A Systematic Literature Review on Contemporary and Future trends in Virtual Machine Scheduling Techniques in Cloud and Multi-Access Computing

Nadim Rana (nadimrana@jazanu.edu.sa)
Jazan University
Fathe Jeribi
Jazan University
Sherif Tawfik Amin
Jazan University
Zeba Khan
Jazan University
Mueen Uddin
University of Doha for Science and Technology
Imed Ben Dhaou
Dar Al-Hekma University

Research Article

Keywords: Cloud computing, Virtualization, SLA, Virtual Machine Scheduling, QoS, Internet of Things, Multi-access computing

Posted Date: April 12th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2792348/v1

Abstract

Due to the extensive migration of business and scientific applications as well as the enormous growth in online data produced by IoT devices, numerous problems have arisen in cloud scheduling. Efficient delivery of resources considering user-defined Service Level Agreement (SLA) and Quality of Service (QoS) can only achieve with efficient and state-of-the-art scheduling methods. In this regard, virtual machine (VM) scheduling has been a highly required method for resource scheduling in the ever-changing cloud and multi-access computing environment (MAC). Based on an examination of recent literature, this investigation intends to provide a comprehensive Systematic Literature Review (SLR) of the methods employed for virtual machine scheduling in cloud computing. Besides, the SLR disseminates the challenges and opportunities in VM design and discusses future researchers' baselines. The SLR investigated the VM scheduling techniques and searched the most relevant research databases online. The authors selected sixty-seven (67) preliminary studies for this review out of 722 articles between 2008 and 2022. A total of 67 articles were reviewed for VM scheduling methods and techniques. The taxonomical results were divided into three major classes; conventional approach, heuristics approach, and meta-heuristic approach. With the observation, this review concludes that a lot of development in VM scheduling techniques in the literature are based on metaheuristics and heuristics methods. At last, many open issues, challenges, and development trends of modern VM scheduling techniques are discussed.

1 Introduction

As a consequence of the advancement of cloud computing, many computing resources are provisioned as utilities on a metered basis to the client over the internet [1, 2]. Based on user demand, the cloud provider may easily and dynamically allocate and release these resources [3]. Virtual Machines (VMs) in the virtual cloud environment, play the most critical role as a resource container with business services encapsulated. As a matter of fact, due to ever-changing conditions, VM scheduling and optimization in a heterogeneous environment remains a challenging issue for cloud resource providers [4]. From the perspective of cloud providers, a massive number of resources are provisioned on virtual machines. In the cloud, thousands of users share the same amount of available resources fairly and dynamically. VM scheduling, at the same time, aims at ensuring the quality of service (QoS) along with cost-effectiveness [5]. Some major issues, supposedly interconnected with Infrastructure-as-a-Service (IaaS) in cloud computing are resource organization [6], data management [7], network infrastructure management [8], virtualization and multi-tenancy [9], application-programming-interfaces (APIs) and interoperability [10], VM security [11, 12] and the load-balancing [13].

Virtual Machine scheduling ensures a balancing scenario in which VMs are allocated to the available Physical Machines (PMs) as per resource requirements [14]. Moreover, VM scheduling techniques are utilized to schedule VM requests of particular datacenters (DC) according to the required computing resources. In essence, the optimization of virtual machine scheduling techniques to achieve efficient and effective resource scheduling gained larger attention of researchers in cloud computing [15].

The present literature in cloud computing scheduling can be categorized using performance matrix, and scheduling methods. The surveys that are based on performance focus on specific issues such as (i) energy-aware scheduling, (ii) cost-aware scheduling, (iii) load balancing-aware scheduling, and (iv) utilization-aware scheduling. The Methods-based survey categorizes (i) VM allocation, (ii) VM consolidation or placement, (iii) VM migration, (iv) VM provisioning, and (v) VM scheduling. The classification, as mentioned above, is discussed in Section 4 of this study. According to the author's best knowledge, several polls and studies have been conducted on the themes that were discussed earlier. However, an extensive study on virtual machine scheduling has been found missing in the available cloud computing literature. Hence, this study tries to do an extensive systematic survey on VM scheduling and presents the following contributions:

- To provide the outline of the techniques in virtual machine scheduling in the same manner as these techniques have been applied in cloud computing.
- To present syntheses of contemporary issues and challenges, and mention the problems related to virtual machine scheduling.
- To present a comparative analysis of VM scheduling methods and parameters in Cloud and mobile access computing (MAC).
- To Evaluate various VM scheduling approaches critically while highlighting their drawbacks and advantages.
- To emphasize the importance of virtual machine scheduling as a baseline for researchers to solve issues in near future.

Extensive examination and analysis of existing literature on contemporary issues and research gaps are crucial for generating ideas. This study tries to disseminate the most relevant VM scheduling techniques and approaches available in the literature and anticipates that they can effectively improve modern VM scheduling methods. This study attempts to present: recent trends, requirements and future scopes in the development of VM scheduling techniques in cloud computing.

The structure of the paper is organized as follows: Section 2 discusses literature reviews in cloud scheduling. Section 3 presents the research methodology. Section 4 illustrates virtual machine management methods and systems models. Section 5 presents the analysis of VM scheduling approaches and their parameters. Section 6 discusses scheduling in mobile edge computing and the validity of the research. Finally, Section 7 illustrates research issues and opportunities Section 8 concludes with the findings of the literature review.

2 Literature Review

Numerous studies are presented in the area of cloud scheduling, and some generic challenges are discussed such as resource scheduling, resource provisioning, and load balancing. Extensive surveys have also been found in the literature on virtual machine scheduling policy - VM placement, VM allocation, VM migration and VM scheduling. However, there is no extensive systematic survey on virtual machine scheduling in current studies. This section refers to some studies in the area of cloud scheduling. When it comes to allocating dynamic, heterogeneous, and shared resources, resource scheduling in cloud environments is considered to be one of the most crucial challenges. To provide reliable and cost-effective access, overloading of those resources must be prevented by proper load balancing and effective scheduling techniques.

Detailed review and classification of load balancing techniques are presented in [16], in which they compare the existing state-of-art techniques on parameters such as model, strength, gap, techniques and future work. Moreover, they have sufficiently analyzed and presented job migration techniques considering their description, merit, and demerits, which play a vital role in achieving fault tolerance. However, the study has not considered some of the job migration studies like predictive and heterogeneous job migration. At the same time, the scope of the study remains within grid computing.

In [17], a comprehensive survey of cloud scheduling algorithms which offer an analysis based on the categorization of some parameters that include; load balancing, energy management, makespan, and many more. The study observed that there is any scheduling algorithm that has the potential to effectively address all parameters of VM scheduling. Furthermore, the study discussed some task scheduling algorithms, limitations, and some future problems. However, the scope of the study is restricted to only grid computing. Similar work in [18, 19] presented a study of scheduling and energy-conscious resource allocation methods with a focus on the quality of service. They mentioned some critical and open challenges in cloud scheduling, particularly energy management in a cloud datacenter. According to their analysis of previous studies, the challenges are enumerated as follows: (1) Processes that are quick and energy-efficient for placing virtual machines and can anticipate workload peaks to prevent performance deprivation in a heterogeneous environment (2) energy-based virtual network topology optimization technique amongst VMs for the best location to lessen network traffic congestion, (3) to properly regulate temperature and energy use, new heat management algorithms, (4) even workloads and workload-aware resource allocation processes, and (5) Scalability and fault-tolerance techniques for virtual machine placement (VMP) challenges that are decentralized and distributed.

