: 2013 IEEE International Conference on Services Computing (SCC), pp. 128–135. IEEE (2013)
118. Nathani, A., Chaudhary, S., Somani, G.: Policy based resource allocation in IaaS cloud. Future Gener. Comput. Syst. **28**(1), 94–103 (2012)
119. Sagbo, K.A.R., Houngue, P.: Quality architecture for resource allocation in cloud computing. In: Service-Oriented and Cloud Computing. pp. 154–168. Springer, Berlin (2012)
120. Wei, G., Vasilakos, A.V., Zheng, Y., Xiong, N.: A game-theoretic method of fair resource allocation for cloud computing services. J. Supercomput. **54**(2), 252–269 (2010)
121. Nguyen, T.-D., Nguyen, A.T., Nguyen, M.D., Van Nguyen, M., Huh, E.-N.: An improvement of resource allocation for migration process in cloud environment. Comput. J. **57**(2), 308–318 (2013)
122. Papagianni, C., Leivadeas, A., Papavassiliou, S., Maglaris, V., Cervelló-Pastor, C., Monje, A.: On the optimal allocation of virtual resources in cloud computing networks. IEEE Trans. Comput. **62**(6), 1060–1071 (2013)
123. Kumar, N., Chilamkurti, N., Zeadally, S., Jeong, Y.-S.: Achieving quality of service (QoS) using resource allocation and adaptive scheduling in cloud computing with grid support. Comput. J. **57**(2), 281–290 (2014)
124. Guo, J., Liu, F., Lui, J.C., Jin, H.: Fair network bandwidth allocation in IaaS datacenters via a cooperative game approach. IEEE/ACM Trans. Netw. **24**(2), 873–886 (2016)
125. Wang, H., Wang, F., Liu, J., Wang, D., Groen, J.: Enabling customer-provided resources for cloud computing: potentials, challenges, and implementation. IEEE Trans. Parallel Distrib. Syst. **26**(7), 1874–1886 (2015)
126. Brummett, T., Galloway, M.: Towards providing resource management in a local IaaS cloud architecture. In: Information Technology: New Generations, pp. 413–423. Springer, Berlin (2016)
127. Lin, C.H., Lu, C.T., Chen, Y.H., Li, J.S.: Resource allocation in cloud virtual machines based on empirical service traces. Int. J. Commun. Syst. **27**(12), 4210–4225 (2014)
128. Pillai, P.S., Rao, S.: Resource allocation in cloud computing using the uncertainty principle of game theory. IEEE Syst. J. **10**(2), 637–648 (2016)
129. Rezvani, M., Akbari, M.K., Javadi, B.: Resource allocation in cloud computing environments based on integer linear programming. Comput. J. **52**(2), 300–314 (2014)
130. Srinivasa, K., Srinidhi, S., Kumar, K.S., Shenvi, V., Kaushik, U.S., Mishra, K.: Game theoretic resource allocation in cloud computing. In: 2014 Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT), pp. 36–42. IEEE (2014)
131. Tyagi, M., Manoria, M.: Secured data storage and computation technique for effective utilization of servers in cloud computing. In: Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems, vol. 1, pp. 531–541. Springer, Berlin (2016)
132. Mell, P., Grance, T.: The NIST definition of cloud computing. Computer Security Division, Information Technology Laboratory (2011)
133. Nabi, M., Toeroe, M., Khendek, F.: Availability in the cloud: state of the art. J. Netw. Comput. Appl. **60**, 54–67 (2016)
134. Hassan, S., Abbas Kamboh, A., Azam, F.: Analysis of cloud computing performance, scalability, availability, & security. In: 2014 International Conference on Information Science and Applications (ICISA), pp. 1–5. IEEE (2014)
135. Buyya, R., Yeo, C.S., Venugopal, S., Broberg, J., Brandic, I.: Cloud computing and emerging IT platforms: vision, hype, and reality for delivering computing as the 5th utility. Future Gener. Comput. Syst. **25**(6), 599–616 (2009)
136. Li, X., Li, Y., Liu, T., Qiu, J., Wang, F.: The method and tool of cost analysis for cloud computing. In: IEEE International Conference on Cloud Computing, 2009, CLOUD'09, pp. 93–100. IEEE (2009)
137. Tziritas, N., Xu, C.-Z., Loukopoulos, T., Khan, S.U., Yu, Z.: Application-aware workload consolidation to minimize both energy consumption and network load in cloud environments. In: 2013 42nd International Conference on Parallel Processing (ICPP), pp. 449–457. IEEE (2013)

Journal: **10586** Article No.: **0684** MS Code: **CLUS-D-16-00024.1** ☐ TYPESET ☐ DISK ☐ LE ☐ CP Disp.:**2016/11/30** Pages: **45** Layout: **Large**

