A Comprehensive Study on Cloud Computing: Architecture, Load Balancing, Task Scheduling and Meta-Heuristic Optimization

Shruti Tiwari and Chinmay Bhatt

Abstract Cloud computing (CC) is evolving computing model with a vast array of heterogeneous autonomous systems by modular computational architecture. Load balancing of activities on the cloud environment is an essential part of distributing services from the data center. CC is agonized by overloading demands because of dynamic computing through the internet. Load balancing must be done to ensure maximum use of the resources in all virtual machines (VM). Task scheduling is a crucial step for improving cloud computing's overall efficacy. Task scheduling is therefore significant to minimize energy usage and increase service providers' benefit by reducing the time required. This work provides a detailed study about the cloud computing architecture, load balancing (LB) mechanism, task scheduling (TS) framework in the cloud environment. Various meta-heuristic optimization techniques have been implemented to manage the load over virtual machines using task scheduling and load balancing terminologies. Various research gaps and issues have been identified from the literary work done by various researchers. This comprehensive study has motivated and provided us future direction to do work in this field.

Keywords Cloud computing · Virtual machines · Task scheduling · Load balancing · Meta-heuristic optimization

1 Introduction

Cloud Computing [1, 2] is an evolving software deployment and maintenance trend that is being embraced by industries like Microsoft, IBM, Google, and eBay. IBM-Blue Cloud framework, eBay Cloud, Google App Engine, as well as Distributed Computing Platform are many conceptual systems and frameworks.

S. Tiwari (✉) · C. Bhatt
Department of Computer Science Engineering, RKDF College, Bhopal, (M.P.), India
e-mail: shruti.tiwari08@gmail.com

J. Hemanth et al. (eds.), *Intelligent Cyber Physical Systems and Internet of Things*,
Engineering Cyber-Physical Systems and Critical Infrastructures 3,
https://doi.org/10.1007/978-3-031-18497-0_11

Cloud Computing is seen as the upcoming development which would influence organizational organizations and also how they handle their IT infrastructure and services. A key researching area is infrastructure and technology which cloud services and deployed models have presented.

The Internet has always been a driving factor in the advancement of different technologies. Cloud Computing is arguably the most debated amongst all these. The cloud computing model has seen an immense shift toward this usage over the last several months and has become a standard in the IT space as it offers its users and suppliers substantial cost savings including new business potential [3]. The benefits that use cloud computing are including:

(1) Minimization in the maintenance cost and its hardware cost.
(2) Accessible to all.
(3) Flexible including automatized procedures in which the client does not have to think about complex problems such as software up-gradation [4, 5].

Although the concept of cloud computing has various variations, this new computing model is defined by certain fundamental concepts. Cloud Computing offers technical resources that are provided on-demand as-a-service through the use of the Web, typically managed across premises. Big Evolution in the field of computer science, a service. CC can efficiently and securely deliver various IT services and resources according to the requirement of the user. Cloud computing offers usage-driven applications. It provides services like Infrastructure as a Services (IaaS), Platform as a Services (PaaS), and Software as a Services (SaaS) [6]. Clients of such resources don't own infrastructure in the remote cloud, yet pay for resources on the per usage, provided that a 3rd party operates and maintains the public cloud. The main principle, therefore, is the virtualization of resources. They rent the physical infrastructure, systems, and software inside a shared structure in a realistic situation. From virtual networks, computing systems, integrated data center, end-user web apps, and web services to immense computing-oriented service, security features will differ.

In many fields of ITs, cloud computing can be used to resolve issues such as GIS, Science Research, Decision Making, ERP, e-Governance Systems [7], Web App Creation, Mobile Technologies, etcetera. As the main back-end computing infrastructure, cloud computing depends on data centers [8]. With the need for cloud computing more dramatically in recent years, geographically dispersed data centers are offering more and more cloud resources to improve the reliability and consistency of services. A huge amount of energy is essential to operate these geographically dispersed data centers, which represent around 15% of the overall cost of a data center (DC) [9].

However, considering data centers as supercomputers with shared resources or making the second stage simple, consistent with servers, new approaches to geographical load-balancing (GLB) as in Fig. 1 [10] mostly concentrate on the first phase, i.e. the proportion of work demands that are assigned to each data center, of employment planning. Moreover, the existing GLB approaches often rely on a specific resource dimension (for example, CPU) and consistent resource need,

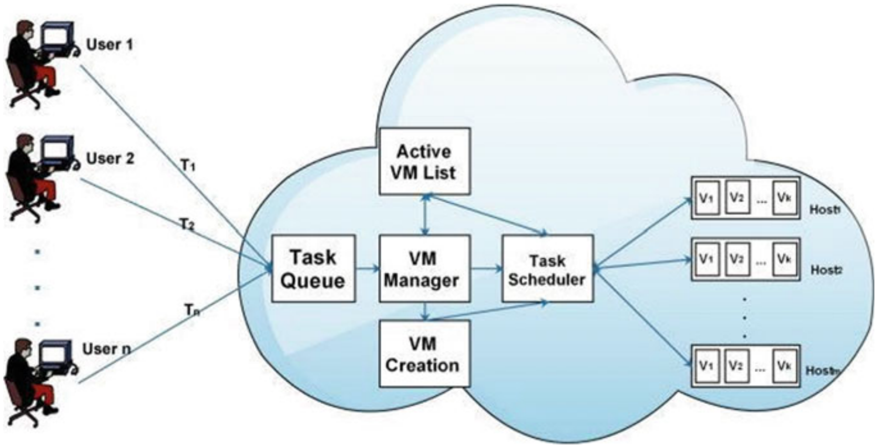

Fig. 1 Cloud load balancing

without taking into account the requirements of many separate jobs and heterogeneous resources. This is significant, in particular, because modern data centers are usually built from a variety of server groups with various processing capacity, size of memory, and storage spaces specifications [11]. The diversity of the market profile for workloads for different occupations exacerbates such heterogeneity further, e.g. CPU intensive numerical computing activities are generally required while high memory support is normally required for database operations [12].

Load must be spread among the cloud-based networks such that there must be no over-utilized or underutilized nodes within the networking process. LB is a common cloud problem that makes it difficult for applications adjacent to QoS (quality of service) measurement to sustain performance according to the SLA (service level agreement) document essential by enterprise cloud providers. Cloud providers have difficulty distributing similar workloads through servers. Effective LB technology can maximize and guarantee high user satisfaction by effectively using VMs resources [13].

In the cloud environment, a load must be balanced through various techniques. Algorithms of LB may be split into Algorithms for dynamic LB and static LB. Load is spread by previous knowledge and information in the Static LB Algorithm. Static algorithms at the same time do not take account of the current workload [14]. These algorithms are for the less work-loaded cloud. Dynamic LB Algorithms take into account current work Cloud load. Centralized and Semi-distributed LB algorithms can be categorized into dynamic LB algorithms [15]. Intelligent strategies like GA (Genetic Algorithms), PSO (Particle Swarm Optimization), ACS (Ant Colony System), and ABC (Artificial Bee Colony) can address problems of load balancing [16]. Various analysis was undertaken in the area of LB and TS in the cloud environments.

The task scheduling [17], known for providing essential cloud service efficiency, currently represents a related hot topic. However, due to inappropriate scheduling, the dilemmas of resources that are underused (unloaded) and overused (overloaded), may emerge, leading respectively to either wastage of cloud resources or a decrease in services. So, the idea to incorporate meta-heuristic algorithms into TS has arisen so that complicated and varied incoming tasks (cloudlets) can be easily distributed within a reasonable time to available resources. The meta-heuristic techniques have shown themselves to be very able to solve scheduling problems, which are met by giving a detailed overview of traditional and heuristic approaches before deeply expanding into common cloud task meta-heuristic methods, then comprehensive systematic review, which includes new taxonomy and advantage of those methods.