Virtual machine migration is a major issue for scheduling. In [20], the paper analyzed current VM migration techniques of thematic taxonomy that underline the commonalities and variances among VM migration schemes concerning certain performance metrics.. Additionally, they look into the difficulties with the VM migration plan, including the heterogeneity of cloud resources, the nature of dynamic workloads, system burden, VM memory size, and the severity of SLA breaches. Considering security is one of the significant concerns in the VM migration process, they suggest some safeguards such as (1) stopping unauthorized parties from accessing VMM; (2) separating VM borders; and (3) network connection security [21].

In another study, [22] investigated live VM migration schemes and present a particular taxonomy to categorize the concerned literature. They investigate storage optimization methods for WAN links, server consolidation rules, DVFS-enabled power optimization, and bandwidth optimization techniques based on their categorization. They also give a comparison of the results of other polls, highlighting some of the crucial factors in virtual machine migration. Their investigation identifies similarities and contrasts across existing VM migration plans based on a set of parameters found in the literature. Their research may be useful for creating intricate designs and optimization strategies for VM migration methods. The mathematical modeling of virtual machine migration strategies, however, is lacking in this research.

In a similar kind of work, Li, Li [23] investigated the scheduling issue for virtual machines in a cloud data center. Additionally, they provide a survey of current technologies, including virtualization, resource scheduling, virtual machine migration, security, and performance assessment in cloud computing. Similar to this, they cover certain upcoming problems and difficulties such as CPU architecture, resource management, upkeep procedures for system security, and performance assessment techniques in a system with several virtual machines. However, the paper lacked categorization, issue formulation, and parametric analysis, and did not conduct a thorough investigation of the methodologies as indicated in earlier research.

Analyzing the cloud computing architecture, Zhan, Liu [24] systematically presented two-level taxonomy of cloud resources. Researchers have critically examined the issue and remedy of cloud scheduling in their review. Additionally, they investigated EC methodologies and talked about several cutting-edge evolutionary algorithms and their potential to solve the cloud scheduling issue. Based on their categorization, they have also identified some of the next problems and research fields, such as distributed parallel scheduling, adaptive dynamic scheduling, large-scale scheduling, and multi-objective scheduling. They have also highlighted some of the most cutting-edge future themes, including the Internet of Things and the convergence of cyber and physical systems with big data. However, they failed to describe the problem's mathematical modeling or include any parametric analysis in the paper.

In another investigation, Xu, Liu [25] described the causes of the performance overhead problem of a virtual machine under several scenarios i.e., from single server-virtualization/ datacenter to multiple and distributed datacenters. The review presents a detailed comparison of contemporary migration techniques and modeling approaches to manage performance overhead problems. However, the authors suggest that there remains a lot to be resolved to ensure the predictable performance of VMs with guaranteed SLA. Similarly, Madni, Latiff [26] examine the difficulties and possibilities in resource scheduling for cloud infrastructure as a service (IaaS). They categorize the previous scheduling schemes according to the issues addressed and performance metrics and present a classification scheme. Furthermore, some essential parameters are evaluated and their strengths and weaknesses are highlighted. Finally, they suggest some innovative ideas for future enhancements in resource scheduling techniques.

One of the significant and recent studies suggesting a taxonomy of the algorithms for load balancing in virtual machine placement is presented by [27] The work is based on several existing models and techniques of load balancing algorithms employed in the virtual machine placement method. Their work summarizes various parameters and their optimization, contribution, gap, future challenges, and improvements. However, the scope of the review is limited to the virtual machine placement problem and ignores VM migration.

Meta-heuristic techniques become a benchmark in cloud computing scheduling because they exhibit efficient and near-to-optimal results in a reasonable time-space. Several types of research have been carried out to assess how well these modern meta-heuristic algorithms perform. In a similar study, Kalra and Singh [28] investigated six major Mete-heuristic optimization techniques namely, Ant Colony Optimization (ACO), Genetic Algorithm (GA), Particle Swarm Optimization (PSO), League Championship Algorithm (LCA) and Bat algorithm. Each Mete-heuristic technique is described in a taxonomical framework, and each technique is compared using some scheduling criteria, such as task awareness, SLA awareness, and energy awareness. Moreover, they have discussed the application of these meta-heuristic techniques and open challenges in the area of grid or cloud scheduling. However, the survey is only limited to specific meta-heuristic techniques and optimization criteria.

In another development, Madni, Latiff [29] investigated the potential of existing state-of-the-art Mete-heuristic techniques for resource allocation in a cloud computing environment for maximizing financial benefit for the cloud provider and minimizing cost for cloud users. In their research, they selected 23 meta-heuristic technique studies between 1954 and 2015. They compared meta-heuristic techniques with traditional techniques to evaluate the performance criteria of the algorithms. They claim that there can be several ways to enhance the performance of these algorithms which can further solve the resource scheduling problem. However, this review resembles the work of [28]. However, the focus of the paper is only on meta-heuristic methods.

Unlike previous studies shown in Table 1, our research presents an extensive (not exhaustive) review of virtual machine scheduling techniques and presents the most appropriate categorization, problem formulation, architecture and future challenges. Then, based on our research, we formalize three questions and choose the most important study from the most trustworthy research database to address them. Furthermore, we delineate the importance of virtual machine scheduling techniques, current issues and challenges, and future direction to support future research.

Table 1
Summary of previous literature in virtual machine scheduling

Previous Reviews	VM Scheduling	Problem formulation	Classification of VM Scheduling	Parametric Analysis	Simulation Tool & Environment	Dataset Available	Architecture	Period Covered
Li et al. [54]							√	2002–2009
Beloglazov et al. [9]	√	√		√		√	√	1991–2012
Rathore and Chana [50]				√				1999–2014
Xu et al.[56]		√	√	√			√	2003–2013
Abdulhamid et al. [51]				√	√		√	2009–2014
Kalra and Singh [58]			√	√				2001–2005
Zhan et al.[55]	√		√		√		√	2003–2014
Ahmad et al. [53]	√		√	√	√		√	1993–2014
Ahmad et al. [52]	√		√	√			√	1997–2015
Madni et al.[59]			√	√				1954–2016
Madni et al. [57]		√	√	√			√	2008–2016
Xu et al. [46]			√	√			√	2008–2016
Our Review	√	√	√	√	√	√	√	2008–2022

3 Research Methodology

According to the guidelines mentioned in [30] and [31], the presented systematic literature review (SLR) employs a tried-and-true procedure to examine earlier research by other researchers, which should provide sufficient details for other researchers to reproduce in the future [32, 33]. Following the best practice and guidelines, this study developed a protocol to accumulate the necessary details for virtual machine scheduling techniques, approaches and their parameters. Three research questions are established based on the analysis of collected literature on the main concerns with VM scheduling in cloud computing. The research questions are presented in the section below.

3.1 Research questions

In this section, the most important problems and challenges related to cloud-based scheduling were discussed, including resource provisioning, resource scheduling, task scheduling, virtual machine scheduling, resource utilization, load balancing, and prospective balancing solutions. Therefore, the effort of this research is to address the following important research questions:

Research Question (RQ1): What is the significance of VM scheduling in light of the increase in cloud usage? RQ1 will try to survey several virtual machine scheduling studies published over the period under study, to underline the importance of virtual machine scheduling along with increasing cloud usage.

RQ2: How many of the current scheduling strategies achieve the primary VM scheduling goals concerning the particular parameters? RQ2's objective is to assess current VM scheduling strategies in a cloud computing system based on the key VM scheduling parameters.

RQ3: What problems and potential solutions were found concerning VM scheduling for upcoming research trends? RQ3's goal is to classify the difficulties in VM scheduling in cloud computing and the methods utilized to ensure QoS in the system.

The specific responses to the questions posed within the scope of this study are obtained through a multi-stage approach. Once the necessity for the research has been established, a standardized process has been used to frame the research topic. The research must go through several processes to adhere to the protocol, including the search request, source selection, quality assessment criteria, extraction, and information analysis approach.