138. Madni, S.H.H., Latiff, M.S.A., Coulibaly, Y.: Resource scheduling for infrastructure as a service (IaaS) in cloud computing: challenges and opportunities. J. Netw. Comput. Appl. **68**, 173–200 (2016)
139. Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I.: A view of cloud computing. Commun. ACM **53**(4), 50–58 (2010)
140. Xiong, K., Perros, H.: Service performance and analysis in cloud computing. In: 2009 World Conference on Services-I, pp. 693–700. IEEE (2009)
141. Faragardi, H.R., Shojaee, R., Tabani, H., Rajabi, A.: An analytical model to evaluate reliability of cloud computing systems in the presence of QoS requirements. In: 2013 IEEE/ACIS 12th International Conference on Computer and Information Science (ICIS), pp. 315–321. IEEE (2013)
142. Bashir, M.B., Abd Latiff, M.S., Ahmed, A.A., Yousif, A., Eltayeeb, M.E.: Content-based information retrieval techniques based on grid computing: a review. IETE Tech. Rev. **30**(3), 223–232 (2013)
143. Patel, P., Ranabahu, A.H., Sheth, A.P.: Service level agreement in cloud computing (2009)
144. Jing, S.-Y., Ali, S., She, K., Zhong, Y.: State-of-the-art research study for green cloud computing. J. Supercomput. **65**(1), 445–468 (2013)
145. Garg, S.K., Buyya, R.: Green cloud computing and environmental sustainability. Harnessing Green IT: Principles and Practices, pp. 315–340 (2012)
146. Abdullahi, M., Ngadi, M.A.: Hybrid symbiotic organisms search optimization algorithm for scheduling of tasks on cloud computing environment. PloS ONE **11**(6), e0158229 (2016)
147. Hooper, A.: Green computing. Commun. ACM **51**(10), 11–13 (2008)
148. Chen, M., Mao, S., Liu, Y.: Big data: a survey. Mobile Netw. Appl. **19**(2), 171–209 (2014)
149. Pecero, J.E., Diaz, C.O., Castro, H., Villamizar, M., Sotelo, G., Bouvry, P.: Energy savings on a cloud-based opportunistic infrastructure. In: Service-Oriented Computing–ICSOC 2013 Workshops, pp. 366–378. Springer, Berlin (2014)
150. Jebalia, M., Ben Letaïfa, A., Hamdi, M., Tabbane, S.: A comparative study on game theoretic approaches for resource allocation in cloud computing architectures. In: 2013 IEEE 22nd International Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), pp. 336–341. IEEE (2013)
151. Kashan, A.H., Karimi, B.: A new algorithm for constrained optimization inspired by the sport league championships. In: 2010 IEEE Congress on Evolutionary Computation (CEC), pp. 1–8. IEEE (2010)
152. Abdulhamid, S.M., Latiff, M.S.A., Madni, S.H.H., Oluwafemi, O.: A survey of league championship algorithm: prospects and challenges. Indian J. Sci. Technol. **8**(S3), 101–110 (2015)
153. Yazdani, M., Jolai, F.: Lion optimization algorithm (LOA): a nature-inspired metaheuristic algorithm. J. Comput. Design Eng. **3**(1), 24–36 (2016)
154. Kashan, A.H.: A new metaheuristic for optimization: optics inspired optimization (OIO). Comput. Oper. Res. **55**, 99–125 (2015)
155. Mirjalili, S.: SCA: a sine cosine algorithm for solving optimization problems. Knowledge-Based Systems (2016)
156. Neshat, M., Sepidnam, G., Sargolzaei, M.: Swallow swarm optimization algorithm: a new method to optimization. Neural Comput. Appl. **23**(2), 429–454 (2013)
157. Rao, R.V., Savsani, V.J., Vakharia, D.: Teaching-learning-based optimization: a novel method for constrained mechanical design optimization problems. Comput. Aided Des. **43**(3), 303–315 (2011)
158. Zheng, Y.-J.: Water wave optimization: a new nature-inspired metaheuristic. Comput. Oper. Res. **55**, 1–11 (2015)

Journal: **10586** Article No.: **0684** MS Code: **CLUS-D-16-00024.1** ☐ TYPESET ☐ DISK ☐ LE ☐ CP Disp.:**2016/11/30** Pages: **45** Layout: **Large**

Journal: 10586 Article: 684

Springer
the language of science

Author Query Form

Please ensure you fill out your response to the queries raised below and return this form along with your corrections

Dear Author

During the process of typesetting your article, the following queries have arisen. Please check your typeset proof carefully against the queries listed below and mark the necessary changes either directly on the proof/online grid or in the 'Author's response' area provided below

Query	Details required	Author's response
1.	As References [18] and [29] are same, we have deleted the duplicate reference and renumbered accordingly. Please check and confirm.	It is ok, its was duplicated reference.
2.	As no significance have been provided for the colors in Tables 13–21, hence, colors have been removed. Please check and confirm the edit and amend if necessary.	All the formatting of Tables are very nice. I have one suggestion for Table 20, which I mention with the highlighted text of Table 20.