More specifically, major contribution keys of this study may be organized as follows:

(1) Comprehensive study about cloud computing evolution and architecture
(2) Studying the mechanisms of existing load balancing.
(3) Providing an overview about task scheduling in cloud computing
(4) Identify important areas where new research can be performed with the optimization principle to better load balancing algorithms.
(5) A systematic analysis is presented on cloud meta-heuristic TS.
(6) Identifying research gaps and issues exist that can be further hurdle to researchers for LB algorithms

The below is paper systematized: Section 2 elaborates the cloud computing concept with their architecture. It also consists of data center networking for VMs. At last, it talks about motivation of study. Section 3 elaborates the review of load balancing techniques, and task scheduling in CC in Sect. 4. Metaheuristic optimization algorithm comparison shows in Sect. 5. Section 6 explains related work also identifies some research gaps and issues and Sect. 7 concludes this work.

2 Cloud Computing: Architecture

Web-based tools and technologies can be found in cloud computing resources. This allows the users to work remotely because the cloud can be used as “Internet”. Therefore, it is not processed as traditional outsourcing. It is also called Massive Computing. In this, the allocation of applications must be dynamic. No hardware or software has to be installed. Cloud computing aims to enable people who have no deep knowledge of information about all technology and applications [18].

2.1 Overview

"Cloud" is a virtualized reusable resource computing pool. It can change or handle a variety of workloads. The "Cloud' term is a tarn of enormous infrastructure possessions, offering various utilities and hardware implementations as well as device software to end-users, alternatively enabling end-users to access their computing needs inconsistent way. The user needs no information about where to discover the device requirement and how the cloud operates. Cloud providers manage all obtainable tools to consumers where 'computing' is defined on an 'SLA'. Cloud computing offers inevitable benefit in sharing over the internet of many configurable system properties and higher-level administrations that may be built with very little board effort. Cloud computing has been developed to promote the use of virtualization technology to allow end-users to use virtual services without infrastructure at a reasonable cost [19]. Cloud Computing Evolution in IT in shown in the Fig. 2.

Now we are describing different modes of cloud computing as follows [21]:

(1) **Public Cloud**: They are managed by CSP which owns facilities and data centers. Infrastructure is located on-site and companies may use pay-as-you-go and on-demand services.
(2) **Private Cloud**: It is only created and operated by certain enterprises however 3rd party companies have control on behalf of the cloud owner to handle it.
(3) **Hybrid Cloud**: It combines an only selection of all kinds of cloud deployments, such as public, private, or community cloud. Core operations are conducted in a private cloud, while a public cloud offers fewer basic resources.
(4) **Community Cloud**: Several organizations or institutions of a shared purpose are sharing the community cloud. Universities that do this for learning and research are typical examples.

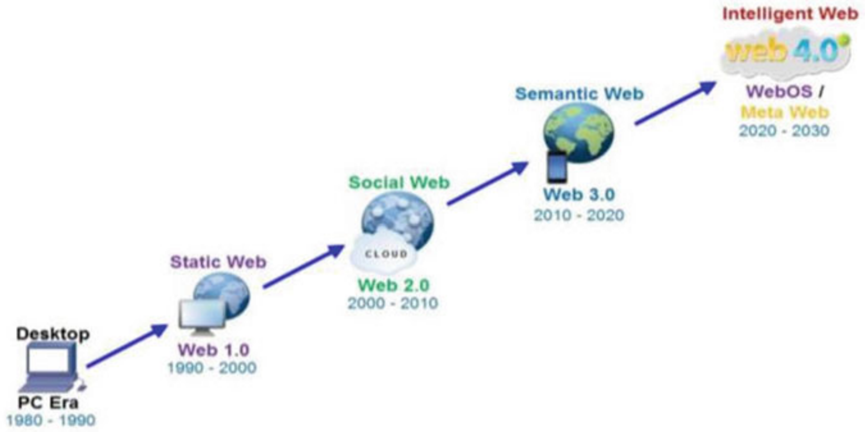

Fig. 2 Cloud computing evolution in IT [20]

2.2 Features of Cloud Computing

Compared with other computing paradigms, CC provides a range of new features [22] and advantages. This section is briefly identified.

- **Scalability and On-Demand Services**: Cloud computing connects customers on request with resources and services. Resources may be scaled across several data centers.
- **QoS (Quality of Service)**: In terms of CPU performance or hardware, memory capacity, and bandwidth, CC may provide QoS for users.
- **User-Centric Interface**: Cloud interfaces are locational isolated and well-defined interfaces including web services and web browsers enable them to be accessed.
- **Autonomous System**: CC applications are user-friendly, transparently controlled decentralized systems. But, software and data inside clouds may be reconfigured and merged automatically into a basic, user-specific platform.
- **Pricing**: Cloud computing requires no investment at the start of the project. There is no need for capital expenses. Users can pay for services and capacities or pay for them, as they require them.

2.3 Architecture

The applications, data, and utilities are all maintained in the cloud over the Internet and run applications, as well as stored data through the provision of software resources on-demand services in CC architecture. The back end and front end of a cloud architecture can be separated. The front end is available to the user by internet connections, enabling user interactions with the system [23]. The back end includes different cloud service models [24], like SaaS, PaaS, and IaaS. Often referred to as 'Layered computing model' is cloud computing architecture [25]. CC architecture can be classified into 4 layers that are hardware layer, infrastructure layer, platform layer, application layer as seen in Fig. 3.

Description of each layer is defined as follows [26]:

- **Hardware Layer**: Cloud is handled with physical resources. Controlling physical servers, switches, routers, the power system is the responsibility of the hardware layer. The implementation of the hardware layer is provided in a data center. There are several servers interconnected by routers and switches in the data center. Hardware layers have several issues including fault tolerance, hardware configuration, traffic management, and management of resources.
- **Infrastructure Layer**: The virtualization layer is named as well. Cloud computing is an important aspect. Infrastructure layers focused on core aspects such as the use of virtualized technologies for the dynamic assignment of resources. The infrastructure layer uses virtualization techniques to collect processing and storage resources and divide physical resources. E.g. Xen, VMware.

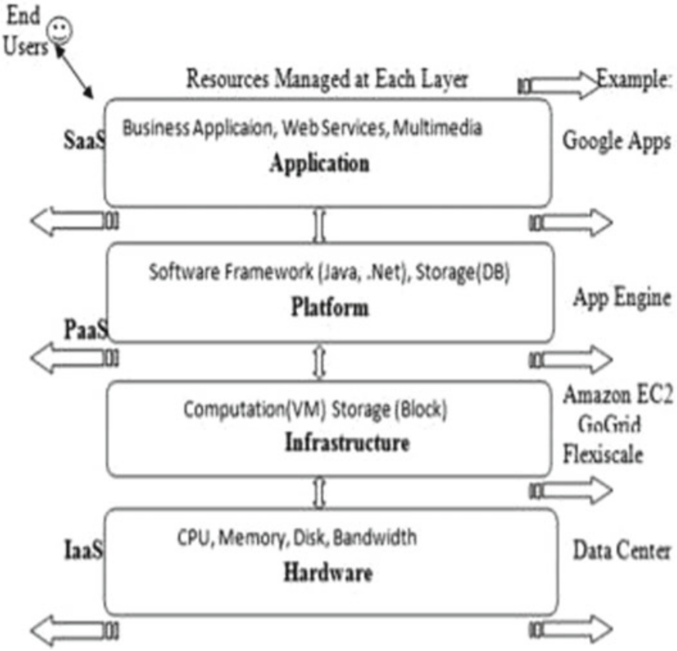
End Users
Resources Managed at Each Layer
Example:
SaaS
Business Applicaion, Web Services, Multimedia
Application
Google Apps
Software Framework (Java, .Net), Storage(DB)
Platform
PaaS
App Engine
Computation(VM) Storage (Block)
Infrastructure
Amazon EC2
GoGrid
Flexiscale
IaaS
CPU, Memory, Disk, Bandwidth
Hardware
Data Center

Fig. 3 Architecture of cloud computing

- **Platform Layer**: This layer is made up of operating system and application framework. It is installed on top of the infrastructure layer. The key principle of the platform layer is CC to reduce overhead for direct deployment in VM containers. E.g., Google App Engine runs on a platform layer to allocate API supports for data storage of various web applications.
- **Application Layer**: It is built on the top level of cloud architecture. It is composed of an actual cloud application. Cloud applications have essential features to achieve better performance, lower operating cost, availability, and scalability.