For respected online academic libraries and databases, search strings or keywords were created by defining keywords, which are based on inclusion and exclusion criteria. The Boolean "OR" and Boolean "AND" operators are used to connect similar and alternative spellings for each of the question elements to define keywords [34]. The search string is created using a combination of synonyms and alternate spellings for each element of the inquiry to find the pertinent topic. The best keywords from our subject have been chosen based on the established search string to obtain the desired outcome from databases. Thus, the terms "Virtual Machine," "VM," "Cloud," "Scheduling," and "Scheduler" have been chosen as the five keywords. The query was defined after going through many processes and assessing the findings of our preliminary study as a pilot to look at the result's coverage. Supposedly, if we used the pilot search from our studies and the original query did not yield the required results, we then modified our search, using terms like "Virtual machine" OR "VM" AND "scheduling," OR "scheduler." The search was carried out in August 2018 and covered the years 2008 to 2022.

3.2 Selection of sources

In the process of article selection, we have chosen some of the most relevant journal articles and conference papers from the most relevant academic databases for our search query. Subsequently, the selected results have been classified based on the publishers. We have searched through Web of Science, Scopus, and Google Scholar as our primary data source search engines. As a result, practically all of the articles published in the most reputable online journals and conferences that have undergone technical and scientific peer review were covered by the search process: Springer Link, ScienceDirect, IEEE Infocom, IEEE -Xplore, ACM-Digital Library, and ICDCS.

3.3 Selection criteria

An assessment method has been followed for the inclusion of the studies based on the prepared quality assessment checklist (QAC) in [35], to assess only specific articles from the peer-reviewed journal published between 2008 and 2022 as mentioned in Fig. 1. Based on the above filtering and analysis of the articles based on the checklist, a list of questions is prepared: (a) Does the research approach depend on the research article? (b) Is the research approach appropriate for the issue covered in the article? (c) Is the analysis of the study adequately done? (d) Does the survey meet the requirements for evaluation?

3.4 Extraction of Data and Quality Evaluation process

We compile the data from the chosen research during the data extraction process for additional analysis. Primarily, we selected a sum of 722 articles from all relevant databases. Then, we read keywords, abstracts and concepts that match our topic of study. Consequently, 88 articles were selected based on abstract and the rest of the studies were discarded. Then, the full body of each article was studied; those studies were not found suitable as the details mentioned inside the text were also removed. After summarizing the studies based on inclusion and exclusion criteria and QAC, 67 articles were selected for our review. Figure 1 demonstrates the overall inclusion and exclusion process followed in this study to identify the most suitable articles. As per the analysis of the retrieved data from relevant sources, a significant amount of growth can be observed in the articles published in the field of cloud scheduling during 2008 and 2022 as mentioned in Fig. 1 (a), (b) and (c). Among them, most of the publications were published in 2018.

3.5 Keyword search

In the formulation of our first question (RQ1), we particularly outlined the importance of virtual machine scheduling and the necessity to improve its mechanisms due to the high rise in data accumulation and resource utilization. Based on this perspective and the growing interest of the researchers in virtual machine scheduling, we only included peer review journal articles and conference papers from the most relevant digital libraries in Table 2. However, since we assumed that researchers and practitioners frequently use journals to obtain knowledge and disseminate new findings, we rejected conference papers that were not from trustworthy sources.

Table 2
Academic Database

Source	URL of the search engines	No. of returned articles
Google Scholar	http://scholar.google.com	360
Web of Science	http://www.webofknowledge.com	81
ACM Library	http://www.acm.org	19
IEEE Xplore	http://ieeexplore.ieee.org	122
Scopus	http://www.scopus.com	39
Springer	http://www.springerlink.com	34
ScienceDirect	http://www.sciencedirect.com	67
Total		722

3.6 Scope of the study

Based on the standards outlined in the study's procedure, the major studies were included. The 67 articles included in this study are further divided into two categories: those that specifically address the VM scheduling challenge and those that examine various problem-solving strategies.

The literature review will provide a solution to:

1. What is the present status of virtual machine scheduling in cloud computing?
2. What are the various methods used in virtual machine management?
3. What types of research are carried out in this area?
4. Why is virtual machine scheduling important in the area of cloud computing?
5. What are the approached prevalent to solving virtual machine scheduling problems?
6. Which approach may be opted for in the current cloud computing scenario?
7. How and why do virtual machine scheduling approaches impact the performance of resource management in cloud data centers?
8. What are the challenges in the design and devilment of virtual machine scheduling techniques in cloud computing?

Here, it is important to mention that, the foremost attention of this study is Virtual Machine (VM) scheduling, its architecture and the techniques used in the literature to solve the VM scheduling problem. Hence, we do not concentrate on the other underlying elements of cloud scheduling like task or job scheduling, workload scheduling, workflow scheduling etc. Also, the study does not consider VM migration in most cases. In the forthcoming section, the VM management methods are explained and the abbreviation used throughout the review is maintained in Table 3 with its illustration.

Table 3
Abbreviation and illustration

Abbreviation	Illustration	Abbreviation	Illustration
SLA	Service-Level-Agreement	SRC-I/O	Share Reclaiming and Collective I/O
SLR	Systematic-Literature-Review	SVS	Synchronization Aware VM Scheduling
DC	Data Center	HEFT	Heterogeneous Earliest Finish Time
VM	Virtual Machine	CDM	Common Deployment Model
PM	Physical Machine	AD	Active Directory
IaaS	Infrastructure as-a-Service	PD	Passive Directory
API	Application Program Interface	KVM	Kernel-based Virtual Machine
QoS	Quality of Services	DVMS	Distributed Virtual Machine Scheduler
PM	Physical Machine	BFD	Breadth First Depth
VMP	Virtual Machine Placement	BALA	Bandwidth-Aware Lago Allocator
VMM	Virtual Machine Management	VSA	VM Scheduling Algorithm
DVFS	Dynamic Voltage Frequency Scaling	GRANITE	Greedy Based Virtual Machine Scheduling Algorithm
WAN	Wireless Area Network	DCN	Data Center Network
EC	Evolutionary Computing	VMSAGE	VM Scheduling Algorithm based on Gravitational Effect
EASE	Energy Efficiency and Proportionality aware Scheduling	FEM	Fairness-aware VM Scheduling Method
SMP	Symmetric Multiprocessing	BFH	Best Fit Heuristic
ACO	Ant Colony Optimization	FHA	Find Host Algorithm
EEVS	Energy Efficient Scheduling	UTC	BAT Algorithm
vCPU	Virtual Central Processing Unit	BPA	Bandwidth Provisioning Algorithm
QAC	Quality Assessment Checklist	PSO	Particle Swarm Optimization
LCA	League Championship Algorithm	MCKP	Multiple Choice Knapsack Problem
GA	Genetic Algorithm	CGDPS	Cost Greedy Dynamic Price Scheduling
CMU	Cumulative Machine Uptime	ACOPS	Ant-Colony Optimization and Particle Swarm-Optimization
MST	Maximum Sustainable Throughput	TLBO	Teaching Learning Based Optimization
ERTE	Time and Resource Efficiency Metric	FCFS	First Come First Serve
PABFD	Power-Aware Best Fit Decreasing	LAVMS	Lock-aware Virtual-Machine Scheduling
VBP-Norm	Vector-Bin Packing Norm-based-Greedy Algorithm	MCT	Minimum Completion Time
CS	Cuckoo Search	MET	Minimum Execution Time
KH	Krill Herd	CIDD	Cloud Intrusion Detection Dataset
SA	Simulated Annealing	UTC	Universal Time Coordinated
DT	Dynamic Thresholds	SOS	Symbiotic Organisms Search
VHEST	Virtualized Homogeneous Earliest Start Time	AWS	Amazon Web Services
PVLOCK	Para Virtual Spinlocks	WOA	Whale Optimization Algorithm
CRTS	Composition Real-Time Scheduling Framework	NP	Non-Probabilistic
EASA	Energy-Aware Scheduling Algorithm	VMM	Virtual Machine Monitor
FC	Fog Computing	IoT	Internet of Things