Thus, this architecture is more modular than other architecture (traditional architecture). Loosely coupled concepts are used in each layer. This architecture allows cloud computing to meet a vast variety of applications and to decrease total costs. No need for the high-power computer to run web-based applications is available in the cloud computing infrastructure.

2.4 Data Center in Cloud Network

Cloud migration platforms are a new and rapid trend. Cloud offers a standardized front-end interface, enabling a large number of applications to be executed on a similar hardware platform. DCs are the backbone of cloud computing development networks. Cloud output depends on the data centers' computing, storage, and network availability [27]. The demands for cloud working load in today's data centers, thus, is reflected in demands for resources.

The data center consists of servers, devices and storage, cooling systems, network devices, power systems, etc. [28]. DCs are for large-scale systems for services like online enterprises, smart grid and mathematical computations. DC is being used to model core services on the cloud infrastructure network. It comprises a group of hosts who handle a group of VM whose roles are to handle "low-level" processing, and a minimum of 1 DC should be set up to commence simulation.

Virtualization technologies [29, 30] are additionally supplied to data centers which allow multiplexed and shared multiple resources between huge no. of users by various time-varying access patterns [31]. DC management is difficult. There are approximately two types of challenge:

(1) *Management of resources,* which focuses on dynamic workloads management provided a resource pool and
(2) *Planning of capacity,* which concentrates on the provision of resources.

DCN (Data Center network) comprises DC also offers data center connections, as defined in its networking topology, the routing and switching equipment as well as the protocols it uses [32]. for the following purposes DCN provides some features to better organize cloud computing [33]:

- DCN allows thousands of data center servers to be efficiently connected such that cloud computing can extend its operation easily by adopting the DCN topology.
- In massive machine-to-machine connectivity, DCN provides traffic reliabilities and efficiencies that generate activities of cloud computing as working loads distributed on servers at data centers.
- DCN embraces different techniques of virtualization that help DCN build virtual machines (VMs), virtual networks, and virtual functions. The DCN should be scalable to isolate and migrate to large numbers of virtual instances.
- Existing DCN research has produced applications in some use cases, including green computing and DC backup, that may also address challenges of cloud infrastructure.

2.5 Virtual Machines in Cloud Computing

Cloud computing includes parallel processing principles and distributed computing to provide shared services via physical server hosting VMs. The service-orientated architecture reduces connectivity costs to collect customer information deliver more flexibility and demand-driven services, etc. The premise behind the generated cloud computing concept is that the processing of information is a public service, which can be achieved more effectively in massive computer farms and storage systems that can be made available worldwide through the Internet. Effective management of

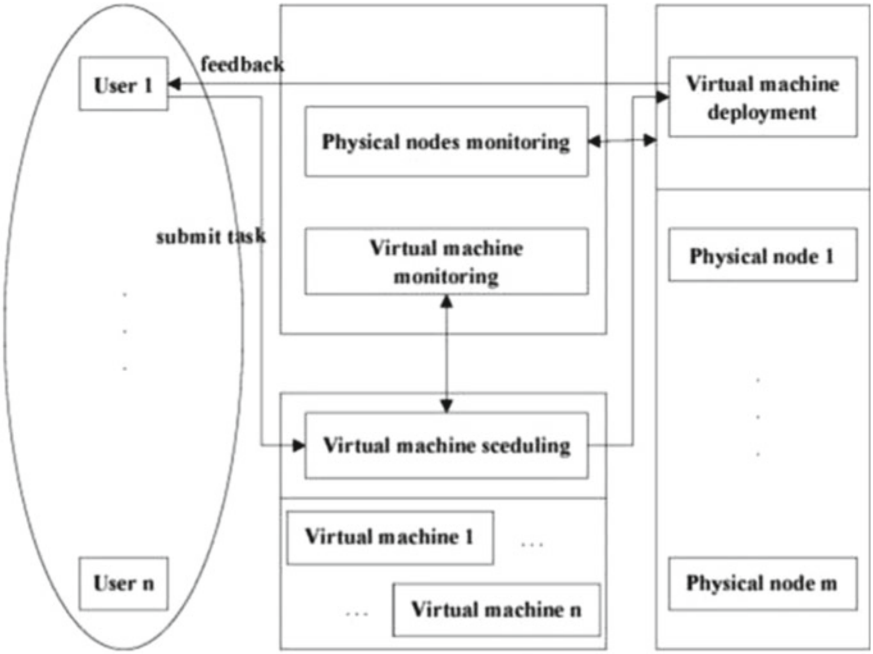
User 1
feedback
submit task
User n
Physical nodes monitoring
Virtual machine monitoring
Virtual machine sceduling
Virtual machine 1
Virtual machine n
Virtual machine deployment
Physical node 1
Physical node m

Fig. 4 Process of VM scheduling [35]

VMs [34] specifically affects the use of the system's resources and QoS. The amount of VM distributing across physical servers seems to be inconsistent over a certain period, given the dynamic nature of the cloud environment. In this case, VMs have to be moved from an overwhelmed server to a load-based server to balance the load.

Figure 4 presents the general cloud data center VM scheduling process. Scheduling can be widely categorized into user applications, control of resources, and scheduling mode.

VM's component handles the allocation of various hosts to various virtual machines so that the computing cores can be allocated (by the host) to VMs. This configuration is based on method, and the default VM assignment policy is 'first-come, first-serve.' Virtualization allows the live migration of VMs [36] with several VMs loaded on several physical machines (PMs) referred to as VMs. An efficient strategy to reduce energy consumption, operational costs, hardware costs, conformance/violation of SLAs, CO_2 emissions also enhance hardware and service reliability, efficiency and hardware life, LB and use of a CCS may be a VM consolidation algorithm. VM consolidation in the cloud environment can essentially reduce energy use and QoS.

2.6 Motivation of Study

The modern cloud computing model offers sophisticated benefits and benefits compared to previous computing paradigms and is adapted, migrated, and adopted by many organizations. Cloud computing has evolved from a prospective logic over the past few years; business is the idea of virtualization to a rapidly increasing sector in the IT sector. Today, recession-hit firms are becoming increasingly aware that they can easily tap into the cloud and quickly access best-of-breed business applications or drastically enhance their infrastructure services at marginal costs. But, some issues, difficulties and effects are still found which scholars, academics and business intelligence (BI) practitioners are presently addressing.

Clouds still present enterprises with security concerns [37] by using cloud computing. Upon details and IT critics behind the firewall, users are still concerned about the susceptibility to attack. Computing in clouds also provides reliability around the clock. There were some cases of outages for a few hours from cloud computing services. Contrary to the conventional computing model, cloud computing uses virtual computing technologies, as users have access to cloud computing services, they can leak hidden information. Attackers will evaluate crucial jobs based on the user's computing task [38]. The growth of cloud computing needs to open standards. The interpretation of most cloud providers with an [39] API, usually well-documented but often special to its execution and thus not interoperable. Heavy transaction-oriented also other applications of data-intensive in which CC can miss sufficient efficiency can be the key to performance. In addition, users far from cloud-based services suffer from high latency and delay. Software and hardware can be saved for companies offering the CC; however, higher latency costs can be incurred. Bandwidth costs may be small for smaller internet applications and not resource-intensive, nonetheless data-intensive applications can rise significantly. Users can be certain that even certain cloud computing services providers can never become irrelevant, or get bought and swallowed up by some bigger company. "Cloud potential providers can retrieve the data and it is imported into replacement application through either format" -Gartner [40]. Cloud service providers are required to comply with legal issues concerning data responsibility and ownership for loss or misuse of data. Legal problems vary from those caused by traditional hosting or outsourcing [41]. These problems and challenges encourage us to focus more on this subject of cloud computing problems.

3 Load Balancing in Cloud Computing

LB is a key problem and issue in cloud environments [42]. The method of allocating and reassigning the load between usable resources is designed to optimize performance, minimize cost and response times, improve efficiency, resource use and save energy. Excellent load balancing strategies can include SLA and user satisfaction.

Therefore, it is a prerequisite to the performance of CC environments to ensure reliable algorithms and processes for load balancing.

LB model is demonstrated in Fig. 5, LB accepts user requests and executes load balancing algorithms to allocate applications to VMs. Load balancer determines that VM's next request must be allocated. The controller of DC is responsible for task management. LB algorithm provides tasks to delegate tasks to appropriate VM. VM manager is responsible for VMs. Summary of LB policies are shown in Table 1.

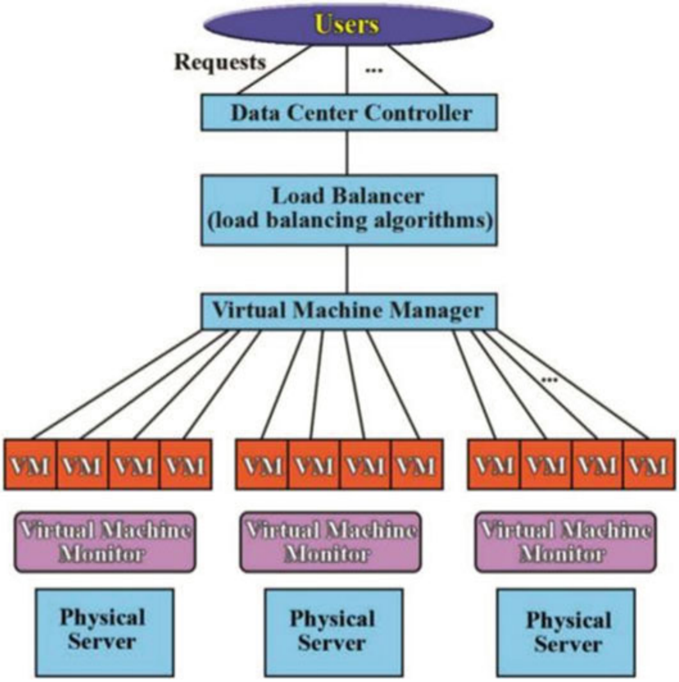

Fig. 5 Load balancing model [43]

Table 1 Summary of LB policies

S. No.	Location policy	Information policy	Selection policy	Transfer policy
1	Find the right partner for transfer	Determine how long information about nodes must be collected	Selecting factors for transferring a task: migration overhead	Includes: task re-scheduling task migration
2	Task monitor availability of resources needed for partner migration	There are three kinds: demand-driven, periodic, and state policies information policy	A no. of remote system calls time of task execution	Depends upon thresholds in terms of load units

Virtualization is leading CC technology. Virtualization aims mostly to share costly hardware between VMs. VM is computer software execution on which operating systems and applications will work. The user requests are processed by VMs. Users are found around the globe and send their requests arbitrarily. For processing, requests must be delegated to VMs. The distribution of tasks is also an important problem in CC. If several VMs are overloaded when others are idle or have some work to do, QoS is reduced. By reducing QoS, consumers are not happy with the system and will never return. The VM is created and managed by a hypervisor or VMM (Virtual Machine Monitor). Multiplexing, suspension (storage), provision (resume) and life migration [44] are four operations of VMM. For load balancing, certain operations are required. In [45] it was specified that load balance would take account of two tasks: allocation of resources and scheduling of tasks. The effect of these 2 tasks is high infrastructure availability, increased use of power, energy savings, reduced cost of resource use, maintaining cloud storage elasticity, and reducing carbon emissions.

3.1 Classification of LB Techniques

This section describes basic testing activities using various approaches to load balancing. The study of the various methods of load balancing for edge computing is explained as follows [46].

Different forms of strategies of LB are seen in Fig. 6. Different types are introduced to provide an improved balance of loads in edge computing, like security, traffic-load based, heterogeneous, optimization-based, joint-load based, heuristic and multi-access based, dynamic load-based, allocation-based, and allocation-based approaches.

3.2 Load Balancing Metrics

The metrics for LB in CC are reviewed in this section and summarized as follows [47]:

- **Response time:** It calculates the overall time required to fulfill a submitted task by the system.
- **Throughput:** This measurement is utilized to quantify the number of processes per unit time performed.
- **Scalability:** An algorithm is capable of uniform LB in the system as per demands, as no. of nodes increases. It is a highly scalable preference algorithm.
- **Makespan:** This metric is utilized to measure maximum completion time or the amount of time a user has attributed the tools.
- **Migration time:** time to switch tasks from overloaded node to undercharged node.

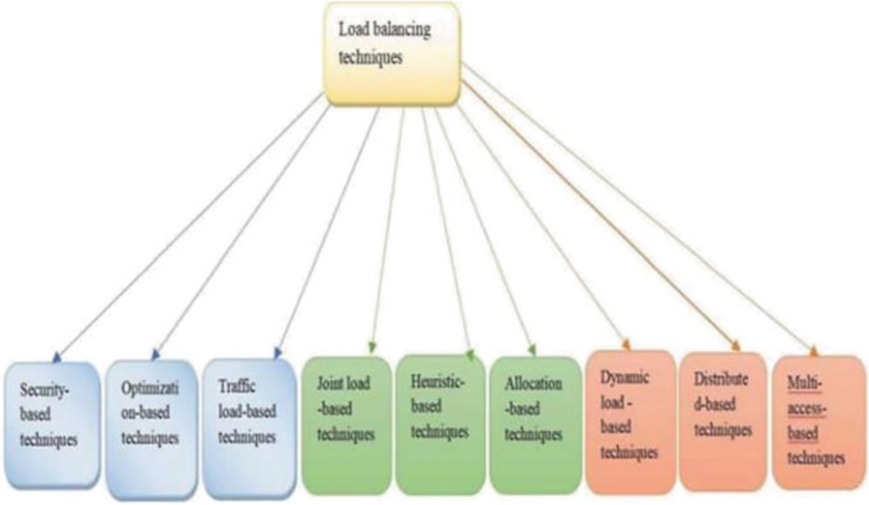
Load balancing techniques
Security-based techniques
Optimization-based techniques
Traffic load-based techniques
Joint load -based techniques
Heuristic-based techniques
Allocation -based techniques
Dynamic load - based techniques
Distributed-based techniques
Multi-access-based techniques

Fig. 6 Categorization of LB techniques in CC

- **Fault tolerance:** It decides the algorithm's capacity to perform LB in event of such nodes or ties breakdown.
- **Performance:** After the LB algorithm, it tests system efficiency.
- **Degree of imbalance:** This measurement the imbalance of VMs.
- **Carbon emission:** The amount of carbon that all materials generated is calculated. Load balancing played an important part in reducing this metric by switching and shutting loads from underloaded nodes.
- **Energy consumption:** It determines how much energy all nodes consume. Load balancing prevents overheating and thereby reduces energy consumption by LB across all nodes.

4 Task Scheduling in Cloud Environment

Cloud computing handles a range of virtualized capabilities that make planning an important component. A customer can use thousands of virtualized properties in the cloud for each task [48]. Manual scheduling is therefore not a workable alternative. The underlying concept behind scheduling consists of distributing tasks (complex and diverse nature) between cloud resources in a manner that reduces time losses and maximizes efficiency by scheduling algorithms. Several study activities have in the past looked at task planning. Cloud operating resources are tracked and loads are computed for any resource before assigning workload or tasks on VMs (resources). If any VM is in over-used mode, a task for those types of resources is not assigned.

The task is a workpiece to be carried out within a given time frame. Inside the cloud, the task is in two forms [49]: a task that is independent and dependent. Task scheduling is the mechanism by which the services are assigned to this task for a certain amount of time. TS is a very common theme in the field of CC. It is used to plan tasks to improve the usage of resources by the allocation of such tasks to definite resources in particular. The main objective of the task algorithm is to increase reliability, enhance service quality, maintain productivity across tasks, and reduce costs. In the TS process, virtual tools are used to their maximum potential. For high performance, effective resource scheduling is essential. The completion period and the cost of completing the task are two different criteria.

4.1 Types of Task Scheduling

Categories of TS are as follows [50]:

- User-Level Scheduling
- Cloud Service Scheduling
- Heuristic Scheduling
- Static and Dynamic Scheduling
- Workflow Scheduling
- Real-Time Scheduling.

4.2 Framework of TS in CC Environment

There is a framework [51] for the TS system in CC. In Fig. 7, users are accessing the cloud environment through the internet. The cloud part shows how the cloud is managing to serve various requests given by the consumers.