4 Virtual Machine Management

Virtual machine management is a solution for virtual machine scheduling in the data center, which enables us to create and deploy the virtual host or VMs, allocate or de-allocate the VMs, mapping the VMs with the PMs to provide better QoS as per user demand. The virtual machine can be managed by different methods to achieve optimal resource utilization and cost saving. Virtual machine management methods consolidate the virtual machines on the physical machines without considering heterogeneity, which is one of the main aspects of modern-day data centers. Since finding the system's heterogeneity is essential to achieving considerable performance and effective resource management, it must be accounted for in designing VM management schemes. Many

studies have been done on management strategies in cloud data centers; however, there is a lot to be explored for the schemes that can improve the effectiveness of data center.

4.1 Classification of VM Management Method

In this section, we put forward the underlined methods for VM management and their possible classifications. According to the investigation of the surveyed literature in this study, the methods or techniques involved in VM management can be classified as VM Scheduling, VM Allocation, VM Placement, VM Migration, VM Consolidation and VM Provisioning. Things to be noted here, these methods are often used interchangeably in the literature, and the distinction between the actual method used becomes challenging to identify. However, the main focus of this study is VM scheduling since an ill-managed virtual machine scheduling on the data center in a heterogeneous environment not only leads to performance degradation of computing resources but also lowers energy efficiency, which results in more energy consumption [36].

This article focuses on virtual machine scheduling since it has the following advantages: scalability, QoS, a particular environment, decreased overheads and latency, enhanced throughput, cost-effectiveness, and a more straightforward user interface. The virtual machine management methods (see Fig. 2), can be classified as below, whereas an overview of virtual machine scheduling is illustrated in Fig. 3.

- VM Scheduling: Allocating a group of virtual machines (VMs) to a group of physical machines is the definition of a virtual machine scheduling problem [37, 38].
- VM Allocation: Allocating the user tasks to virtual machines is known as "VM allocation," and it often takes CPU, network, and storage requirements into account [39].
- VM Placement: It's a method for deciding which virtual machines (VMs) belong to which physical machines [40].
- VM Migration: Relocating a virtual machine means shifting it from one server or storage facility to another [41].
- VM Consolidation: As a result of the strategic placement of the VMs, we may reduce the number of necessary PMs [42].
- VM Provisioning: Configurable actions linked to deploying and personalizing virtual machines following organizational needs [43].

4.2 Systems Model of VM Scheduling

Figure 3 demonstrates the association between VMs and PMs. A sequence of all the PMs in the system here is represented as;$\rho = \{\rho_1, \rho_2, \ldots, \rho_{\aleph}\}, \aleph$ is the number of PMs, $\rho_i (1 \leq i \leq \aleph)$which represents PM i. Whereas, VMs set on the PM $\rho_i v_i = \{v_{i1}, v_{i2}, .., v_{imi}\}$ in which m_i is the number of assigned VMs on PM i. Considering $= \{ _1, _2, \ldots, _{\aleph}\}$ is the solution set which can be generated after the deployment of the VM v on each physical machine. Hence, $_i$ is the resultant solution set when VM v is mapped to PMρ_i.

4.2.1 The formulation of the load

A workload of a PM generally can be derived by summing-up the workloads of the VMs executing on it. We presume the finest time examined by previous data isτ. That is the period of τ from the existing time in the monitoring zone by previous data. According to the changing policies of PM workload, we can distribute the time τ into n times. Therefore, we define $\tau = [(t_1 - t_0), (t_2 - t_1), \ldots, (t_n - t_{n-1})]$. The equation states that, according to the changing policies of PM workload, the time τ is distributed into n smaller time intervals. In this notation, $(t_1, t_2, \ldots, t_{n-1})$ represent the end points of the n time intervals, and t_0 is the starting point. The values in the brackets represent the duration of each time interval, calculated as the difference between consecutive end points $(t_i - t_{i-1})$. The sum of all the duration of the time intervals is equal to the total duration of the period τ.

In the explanation, $(t_k - t_{k-1})$ signifies time k. Supposing the workload of VMs is fairly constant every time, then we can define the workload of VM numberin period k is v (i, k). Thus, we can determine that in cycle τ, where n is the number of instances in the index I and workload(i) is the workload value for the ith instance. So, the mean workload of the VM v_i on PM ρ_i is

$$v_i(\bar{i}, \tau) = \frac{1}{\tau} \sum_{k=1}^{n} v(i,k) \times (t_k - t_{k-1})$$

1

Going by the system policy, the workload of a PM is generally derived by summing-up workloads of the VMs executing on it. Hence, we can assume that the workload of the PM p_i where mi is the number of VMs on PMj.

$$\rho(i,\tau) = \sum_{j=1}^{m_i} v_i(\bar{j}, \tau)$$

2

The present VM requires placement as v. Then the previous VM configuration is required by the current schedular, and the estimation of the workload of the VM is v' based on historical data. Therefore, when v is mapped to PM, the workload of each PM should be

$$\rho(i,\tau)' = \begin{cases} \rho(i,\tau) + v' \text{ After deploy} v \\ \rho(i,\tau) \, Others \end{cases}$$

3

Typically, when v is allocated to ρ_i, there will be some variations in the system workload. Consequently, to achieve load balancing, we must do load adjustments. The load discrepancy of the mapping solution $_i$ in time τ after v is arranged to ρ_i as

$$\sigma_i(\tau) = \sqrt{\frac{1}{\aleph}\sum_{i=1}^{\aleph}\left(\bar{\rho}(\tau)\prime - \rho(i,\tau)\prime\right)^2}$$

4

where

$$\bar{\rho}(\tau)\prime = \frac{1}{\aleph}\sum_{i=1}^{\aleph}\rho(i,\tau)\prime$$

5

5 Vm Scheduling Approaches

As per the analysis of this review, the literature is categorized into three classes of approaches applied to solving VM scheduling problems. The first category of literature is based on the traditional approach, in which researchers have used basic and generic approaches for scheduling. The second types are heuristic approach, where researchers have applied heuristic techniques to solve their specific optimization problems. Whereas, the third class of approach is the metaheuristic approach, which implies modern intelligent algorithms and techniques for solving complex engineering problems. The distribution of the approaches is mentioned in Fig. 4 and Fig. 5. Table 4–6 present a comprehensive analysis of surveyed literature of each approach. We discuss each approach in detail in this section.

5.1 Conventional approach

Efficient virtual machine scheduling techniques are proven to be efficient in solving problems, such as high response time taken by tasks, distribution of the VMs on the physical hosts to achieve optimal load balancing, equal resource consumption, and server consolidation in data centers. The mentioned problems are addressed using Best-Fit and Worst-Fit algorithms, which follow two mechanisms. The reaction time is reduced by a factor of ($\log_n$) for the best-fit method, and by a factor of ($\log_n$) for the worst-fit method (1). In the worst-fit technique, the load on the PMs is equally distributed, but it requires additional VMs, such that every single host has to execute the processes. Then, in the best-fit process, every physical machine has equal resources left out for the execution of the remaining tasks. Better response times and more evenly distributed workloads on VMs are what the simulations suggest is possible. However, in the mentioned scheduling technique, they did not consider VM migration for the underutilized or overutilized host [44].

The paper elaborates the distinction between VMs scheduling and processor task scheduling in a traditional computing environment. Also, it points out some key advantages and challenges of VMs scheduling. The proposed gang scheduling-based co-scheduling algorithm works in two fashions. Firstly, the algorithm schedules the coherent processes to run simultaneously on different processors. At the same time, it maps the related virtual CPUs (vCPUs) to the real processors. The simulation results exhibit faster execution of processes that execute on VMs and display higher performance and avoid unnecessary VM blocks [45].

Hu, Jin [46] presented a novel scheme for virtual machine scheduling using live migration of virtual machines to the under-loaded server clusters. The scheme named Magnet shows a better reduction in energy saving and is applied to both homogeneous and heterogeneous physical machines in the data center. The scheme also claimed an appositive impact on average job slowdown and a negative impact on the execution time for task processing. The authors of [47] measured the performance of interactive desktops and try to solve the latency peak problem that arises during server peak workload. The proposed method enhances the XEN credit scheduler to analyze the latency for peak operation. They claim to reduce latency and frequency by their scheduler in comparison to the default one.

Von Laszewski, Wang [48] anticipated the Dynamic Voltage Frequency Scaling (DVFS) technique to analyze the problem of energy consumption in computer clusters. The proposed design focuses on the allocation of VMs on the DVFS-enabled clusters. The simulation results show an acceptable reduction in energy consumption. Lago, Madeira [49] presented an optimization algorithm for virtual machine scheduling considering bandwidth constraints in a heterogeneous network environment. There techniques work in two steps, first they used Find Host Algorithm (FHA) to find the optimum host to allocate the available virtual machine which is executed by the cloud broker. Secondly, the Bandwidth Provisioning Algorithm (BPA) is used to provision the network bandwidth for the VM which is to be run on the host machine. In the simulation results the proposed algorithm shown significant reduction in energy saving and a better makespan.

In VM scheduling of heterogeneous multicore processor environment, two key issues are significant to achieve an efficient performance. Characteristics of VM for optimum VM placement at the suitable core and the actual source of delay to eliminate the impede cloud performance. The authors of [50] developed a plan to allocate resources among the several virtual machines. The authors discuss performance dependence on the physical host and responsiveness to CPU clock frequency. The simulation outcomes show that the proposed scheduling policy is effective in energy saving in a cloud environment. In a cloud data center excessive amount of energy is consumed by the virtual machine scheduler. Knauth and Fetzer [51] suggested the energy-aware scheduling algorithm

OptSched to minimize energy-saving problems in cloud computing. Simulation results show that the enhanced method can significantly reduce CMU up to 61.1% when compared with the default scheduler round-robin and is considered the best fit in OpenStack, OpenNebula, and Eucalyptus.

One other study proposes a credit-aware virtual machine scheduling method to reduce data center overhead. The mechanism seems to be easy to implement with a simplified design. However, the experimental result does not show optimal performance in all cases and is even not implanted in the real cloud [52]. In stream data processing, the demand of the workloads verily changes over a period of time. To maintain seamless processing the VMs need to allocate and deallocate frequently by the Virtual Machine Manager (VMM). In this so-called steam processing scenario, maintaining QoS is a challenging task and requires adaptive scheduling techniques to handle uncertainties.

Imai, Patterson [53] provided a proactive elastic VM scheduling framework to forecast the arrival of workloads, when the estimation is done for the arrival of the highest workload the minimum amount of VMs is allocated to handle that workload. To know the uncertainties from VM and application they have used MST (maximum sustainable throughput) model. The authors applied their framework on three different workloads and were able to achieve 98.62% of QoS satisfaction and 48% less cost in comparison to static scheduling.

On the other hand, there is a high possibility to discover a high amount of content similarity and identical disk blocks with a similar operating system and the same host with the help of VMs scheduling. The researcher observed that a similarity between VM images can be as high as 60–70% which causes a reduction in the amount of data transfer in the VM deployment process. Based on the above notion, Bazarbayev, Hiltunen [54] developed a content-based scheduling scheme to reduce the network congestion which is related to the VM disk images transfer process inside data centers. Data center network usage and congestion are significantly reduced as a result of the algorithms' evaluation, which shows a reduction in data transfer of up to 70% during the processes of VM migration and virtual disk image transfer.

Table 4
Analysis of Conventional approach used in virtual machine scheduling

Reference	Problem addressed	Algorithm / Technique	Improvement / Achievement	Weakness / Limitation	Tool / Hypervisor
Knauth and Fetzer [86]	VM Scheduling	OptSched Technique	• Improve energy saving • Reduce machine uptime	• Does not work on real cloud • Low resource utilization	Python
Pegkas et al. [87]	VM Scheduling	Credit based algorithm	• Improve response time • Minimize the finish time	• Low performance in all cases	Python
Takouna et al. [85]	VM Scheduling	VM scheduling policy	• High energy saving	• Used basic DVFS mechanism • Heterogeneous VMs	Xen hypervisor
Imai et al. [88]	Elastic VM scheduling techniques	New framework for proactive elastic VM scheduling	• Better quality of services	• Single objective application • Low scalability	Not mentioned
Bazarbayev et al. [89]	VM Scheduling and Placement	Content based scheduling algorithm	• Improve network utilization • Reduce network congestion	• High response time	Ubuntu Server
Rathor et al. [79]	VM Placement technique, Load Balancing, Server Consolidation	Best-fit and Worst-fit algorithm	• Reduce response time • Better resource utilization • Cost saving	• Do not consider underutilized host and over utilized host for migration	CloudSim
Salimi et al. [80]	Scheduling advantages and optimization	Virtual Processor co-scheduling method	• Increase system performance	• The work performs only 4 tasks	CloudSim
Lago et al. [84]	VM Scheduling	Dynamic Voltage Frequency Scaling (DVFS)	• Energy saving	• Heterogeneous virtual machines	CloudSim
Hu et al. [81]	VM Scheduling	Magnet	• Reduce Energy	• High execution time	Xen hypervisor
Xia et al. [82]	VM Scheduling	Not mentioned	• Reduce latency	• Basic approach adopted	Xen hypervisor
Von Laszewski et al. [83]	VM Scheduling	Not mentioned	• High energy saving • Reduce CO2 emission	• Result showed discrepancy in data	nBench, Linux and DVFS-SIM / OpenNebula

5.2 Heuristic approach

The heuristic approach to handling complex optimization problems is explained as trying to find a probable number of solutions to an NP-hard problem and suggest the best solution to achieve some specific objective function. It is mostly bound with hard and soft constrains which must not be overlooked in the optimization design. Heuristic approaches perform where traditional approaches fail; especially in the high dimensional or multimodal space when a problem can be addressed using more than one solution. In this context, many researchers have applied heuristic approaches in their work and achieved effective solutions to their problems. Table 5 shows a descriptive analysis of the heuristic approach. We discuss here the heuristic approaches used to solve VM scheduling problems.

In the SMP (Symmetric Multiprocessing) virtual machine scheduling, dynamic load balancing and CPU capping techniques are used which consequently results in a significant number of inefficiencies in parallel workloads. In a virtualized system, where the tenants rent the resources, fairness among them considered being the key to success in running their applications effectively. However, the available virtualization platforms do not implement fairness in a condition where some VMs contain several virtual CPUs running on different CPUs. Based on this method, Rao and Zhou [55] developed an innovative vCPU scheduling technique namely Flex, that applies fairness at the virtual machine level and also increases the effectiveness of parallel running applications on the host servers.