(1) **SLA Monitor:** First, a client sends the service request. SLA monitor then reviews the requested QoS application before deciding whether the submission should be accepted or denied. SLA Monitor is responsible for monitoring the success of the job submitted and disciplinary measures must be taken promptly if there are any violations.
(2) **Resource Discovery and Monitoring:** Resource search can essentially be defined as the job of the manufacturer to locate the right resources to satisfy incoming customer demands. Cloud computing's main advantages are the ability to receive and distribute data on demand, so the management of resources must be continuous.
(3) **Task Scheduling**: The input of the algorithm for task scheduling is usually an abstract model. This abstract model describes tasks without defining were tools to perform the tasks are physically placed.
(4) **Reschedule:** If a task cannot be done due to a processor malfunction or other problem, the incomplete task in the next calculation will be rescheduled.

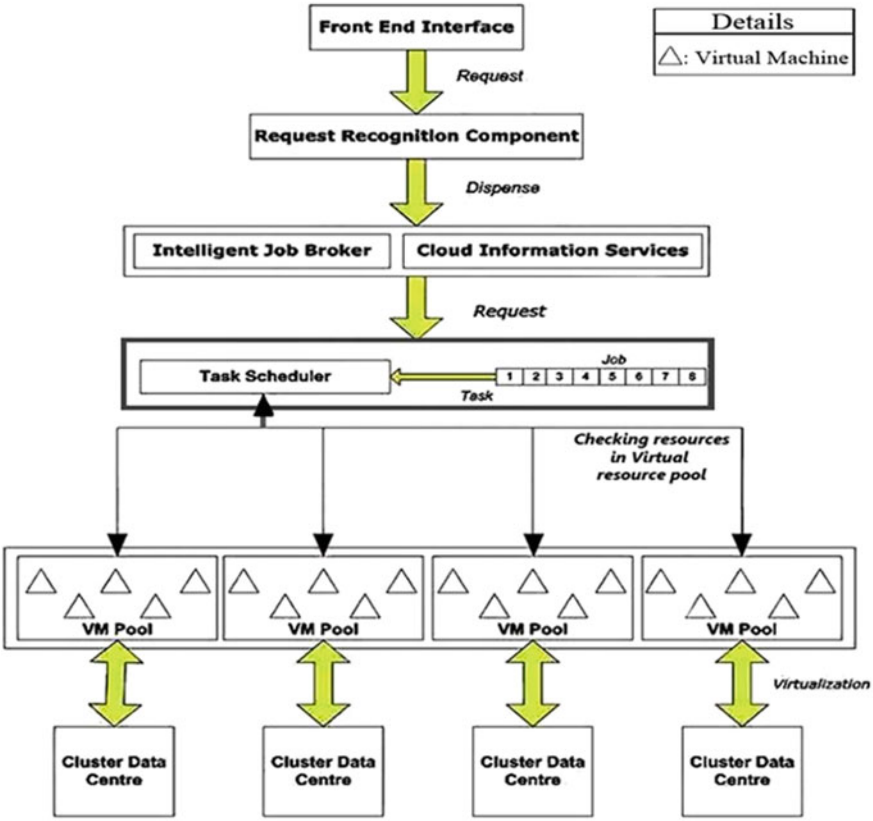

Fig. 7 Framework of TS in CC environment

(5) **Scheduling Optimizer**: Several relevant candidates are illuminated after the information is obtained about the available resources in the cloud. The framework for resource selection chooses a solution that meets all demands and optimizes infrastructure usage. The allocation of resources can be made using an algorithm for optimization. Various optimization methods can be used by common and well-known techniques, including simple metaheuristic algorithms. Examples include GA, ACO and PSO for the cloud.

(6) **Advanced Resource Reservation Monitor**: The Advanced Resource Reservation Control ensures QoS connections in the data center to different services. In the future, users will protect the necessary services. This is necessary if some processors need to be run to complete critical time applications such as real-time and Workflow applications for parallel applications. Providers can more reliably forecast potential demand and use.

(7) **Data Centers and VMs**: Data centers are the infrastructure or hardware part where all the physical servers present. Physical servers are not used as viable

users. Through virtualization, they are converted into different VMs and users' jobs are going to run on these VMs in scheduling tasks.VM manager manages the VMs.

5 Meta-Heuristic Optimization Techniques

Due to the usefulness of these metaheuristic algorithms in solving massive computational and complex problems [52] has been enormously common in the past 20 years. There are some useful metaheuristic characteristics such as:

(1) These algorithms are not unique to problems.
(2) Effectively explore a meta-heuristic algorithm in search areas to find (near) optimal solution or sub-optimal solution of NP-complete problems.
(3) Approximate and typically non-deterministic meta-heuristic algorithms.

Metaheuristic algorithms are problem-independent and can be used to solve issues in a variety of domains. Metaheuristic approaches are standard techniques for solving NP-hard optimization problems.

$$\text{Meta-heuristic} = \text{Heuristic} + \text{Randomization}.$$

Many meta-heuristic algorithms are available in the cloud environment to find an approximate (suboptimal) solution to an NP-Complete problem in a short amount of time. TS is one of the problems due to the vast space for a solution which takes a long time to find an optimal solution. Diverse choice of meta-heuristic algorithms like PSO, ACO, GA, Artificial Bee Colony (ABC), Simulated Annealing (SA), Differential evolution algorithm (DEA), BAT optimization, Bacteria Foraging Optimization, Firefly optimization, Cat swarm optimization, Lion optimization, Cuckoo search, etc.

We also addressed and analyzed numerous task scheduling algorithms in the cloud environment with a meta-heuristic method and their benefits and drawbacks (demonstrated in Table 2). In evaluating precision solutions to a specific problem, the right choice of the optimization algorithm is greatly helpful. At first, we will analyze the PSO algorithm since it is easier to converge the PSO algorithm and the complexity than other metaheuristic algorithms.

6 Literature Survey

1. Oliveira et al. [3] Recently, a great many projects have carried out extensive cloud research, including load balancing and task scheduling. These works illustrate some of the core problems facing state-of-the-art load balancing

Table 2 Summary of the tested meta-heuristic algorithms, including their benefits and drawbacks

S. No.	Author (Year)	Algorithm	Nature of tasks	Techniques	Benefits	Drawbacks
1	Elaziz et al.	Hybrid algorithm	Independent	MSDE algorithm (MSA with DE)	Enhance makespan system time and throughput	A single objective (focused only time) is the algorithm performed and does not take account of other QoS parameters such as reliability, cost, energy
2	Adhikari et al.	BAT	Independent	LB-RC algo via BAT	Enhance cost, makespan time, resource utilization, reliability, etc	Algorithm doesn't discuss compromise among time and cost QoS parameter
3	Sharma and Kumar PSO-COGEN	PSO-COGENT algorithm	Independent	Modified PSO depends on APSO-VI	Enhance time, throughput, energy consumption, QoS, cost parameters, etc	QoS parameters such as availability, reliability, etc. cannot be taken into consideration. VMs can be overloaded when applications of VMs have been improperly described
4	Alaaeldin and Almezeini	Lion algorithm	Independent	TS by lion	Enhance cost, makespan time, resource usage and imbalance degree	Algorithm considers no limit and did not address the compromise solution

(continued)

Table 2 (continued)

S. No.	Author (Year)	Algorithm	Nature of tasks	Techniques	Benefits	Drawbacks
5	Addya et al.	SA	Independent	MVMP algorithm via SA	enhances, energy consumed, profit and execution time	The solution to static VM migration takes time
6	Zamanifar and Rastkhadiv	ABC	Independent	Agile task handling method via ABC	Enhance makespan time and degree of imbalance	Algorithms are only used to carry out independent tasks and main performance parameters such as reliability, energy, and cost, etc. are not considered, where the deadline is a restraint
7	Lin et al.	Binary PSO	NA	Modified sigmoid transfer function	Enhance time of High utility itemset mining	The sigmoid function proposed would not account for the last step of the operation
8	Pacini et al.	ACO	Independent	2 Level cloud scheduler method by ACO algorithm	Enhance response time, throughput and makespan time	Algorithms do not take into account the time and tasks priority; indirect communication mechanisms are utilized to communicate information (pheromone)

(continued)

Table 2 (continued)

S. No.	Author (Year)	Algorithm	Nature of tasks	Techniques	Benefits	Drawbacks
9	Kaushal and Verma	PSO based scheduling	Workflow applications	Bi-objective priority-based PSO	Reduce execution costs and time at the same time	Energy usage and other effective QoS are not taken into account
10	Ramezani and Khadeerhussain	PSO based scheduling algorithm	Independent	task migration, Cloud resource broker method	Enhance rejection ratio time, expense and task	Gbest is affected by local minima and energy utilization is not taken into account
11	Jana and Kulia	TBSLB- PSO	Independent	Task migration from overloaded VMs to under-loaded VMs	Reduce task transfer time and makespan time	Performance of algorithm suggested is not related by other state-of-the-art algorithms
12	Krishna and Babu	ABC	Independent	Task migration based LB using ABC	Time of response, runtime and performance, number of migrated tasks, throughput	Only non-preemptive tasks and deadline of tasks are not considered for proposed algorithm
13	Huo and Zhan	Hybrid PSO	Independent	SA with PSO is added	Enhance algorithm search capabilities with SA	After adding SA and PSO the complexity of the algorithm is improved

applications. Several task scheduling mechanisms have been defined by many researchers for load balancing.

2. Jena [53], focused on TSPSO (task scheduling using multi-objective nested particle swarm optimization) to enhance energy and processing time. Outcome from TSPSO was simulated through cloud platform for open source applications (CloudSim). The findings were finally related to the existing scheduling algorithms and found an optimum balance outcome for different goals with the proposed algorithm (TSPSO) [54].
3. Wang and Zhou [55], Proposed a technique for solving the MapReduce load imbalance issue created by the use of the default partition algorithm of the Hadoop platform. This suggested approach will optimize the activities and balance inputs of a decrease process in the map phase by using multiple partition techniques. Moreover, this suggested technique will use idle nodes to completely offset high load nodes, to attain optimum work scheduling throughout the reduction phase execution method [56].
4. Velde and Rama [57], In this cloud partitioning strategy, LB and resource optimization are effectively extended. Besides the optimal time of refresh that sets the state adjustments of the data center partitions for efficient use of resources. CloudSim is a simulation platform utilized to create prototype applications to prove hypothesis proof. Results have shown that this approach increases cloud resource efficiency, performance, and optimization [53].
5. Basha and Padmavathi [58], proposed an innovative, dynamic, and elasticity algorithm perform LB by ACO to perform load Balance between Systems existing in DCs [55].
6. Vijayakumar and Kanthimathi [59], The suggested scheme used additional virtual machines using genetic methods to address the requests of the highest virtual machines. The allocation of the best virtual machines will deal with demands very efficiently and quickly. During execution, the load might be balanced via the ACO technique if those VMs were overloaded by requirement. The above strategy will share the extra charge with other virtual machines, gently loaded and idle. On the other hand, after their work completion or in idleness total energy consumption is enhanced by turning away VMs [57].
7. (MPSO) hybridization and an improved Q-learning algorithm called QMPSO. Hybridization is done to change MPSO velocity by pbest and gbest depends upon the best action provided by improved Q-learning. The objective of hybridization is to improve machine performance by balancing the charge among VMs, optimize VM's output and maintain the balance between task priorities by maximizing the time of work. By comparing QMPSO results obtained during the simulation process with the existing LB and scheduling algorithm, the robustness of this algorithm is evaluated. The simulation compared with real results of the platform reveals that the proposed algorithm outperforms its competitor [60].
8. Malik and Bansal [61], An optimized approach to scheduling model and resource cost scheduling model labeled MultiFaceted Optimization Scheduling Framework (MFOSF) was timely in this study to resolve the restrictions and

change solution quality. The resource Cost model shows the relation between customer budget and production costs during scheduling. PSO can be used to achieve a model focused on optimizing efficiency and cost. There have been several simulations to test this approach using 4 different metrics (a) Makespan (b) Cost (c) Resource Utilization (d) Deadline. Based on the above parameters, experimental results demonstrate that the MFOSF-PSO approach was better than other models and in the best-case scenario increased by 57.4% [61].

9. Devaraj et al. [62], As a firefly hybrid and Improved Multi-objective Particle Swarm Optimization (IMPSO) technology, abbreviated as FIMPSO, the latest LB algorithm is proposed for use. This technique uses FF (Firefly) algorithm to limit search space to detect the improved response using the IMPSO technique. The proposed FIMPSO algorithm generated and improved effective average load for crucial measures, including proper use of resources and response time for tasks. The simulation result has shown the efficient performance of the FIMPSO model suggested compared to other methods. From the resulting simulation, the FIMPSO algorithm is understood to have generated a successful outcome with an average response time of 13.58 ms maximum, overall CPU usage of 98%, 93% memory utilization, 67% reliability and 72% throughput, along with 148, which were superior to all other techniques compared [62].
10. Miao et al. [63], The new algorithm called adaptive Pbest Discrete PSO (APDPSO) for PSO-based static load balancing was proposed to combat this problem. Good solution held in the external archive is used to upgrade the particles' personal best positions and to update velocity and direction vectors of particles by proposing a probability and comparability approach for PSO. MATLAB and CloudSim systems perform simulation experiments with random synthetic tasks. The findings revealed that our proposed algorithm dramatically increased swarm convergence and diversity and reduced the level of load imbalance relative to state art in the field [63].

6.1 Research Gaps and Issues

This section addresses the study deficiencies of cloud computing that need to be addressed in terms of existing load-balancing techniques.

- Lightweight security solutions were not considered for increased LB efficiency [64].
- Method [65] failed to implement the paradigm built-in CC and real field of joint edge computing. The procedure failed to increase QoE and reduce incoming traffic based on the adaptive bitrate approach for mobile video streaming and optimized edge caching [66].
- The cross-layer approach on heterogeneous networks was more complicated [67].
- The approach in [68] neglected to take into account MEC server security levels, which is very important to industry data.

- In [69], the complexity was not properly calculated and the reliability was not improved. In [70] the strategy to balance the positive and negative impacts on the migration of services has not been established.
- The method in [71] needed additional major structures for orchestration to improve the system performance.
- Performance has been improved, but the allocation of vehicles was highly imbalanced, leading to inefficient QoS for a vehicle to infrastructure (I2V) communication [72].
- However, the forecast was inaccurate because the task length was smaller and the training set was not included [73]. Large-scale heterogeneous MEC problems for optimizing training speed were not taken into consideration [74].
- In [75], live video and car navigation services could not be used. No further data sets were considered to improve the system's performance [76].
- The load-aware user associations that contain dynamic settings were not taken into account [77]. The transmitting power and rates of SCBSs were not taken into account in [78] to achieve minimum energy consumption.
- Because of many interferences and noises, the communication channel was not perfect [79]. They are normally unable to understand the degradation of QoE [80]. The process of the reorder did not however alter the efficiency of the load balancing [81].

The challenges of these approaches are taken into account as a reason for developing a new load-balancing method.

7 Conclusion

AC is the product of the Virtualization concept-enabled grid computing. LB is a challenge in the cloud environment, as it is a complex task to share the task equally between VMs. LB is the main CC mechanism that prevents overloading nodes. Load balancing stabilizes service quality (QoS), covering response time, cost, throughput, efficiency, and the use of resources. At the height of time, servers cannot manage incoming requests with no. of VMs available, meaning that some additional virtual machines needed to be run without fault or delay. The benefit of switching to a virtual environment is important, but there are significant efficiency and cost optimization issues in cloud computing due to different types of need for resources and multiple tasks. The success standard and requirements of SLA to be retained therefore become very difficult because of these constraints. Various alternatives to the planning model and resource cost timeline model have been applied to solve these limitations and to change solution quality. Different meta-heuristic optimization strategies were applied to conduct TS in the CC environment to manage demand on virtual machines.