In other progress, an efficient dynamic VM scheduling- the algorithm is developed to address the energy-consumption problem with the concentration of deadline constraints [56]. The study presents a robust energy-efficient scheduling technique namely EEVS, which can be capable of dealing with various

physical nodes and equally performs in a dynamic voltage environment. Furthermore, the algorithm considers scheduling periods and optimal performance-power ratio as performance parameters. Experiment analysis shows that in the best instances, VMs can reduce their energy consumption by over 20% while increasing their processing power by 8%.

Quang-Hung and Thoai [57] Time and Resource Efficiency Metric (ERTE) is a suggested technique for scheduling virtual machines that take energy efficiency into account to reduce data center idle time. In addition, the suggested approach was evaluated in terms of power consumption alongside two state-of-the-art algorithms: power-aware best fit decreasing (PABFD) and vector bin packing norm-based greedy algorithm (VBP-Norm L1/L2). Based on experimental results, the suggested scheduling method not only improves performance by 48% but also reduces average energy usage by 49%.

In the virtualized environment and with the presence of an intensive mixed workload, reducing energy consumption is considered one of the challenging tasks. Xiao, Hu [58] to reduce the energy consumption caused by I/O virtualization, a mixed-workload energy-aware virtual machine scheduling technique was developed. Additionally, they developed a novel scheduler called SRC-I/O by fusing two newly designed techniques: share-reclaiming and communal I/O. Both the share-reclaiming method and the collective I/O method aim to increase CPU utilization and reduce context-switching costs due to I/O-intensive workloads, respectively. Simulation results reveal that the SRC-I/O scheduler outperforms its rival on a different performance matrix.

Increases in virtualization technologies have allowed for massive VM consolidation in data centers. Services that depend on rapid responses could be hampered by a lack of availability if they didn't have access to latency-sensitive task support. In this regard, Kim, Lim [59] accommodate latency-sensitive tasks, it is necessary to devise a priority-based virtual machine scheduling method that takes into account the needs of guests. The provided method schedules the required VMs for workload allocation based on the priority of the VMs and the current state of the guest-level tasks running on each VM. In addition, it selects for scheduling those virtual machines (VMs) that are capable of running latency-sensitive applications with the quickest possible response to I/O events. [60] reduce the virtual machine's carbon footprint by putting forward a cognitive scheduling method based on its camera's eye. The suggested method seeks to identify the optimal PM to allocate to a virtual machine so that it may be run within a specified response time. When compared to other algorithms, this one is 17% more efficient at saving power. Due to SLA violations of up to 14%, the proposed algorithm does not achieve optimal performance with response time.

Due to high flexibility and cost-effectiveness, multiple applications run concurrently on the virtual cloud. Running tightly-coupled parallel applications is a feasible solution over the clustered cloud environment for better resource utilization. However, due to over-commitment in the cloud and ignorance of the synchronization constraint of VMs by Virtual Machine Monitor (VMM), performance degradation is taken into consideration in recent research. To overcome this problem, Wu, Lu [61] emphasized the role of dynamic workload on the VM in a Data Center Network (DCN) and presented a VM scheduling to improve the elasticity as a new QoS parameter. A new precedence-constrained parallel Virtual Machines (VM) consolidation algorithm is anticipated by [62], which tends to improve the resource utilization level of physical machines, and also display minimum energy consumption. Simulation results show their algorithm performs better in comparison to Heterogeneous Earliest Finish Time (HEFT), in reducing energy and makespan time of the services.

Saravanakumar and Arun [63] proposed a Common Deployment Model (CDM) based on a brokering mechanism to manage virtual machines in cloud data centers efficiently. After a task has been completed, the current state of the virtual machine (VM) is preserved using the active directory (AD) and passive directory (PD). These folders are used for two processes, VM migration and VM rollback, and ensure that virtual machines have the correct configuration mapping of the physical computers. The suggested model takes into account VM downtime for various job kinds. When it comes to managing unused virtual machines (VMs) in a repository, the CDM model is contrasted with the iCanCloud concept. Keeping the inactive VM in the hypervisor eliminates the latency issue that arises when moving VMs between the hypervisor and the VM repository. The experimental results show that the CDM-based model takes less latency in VM management. They proposed two algorithms for VM scheduling and VM placement to achieve effective utilization of VM. Further, they have compared both algorithms with different scheduling and placement algorithms respectively. VM scheduling algorithms show a better result when compared to other algorithms regarding CPU utilization. Whereas, VM placement resulted in better improvement in terms of completion time of VM placement and resource utilization.

I/O performance degradation is a common phenomenon in a virtualized environment. The virtual machine is not able to distinguish the different processes coming from the same physical machine. Since the process information is located in the higher layers, getting it can be challenging. To address this problem, Xie, Cao [64] suggested a disk predictive scheduling method that takes into account running processes be used to solve the disk I/O issue. With the assistance of a predictive model, the VMM in this approach learns about the process and then uses that knowledge to categorize the I/O request. The connection between a process and its address space is used to infer the process's level of awareness. The simulation results validate the practicability of the proposed strategy and highlight the subsequent increase in disk I/O speed.

In a multi-core virtualized environment, Symmetric Multiprocessing (SMP) is increasingly being used for efficient resource utilization and performance degradation. There a separate scheduler exists in the hypervisor as well as in the guest host resulting in a problem of double scheduling. To overcome this problem Miao and Chen [65] evaluated a scheduling scheme FlexCore using vCPU ballooning. The scheme dynamically adjusts the number of vCPUs of a VM at runtime and eliminates unnecessary scheduling within the hypervisor layer to considerably improve the performance. The experimentation is done on a KVM-based hypervisor that shows that the average performance improvement is approximately 52.9%, ranging from 35.4–79.6% for a 12-core Intel machine for PARSEC applications. In a similar progress, Kertesz, Dombi [66] presented an improved pliant-based VM scheduling scheme for solving energy consumption problems. The authors in their work utilized industrial application workloads to evaluate the performance of their improved CloudSim framework. The results depict a significant improvement in energy saving and a better trade-off in execution time.

Due to the hard scalability problem in a distributed virtualized cloud environment, it is difficult to manage VMs by virtual machine In-charge on a pool of physical machines. It becomes worse in the case of VM image transfer. In this regard, Quesnel, Lèbre [67] provided a new Distributed Virtual Machine

Scheduler (DVMS), which acts as a decentralized and preemptive scheduler in a massive-scale distributed environment. As shown in the results, the elements of the validation approach are sufficiently solving the resource violation problem.

In another kind of progress, Adhikary, Das [68] suggested a distributed and localized VM scheduling algorithm (VSA) to cater to energy consumption problems in data centers. The proposed algorithm functions as an intra-cluster and inter-cluster scheduling and addressed some major parameters such as energy, resource estimation, and availability. It schedules VMs in a way that energy consumption is minimized for both servers as well as networking devices. The results show that the algorithm outperforms compared to other existing algorithms concerning energy reduction.

VM consolidation is often used to solve energy consumption problems. Secondly, energy consumption can also be managed by sending the real-time resource requirement to the VMM and controlling the frequency of recourse demand. In that essence, a power-aware framework is introduced for compositional real-time scheduling. The method encapsulates each VM into a single component to minimize resource utilization and thus reduce energy. The framework is implemented on Xen hypervisor on Linux kernel and results in better performance [69].

Efficient virtual machine scheduling increases the performance of the data center and increases the profitability of the cloud providers. In this regard, Li, Garraghan [70] offered a greedy-based virtual machine scheduling algorithm GRANITE to reduce datacenter energy consumption following two major strategies VM placement and VM migration. They have used computational fluid dynamics techniques to address the cooling model of the datacenter. Moreover, they claim to address the CPU temperature for the first time along with the other infrastructure devices and nods. The approach outperforms when compare to other contemporary algorithms in reducing the energy overhead while balancing the SLA.

In a different work, Li, Zhu [71] improved the deficiency of the semi-homogeneous tree to a general heterogeneous tree as its optimal solution. The proposed maximum elasticity scheduling both maximum elasticity computation and maximum elasticity communication using a hose model.

Inspired by the gravitational model of physics, Xu, Zhang [72] presented a Virtual Machine Scheduling Algorithm based on Gravitational Effect (VMSAGE) to handle the issue of energy consumption in data centers. This work is the extension of [73] in which the authors presented a heuristic-based approach for VM scheduling for Fog-cloud. To assure optimum utilization of the resources, their method addressed the issue of load balancing and achieved better resource utilization on the edge network.

Table 5
Analysis of Heuristic approach used in virtual machine scheduling

Reference	Problem addressed	Algorithm / Technique	Improvement / Achievement	Weakness / Limitation	Tool / Hypervisor
Rao and Zhou [90]	Dynamic VM scheduling	Symmetric Multiprocessing (SMP) based VM scheduling scheme (Flex)	• Achieve fair CPU allocation	• Low performance	Xen 4.0.2
Ding et al. [91]	VM scheduling	Energy efficient VM scheduling (EEVS)	• Reduce energy consumption • Low execution time	• Power penalties of status transitions • VM migrations are ignored	Not mentioned
Quang-Hung and Thoai [92]	VM Scheduling	ETRE Algorithm	• Low total busy time of all PMs	• Does not consider other parameters	CloudSim
Xiao et al. [145]	VM Scheduling	Shared Reclaiming with collective IO Scheduler (SRC-I/O)	• Minimum CPU utilization • Better energy efficiency	• Low scalability	Xen Hypervisor
Beloglazov and Buyya [93]	Dynamic VM scheduling	Dynamic Thresholds (DT)	• Improve energy consumption • High level of SLA	• Applied on single core CPU only	CloudSim
Kim et al. [94]	VM Scheduling	Priority-based scheduling scheme	• High timeliness and CPU fairness • Low response time	• Required kernel modification to implement the scheme • Used open-source OS that may encounter security issues	Xen 3.0.4, para-virtualized Linux 2.6.16
Zhao et al. [95]	VM Scheduling	Vision cognition algorithm	• Improve energy saving	• SLA violation	CloudSim
Wu et al. [96]	VM Scheduling	Synchronization aware VM scheduling algorithm (SVS)	• High application performance • Better execution time	• Low resource utilization	Xen hypervisor
Ebrahimirad et al. [97]	VM Scheduling	Virtualized homogeneous earliest start time (VHEST)	• Improved utilization • Makespan reduction • Reduced power consumption	• Homogeneous virtual machines	VDCS (Virtualized Data Center Simulator)
Saravanakumar and Arun [98]	VM Scheduling and VM Placement	Common Deployment Model (CDM)	• High resource utilization	• Compared with iCanCloud	CloudSim
Xie et al. [99]	VM Scheduling Scheme	Process-aware predictive scheduling	• Improve disk I/O speed of the process	• Based on only Xen hypervisor	Xen Hypervisor
Kim et al. [100]	VM Scheduling Scheme	Task aware VM scheduling scheme	• High performance	• One vCPU on single CPU • Does not consider task migration and synchronization issues	Xen Hypervisor
Miao and Chen [101]	VM Scheduling Scheme	FlexCore scheduling scheme	• High performance	• Does not consider VM migration	KVM Hypervisor
Kertesz et al. [102]	VM Scheduling	Pliant-based virtual machine scheduling	• Low execution time	• Cost saving for provider's only	CloudSim
Quesnel et al. [103]	VM Scheduler	Distributed virtual machine scheduler (DVMS)	• High system utilization	• Does not show the QoS improvement	KVM hypervisor
Adhikary et al. [104]	VM Scheduling	Virtual machine scheduling algorithm (VSA)	• Better energy conservation	• Worked for network devices with fixed experiment condition	CloudSim
ho Seo et al. [105]	VM Scheduler	Composition real-time scheduling framework (CRTS)	• Low power consumption	• Worked with only two VMs	RT –Xen Hypervisor

Reference	Problem addressed	Algorithm / Technique	Improvement / Achievement	Weakness / Limitation	Tool / Hypervisor
Li et al. [106]	VM Scheduling	Greedy based holistic approach (GRANITE)	• Reduce Energy consumption • Low SLA violation	• Do not compare with benchmark systems	CloudSim
Wu et al. [107]	VM Scheduling	Maximum elasticity scheduling	• High computation time • High communication elasticity	• Feasible for cloud data center network (DCN)	Matlab
Li et al. [108]	VM Scheduling	Hierarchical VM placement algorithm	• High resource utilization	• Homogeneous VMs	Matlab
Xu et al. [109]	VM Scheduling	Gravitational effect based virtual machine scheduling (VMSAGE)	• Reduce energy consumption • Minimize migration time	• Compare with on conventional BFS and DVFS	CloudSim
Xu et al. [111]	VM Scheduling	HSM scheduling method	• Better load balancing • Improve resource utilization	• Compare with on conventional FFD and BFD	CloudSim
Lago et al. [112]	VM Scheduling	Bandwidth-aware lago allocator (BALA)	• Low energy consumption • Low makespan	• Performance degradation	CloudSim
Al-Dulaimy et al. [113]	VM Scheduling	Multiple choice knapsack problem (MCKP)	• Reduce energy consumption • Improve PMs utilization	• Live migration cost overhead	CloudSim
Xu et al.[114]	VM Scheduling	Cost-greedy dynamic price scheduling algorithm (CGDPS)	• Enhance execution time • Improve cost saving • Increase fairness of users	• CPU cores cannot be allocated to more than one VM	Not mentioned
Yu et al. [115]	VM Placement	Lock-aware virtual machine scheduling scheme (LAVMS)	• Reduce CPU waiting time • High performance	• Implemented on limited virtual machine	Xen-based prototype
Qiu et al. [116]	VM Scheduling	Energy efficiency and proportionality aware scheduling (EASE)	• Low energy consumption • High completion time	• Work for few numbers of VMs	KVM/QEMU
Xu et al. [118]	VM Scheduling	VM scheduling heuristics	• Reduce energy consumption • High performance	• Implemented does not support VM migration	CloudSim
Xing et al. [117]	VM Scheduling	Fairness-aware VM scheduling method (FEM)	• Improve fairness • High power saving	• Low scalability • Low resource utilization	CloudSim
Xu et al. [119]	VM Scheduling for WMAN	MFEA Scheduling technique	• Energy saving	• Resource wastage	Not mentioned
Wan, Dang [81]	System queuing scheduling model	Particle optimization	• Performance and Cost	• Semi Metaheuristic approach	Matlab
Qi, Chen [82]	VM Scheduling	QVMS using NSGA-III	• Energy and downtime	• Increased migration cost	NA

Reference	Problem addressed	Algorithm / Technique	Improvement / Achievement	Weakness / Limitation	Tool / Hypervisor
Saravanakumar, Geetha [83]	Clustering based VM scheduling	Cloud radio access network (C-RAN)	• Network-overhead, allocation time	• Data size volume constraints ignored, Work in only homogeneous environment	CloudSim
Xu, Xu [84]	VM Scheduling	Greedy-based best fit decreasing (GBFD)	• QoS	• Dynamic workload overlooked	CloudSim

Using Virtual Machine Management (VMM) strategy, Al-Dulaimy, Itani [74] anticipated an improved energy-efficient VM scheduling technique for dynamic consolidation and placement of the virtual machines in data centers. In this strategy, Multiple Choice Knapsack Problem (MCKP) first decides the set of VMs to migrate from the under loaded and overloaded PM criteria. Then VM selection is performed from the generated candidate solutions, and finally, these selected VM is placed on the number of PMs. The proposed method outperforms when compare to similar strategies in terms of energy saving.