This survey will offer an idea of the newest developments in LB in cloud systems and will provide an analysis of the technological advances, company benefits and growing use of data centers, and future trends. Amazon EC2, Microsoft Azure, and

Google App Engine are compared. This article highlights research gaps and load handling problems in CC andsummarizes the superiority of CC. The future analysis is to find effective methods for calculating the cloud results.

References

1. Academic paper http://www.mcs.csueastbay.edu/~lertaul/Cloud%20Security%20CamREADY.pdf
2. Chen G, Lu J, Huang J, Wu Z (2010) SaaAS—the mobile agent-based service for cloud computing in internet environment. In: Sixth international conference on natural computation, ICNC 2010. IEEE, Yantai, Shandong, China, 2010, pp 2935–2939. ISBN: 978-1-4244-5958-2
3. Oliveira D, Baião F, Mattoso M (2010) Towards taxonomy for cloud computing from an e-science perspective. In: Cloud computing: principles, systems, and applications (to be published). Springer, Heidelberg
4. Singh G, Sood S, Sharma A (2011) CM-Measurement facets for cloud performance. IJCA 23(3). Lecturer, Computer Science and Engineering, Eternal University, Baru Sahib (India)
5. Ertaul L, Singhal S (2009) Security challenges in cloud computing. California State University, East Bay
6. Global Netoptex Incorporated.—Demystifying the cloud. Important opportunities, crucial choices, pp 4–14. Available: http://www.gni.com. 13 Dec 2009
7. Gulshan S, Kalra M (2014) A novel approach for load balancing in cloud data center. 978-1-4799-2572-8/14/$31.00 c_2014. IEEE
8. Wu H, Ding Y, Winer C, Yao L (2010) network security for virtual machines in cloud computing. In: 5th International conference on computer sciences and convergence information technology, pp 18–21, Seoul, Nov 30–Dec 2, 2010. ISBN: 978-1-4244-8567-3
9. Kliazovich D, Arzo ST, Granelli F, Bouvry P, Khan SU (2013) eSTAB: energy-efficient scheduling for cloud computing applications with traffic load balancing. In: Green computing and communications (GreenCom). IEEE, pp 7–13
10. Greenberg A, Hamilton J, Maltz DA, Patel P (2008) The cost of a cloud: research problems in data center networks. ACM SIGCOMM Comput Commun Rev 39(1):68–73
11. Lu X, Kong F, Yin J, Liu X, Yu H, Fan G (2015) Geographical job scheduling in data centers with heterogeneous demands and servers. In: 2015 IEEE 8th international conference on cloud computing, pp 413–420. https://doi.org/10.1109/CLOUD.2015.62
12. Armbrust M, Fox A, Griffith R, Joseph AD, Katz R, Konwinski A, Lee G, Patterson D, Rabkin A, Stoica I et al (2010) A view of cloud computing. Commun ACM 53(4):50–58
13. Wang W, Li B, Liang B (2013) Dominant resource fairness in cloud computing systems with heterogeneous servers. arXiv preprint arXiv:1308.0083
14. Shafiq DA, Jhanjhi NZ, Azween Abdullah (2021) Load balancing techniques in cloud computing environment: a review. J King Saud Univ Comput Inf Sci. https://doi.org/10.1016/j.jksuci.2021.02.007
15. Khan S, Sharma N (2014) Effective scheduling algorithm for load balancing (SALB) using Ant Colony
16. Optimization in cloud computing. Int J Adv Res Comput Sci Softw Eng 4(2)
17. Kushwah P (2014) A survey on load balancing techniques using ACO algorithm. (IJCSIT) Int J Comput Sci Inf Technol 5(5):6310–6314
18. Farrag AAS, Mahmoud SA, EI Sayed M, EI- Horbaty (2015) Intelligent Cloud Algorithms for Load Balancing problems: a survey. In: 2015 IEEE seventh international conference on intelligent computing and information systems (ICICIS ‘2015)
19. Houssein EH, Gad AG, Wazery YM, Suganthan PM (2021) Task scheduling in cloud computing based on meta-heuristics: review, taxonomy, open challenges, and future trends. Swarm Evol Comput 62:100841. https://doi.org/10.1016/j.swevo.2021.100841

20. Chen Y, Paxson V, Katz RH (2010) What's new about cloud computing security. In: University of California, Berkeley Report No. UCB/EECS-2010-5, January 2010
21. George SS, Suji Pramila R (2021) A review of different techniques in cloud computing. Mater Today Proc. https://doi.org/10.1016/j.matpr.2021.02.748
22. Radarnetworks and Novaspivak. http://radarnetworks.com
23. Odun-Ayo I, Ananya M, Agono F, Goddy-Worlu R (2018) Cloud computing architecture: a critical analysis. In: 2018 18th international conference on computational science and applications (ICCSA), 2018, pp 1–7. https://doi.org/10.1109/ICCSA.2018.8439638
24. Prasad MR, Lakshman Naik R, Bapuji V (2013) Cloud computing: research issues and implications. Int J Cloud Comput Serv Sci (IJ-CLOSER) 2(2):134–140
25. Verma, Kaushal S (2011) Cloud computing security issues and challenges: a survey. (July):445–454
26. Tiwari K, Chaudhary S, Shanu K (2015) Survey paper on cloud Computing. In: International conference on emerging trends in technology, science and upcoming research in computer science, Apr 2015, pp 1777–1782
27. Zhang Q, Cheng L, Boutaba R (2010) Cloud computing: state-of-the-art and research challenges. J Internet Serv Appl I:7–18
28. Joshi S, Kumari U (2016) Cloud computing: architecture and challenges. Mody Univ Int J Comput Eng Res 1(1):56–60
29. Birke R, Chen LY, Smirni E (2012) Data centers in the cloud: a large-scale performance study. In: 2012 IEEE Fifth international conference on cloud computing, pp 336–343. https://doi.org/10.1109/CLOUD.2012.87
30. Abts D, Felderman B (2012) A guided tour through data-center networking. Queue 10(5):10:10–10:23
31. Wood T, Tarasuk-Levin G, Shenoy PJ, Desnoyers P, Cecchet E, Corner MD (2009) Memory buddies: exploiting page sharing for smart colocation in virtualized data centers. In: VEE, 2009, pp 31–40
32. Mehta S, Neogi A (2008) ReCon: a tool to recommend dynamic server consolidation in multi-cluster data centers, In: NOMS, 2008, pp 363–370
33. Chen Y, Das A, Qin W, Sivasubramaniam A, Wang Q, Gautam N (2005) Managing server energy and operational costs in hosting centers. In: SIGMETRICS, 2005, pp 303–314
34. Wanga B, Qi Z, Maa R, Guana H, Vasilakos AV (2015) A survey on data center networking for cloud computing. Comput Netw 91:528–547
35. Wang A, Iyer M, Dutta R, Rouskas GN, Baldine I (2013) Network virtualization: technologies, perspectives, and frontiers. J Lightwave Technol 31(4):523–537
36. Sahu Y, Agrawal N (2015) A survey paper: cloud computing and virtual machine migration. IJCSN J 4(4):577–581
37. Liu L, Qiu Z (2016) A survey on virtual machine scheduling in cloud computing. In: 2016 2nd IEEE international conference on computer and communications (ICCC), 2016, pp 2717–2721. https://doi.org/10.1109/CompComm.2016.7925192
38. Zolfaghari R, Sahafi A, Rahmani AM, Rezaei R (2021) Application of virtual machine consolidation in cloud computing systems. Sustaine Comput Inf Syst 30:100524. https://doi.org/10.1016/j.suscom.2021.100524
39. Mills E (2009) Cloud computing security forecast: clear skies
40. Jiang J, Wen W (2010) Information security issues in cloud computing environment. Netinfo Secur. https://doi.org/10.3969/j.issn.1671-1122.2010.02.026
41. Clark C, Fraser K, Hand S, Hansen JG, Jul E, Limpach C, Pratt I, Warfield A (2005) Live migration of virtual machines. In: Proceedings of NSDI'05. Berkeley CA, USA, 2005. USENIX Association, pp 273–286
42. Gartner, Seven cloud-computing security risks. http://www.infoworld.com. 02 July 2008
43. Prasad MR, Gyani J, Murti PRK (2012) Mobile cloud computing implications and challenges. IISTE J Inf Eng Appl (JIEA) 2(7):7–15. http://iiste.org
44. Jadeja Y, Modi K (2012) Cloud computing—concepts, architecture and challenges. In: International conference on computing, electronics and electrical technologies [ICCEET]

45. Gupta H, Sahu K (2014) Honey bee behavior based load balancing of tasks in cloud computing. Int J Sci Res 3(6)
46. Hwang K, Dongarra J, Fox GC (2013) Distributed and cloud computing: from parallel processing to the Internet of Things
47. Ivanisenko IN, Radivilova TA (2015) Survey of major load- balancing algorithms in distributed system. In: Information technologies in innovation business conference (ITIB)
48. Pydi H, Iyer GN (2020) Analytical review and study on load balancing in edge computing platform. 2020 Fourth international conference on computing methodologies and communication (ICCMC), 2020, pp 180–187. https://doi.org/10.1109/ICCMC48092.2020.ICCMC-00036
49. Jafarnejad Ghomi E, Masoud Rahmani A, Nasih Qader N (2017) Load-balancing algorithms in cloud computing: a survey. J Netw Comput Appl 88:50–71. https://doi.org/10.1016/j.jnca.2017.04.007
50. Arunarani AR, Manjula D, Sugumaran V (2018) Task scheduling techniques in cloud computing: a literature survey. Future Gener Comput Syst. https://doi.org/10.1016/j.future.2018.09.014
51. Abdul Qadir OS, Ravi G (2020) A survey on task scheduling algorithms in cloud computing. Int J Innovations Eng Technol 15(4):29–35. https://doi.org/10.21172/ijiet.154.06
52. Sharma P, Shilakari S, Chourasia U, Dixit P, Pandey A (2020) A survey on various types of task scheduling algorithm in cloud computing environment. Int J Sci Technol Res 9(01):1513–1521
53. Jena RK (2015) Multi objective task scheduling in cloud environment using nested PSO framework. Procedia Comput Sci 57:1219–1227. https://doi.org/10.1016/j.procs.2015.07.419
54. Venu G, Vijayanand KS (2020) Task scheduling in cloud computing: a survey. Int J Res Appl Sci Eng Technol (IJRASET) 8(V):2258–2266
55. Wang S, Zhou H (2016) The research of MapReduce load balancing based on multiple partition algorithm. In: 2016 IEEE/ACM 9th international conference on utility and cloud computing (UCC), pp 339–342
56. Kumar M, Sharma SC, Goel A, Singh SP (2019) A comprehensive survey for scheduling techniques in cloud computing. J Netw Comput Appl. https://doi.org/10.1016/j.jnca.2019.06.006
57. Velde V, Rama B (2017) Simulation of optimized load balancing and user job scheduling using CloudSim. In: 2017 2nd IEEE international conference on recent trends in electronics, information and communication technology (RTEICT), pp 1379–1384. https://doi.org/10.1109/RTEICT.2017.8256824
58. Padmavathi M, Basha SM (2017) Dynamic and elasticity ACO load balancing algorithm for cloud computing. In: 2017 International conference on intelligent computing and control systems (ICICCS), pp 77–81. https://doi.org/10.1109/ICCONS.2017.8250571
59. Kanthimathi M, Vijayakumar D (2018) An enhanced approach of genetic and ant colony based load balancing in cloud environment. In: 2018 International conference on soft-computing and network security (ICSNS), pp 1–5. https://doi.org/10.1109/ICSNS.2018.8573608
60. Kumar KP (2018) gravitational emulation-grey wolf optimization technique for load balancing in cloud computing. In: 2018 Second international conference on green computing and Internet of Things (ICGCIoT), pp 177–184. https://doi.org/10.1109/ICGCIoT.2018.8753108
61. Bansal M, Malik SK (2020) A multi-faceted optimization scheduling framework based on the particle swarm optimization algorithm in cloud computing. Sustain Comput Inf Syst 28:100429. https://doi.org/10.1016/j.suscom.2020.100429
62. Devaraj AFS, Elhoseny M, Dhanasekaran S, Laxmi Lydia E, Shankar K (2020) Hybridization of firefly and improved multi-objective particle swarm optimization algorithm for energy efficient load balancing in cloud computing environments. J Parallel Distrib Comput 142:36–45. https://doi.org/10.1016/j.jpdc.2020.03.022
63. Miao Z, Yong P, Mei Y, Quanjun Y, Xu X (2021) A discrete PSO-based static load balancing algorithm for distributed simulations in a cloud environment. Future Gener Comput Syst 115:497–516. https://doi.org/10.1016/j.future.2020.09.016
64. Li J, Luo G, Cheng N, Yuan Q, Wu Z, Gao S, Liu Z (2018) An end-to-end load balancer based on deep learning for vehicular network traffic control. IEEE Internet of Things J 6(1):953–966

65. Dong Y, Xu G, Ding Y, Meng X, Zhao J (2019) A 'joint-me'task deployment strategy for load balancing in edge computing. IEEE Access 7:99658–99669
66. Liu J, Shou G, Liu Y, Hu Y, Guo Z (2018) Performance evaluation of integrated multi-access edge computing and fiber-wireless access networks. IEEE Access 6:30269–30279
67. Liu L, Chan S, Han G, Guizani M, Bandai M (2018) Performance modeling of representative load sharing schemes for clustered servers in multiaccess edge computing. IEEE Internet of Things J 6(3):4880–4888
68. Niu X, Shao S, Xin C, Zhou J, Guo S, Chen X, Qi F (2019) Workload allocation mechanism for minimum service delay in edge computing-based power Internet of Things. IEEE Access 7:83771–83784
69. Xu X, Fu S, Cai Q, Tian W, Liu W, Dou W, Sun X, Liu AX (2018) Dynamic resource allocation for load balancing in fog environment. Wirel Commun Mobile Comput 2018
70. Fahs A, Pierre G (2019) Proximity-aware traffic routing in distributed fog computing platforms
71. Lee H, Kwon B, Kim S, Lee I, Lee S (2015) Theoretical-analysis-based distributed load balancing over dynamic overlay clustering. IEEE Trans Veh Technol 65(8):6532–6546
72. Fernando N, Loke SW, Rahayu W (2016) Computing with nearby mobile devices: a work sharing algorithm for mobile edge-clouds. IEEE Trans Cloud Comput (2016)
73. Lu H, Gu C, Luo F, Ding W, Liu X (2020) Optimization of lightweight task offloading strategy for mobile edge computing based on deep reinforcement learning. Future Gener Comput Syst 102:847–861
74. Fan Q, Ansari N (2018) Towards traffic load balancing in drone-assisted communications for IoT. IEEE Internet Things J 6(2):3633–3640
75. Han T, Li S, Zhong Y, Bai Z, Kwak K-S (2019) 5G software-defined heterogeneous networks with cooperation and partial connectivity. IEEE Access 7:72577–72590
76. Li C, Wang YaPing, Tang H, Zhang Y, Xin Y, Luo Y (2019) Flexible replica placement for enhancing the availability in edge computing environment. Comput Commun 146:1–14
77. Li C, Sun H, Chen Y, Luo Y (2019) Edge cloud resource expansion and shrinkage based on workload for minimizing the cost. Future Gener Comput Syst 101:327–340
78. Asif-Ur-Rahman M, Afsana F, Mahmud M, Shamim Kaiser M, Ahmed MR, Kaiwartya O, James-Taylor A (2018) Toward a heterogeneous mist, fog, and cloud-based framework for the internet of healthcare things. IEEE Internet of Things J 6(3):4049–4062
79. Bulkan U, Dagiuklas T, Iqbal M, Huq KMS, Al-Dulaimi A, Rodriguez J (2018) On the load balancing of edge computing resources for on-line video delivery. IEEE Access 6:73916–73927
80. Ramaswamy L, Liu L, Iyengar A (2007) Scalable delivery of dynamic content using a cooperative edge cache grid. IEEE Trans Knowl Data Eng 19(5):614–630
81. Wan J, Chen B, Wang S, Xia M, Li D, Liu C (2018) Fog computing for energy-aware load balancing and scheduling in smart factory. IEEE Trans Ind Inf 14(10):4548–4556