In a similar work, Xu, Liu [75] investigated the VM scheduling problem and proposed an incentive-aware scheduling technique for both cloud providers and cloud users with a guaranteed QoS. In this work, the improved meta-heuristic method namely Cost Greedy Dynamic Price Scheduling (CGDPS) prioritizes the VM requests as per the user demand and generates several candidate solutions. Finally, the VMs are assigned to the candidate node with minimum computation cost. The comparative results show a competitive improvement in user satisfaction.

In the study of Yu, Qin [76] a synchronization problem in VM scheduling is addressed to avoid the extra-long waiting time assigned to a vCPU for lock spins. The proposed Lock-aware Virtual Machine Scheduling (LAVMS) provides additional scheduling chances for processors to avoid locks. The method ensures the scheduling without wasting the waiting time of the vCPU. The scheme outperforms when compare to the contemporary para-virtual-spinlocks (PVLOCK) in terms of performance. Along the same lines, Qiu, Jiang [77] introduced an energy efficiency and proportionality-aware VM Scheduling framework (EASE). The framework set out the standard benchmarking as per the specified configuration components of the servers. Again, it addresses the real workload which again configuration centric to the servers. Then, real-time server data is collected, efficiently is identified, and finally workload classification is performed to achieve optimum VM scheduling. The simulation results depict a significant reduction in energy and completion time up to 49.98% and 8.49% respectively in a homogenous cluster. Similarly, in heterogeneous clusters, it has been observed 44.22% and 53.80% respectively.

Considering resource provisioning a major concern for Internet of Things (IoT) applications [78] adapted a fairness-aware VM scheduling method (FEM) to achieve fairness and energy saving. Therefore, the system is designed and evaluated on three IoT datasets and compared with the benchmark energy-efficient VM scheduling (EVS). The experimented graphs show superior performance in resource- fairness and power saving. In the same context, Xu, Zhang [79] considered the balancing scenario between energy saving with guaranteed performance and introduced a novel VM scheduling technique for Cyber-physical system. The joint-optimization model-based method utilizes the live migration of the VMs to underloaded PMs to offload the overhead consequently reduced power consumption and performance degradation. [80] examined the power management problem in Wireless Metropolitan Area Network (WMAN) and put forward a VM mapping strategy to reduce power consumption. The proposed method namely MFEA is optimized to reduce the number of VMs on the physical servers after migrating the underutilized VMs. The experimental graph shows comparable energy reduction with other benchmark techniques.

In a different work, Wan, Dang [81] proffered a system queuing scheduling model to analyze the performance of the cloud systems by switching off and on the (hot and cold shutdown) of the VMs. The proposed method uses multi-objective particle optimization to optimize the most critical parameters in the cloud scheduling process, such as performance and cost. However, the heuristics approach is not used in the true sense and the description is lacking. Similarly, Qi, Chen [82] developed a QoS-aware cloud scheduling system by applying the NSGA-III algorithm to find the optimal VMs to migrate on the PMs in the cyber-physical system (CPS). The algorithms generate multiple VM scheduling solutions and select the best strategy to map the VMs. In another work, Saravanakumar, Geetha [83] proposed a VM clustering method to monitor the performance measure of the VM metrics such as network-overhead cost. It dynamically allocates the submitted tasks to the VMs to deal with the network overload problem and reduce the allocation time. However, the proposed method lakes in dealing with the volume of the data size constraints. Furthermore, Xu, Xu [84] addressed one of the significant factors called reliabilities in VM scheduling and presented a fault tolerance scheduling system while satisfying several QoS. They designed a greedy-based technique to identify suitable computer nods to execute the user's tasks with improved performance.

5.3 Meta-heuristic approach

The distinction between heuristic and meta-heuristic is overwhelming. Both, heuristic and meta-heuristic approaches are used to solve high-dimensional and multi-model problems and provide near to optimal solutions for a problem. Heuristic approaches are problem specific, whereas, meta-heuristic approaches are more generalizable and adaptable. The latter can guide, modify, and hybridize with other heuristic approaches in the process of local optima generation [85].

Nature-inspired meta-heuristics contain immense power in solving complex engineering problems. Meta-heuristic approaches have unique features in striking a balance between exploration and exploitation phases, and in avoiding local optima stagnation [86]. Due to these unique and promising features, researchers around the world prefer using meta-heuristic approaches in their efforts to solve optimization problems. In this section, we discuss the most relevant metaheuristic approaches used in solving VM solving problems. Table 6 shows a brief analysis of the meta-heuristic approach. Furthermore, the parameters used in this surveyed literature are shown in Table 7. In Fig. 6 and Fig. 7, the distribution of the literature based on parameters used in numbers and percentages is mentioned. Figure 8 and Fig. 9 show a comparison of single-objective and multi-objective optimization problems used in the literature. Table 8 maintains a list of available datasets for cloud computing. Here, in this subsection, we talk about using various meta-heuristic techniques to address VM scheduling issues.

In a cloud environment that has been virtualized, the incoming requests frequently change. The types of requests a virtual machine (VM) may get and the tasks it will carry out are unknown to the system. Therefore, a technique either considers a fixed number of tasks or requires detailed information about the

tasks has become insignificant. In this regard, Cho, Tsai [87] introduced a hybrid meta-heuristic approach that incorporates ACO and PSO, two highly developed algorithms, to tackle the VM scheduling problem. To anticipate incoming workload and adapt to changeable settings, the proposed ACOPS algorithm employs previously stored information on the server. To save computing time, it does not require any more job information and disproves unmet scheduling needs. The simulation graphs demonstrate that the suggested algorithms outperform other comparable systems and have a balanced cognitive burden. In a cloud environment that has been virtualized, the incoming requests frequently change. The types of requests a virtual machine (VM) may get and the tasks it will carry out are unknown to the system. Gondhi and Sharma [88] developed a VM allocation problem solution based on the ACO algorithm. The authors modified the ACO by using a local search algorithm to maximize the allocation result because they believed that the combinatorial problem of bin packing was NP-hard.

Table 6
Analysis of Meta-heuristic approach used in virtual machine scheduling

Reference	Problem addressed	Algorithm / Technique	Improvement / Achievement	Weakness / Limitation	Tool / Hypervisor
Hu et al. [125]	Load balancing	Genetic algorithm (GA)	• Reduce load imbalance • Low migration cost	• High makespan	OpenNebula and C++
Kumar and Raza [126]	VM Scheduling and Placement	Particle swarm optimization-based policy	• Reduce resource wastage • High server utilization	• Low performance	Eclipse Kepler 2
Cho et al. [123]	VM Scheduling	ACO-based Vm scheduling (ACOPS)	• Improve resource utilization	• Work on single objective • Homogeneous synthetic cloud	Test-bed@NCKUEE
Gondhi and Sharma [124]	VM Allocation	Local search-based Ant colony optimization	• Reduce energy consumption • Better resource utilization	• Only one optimal solution • Compared only with BFD	CloudSim
Liu et al. [127]	VM Scheduling	Adaptive penalty function (CGA)	• Improve deadline constraint • Save execution cost	• Independent task	WorkflowSim
Wang et al. [128]	VM Scheduling	Improved teaching learning-based optimization scheduling strategy (TLBO)	• High energy saving	• Does not compare with benchmark algorithms	Not mentioned
Qin et al. [129]	VM Scheduling strategy	Semi sleep mode VM scheduling	• High energy saving • Improve average latency	• Does applied on real time workload • No comparison shown	Matlab 2010a
Xu and Li [130]	VM Scheduling methods	Learning effects models	• High execution time • Reduce makespan	• Work for single VM only • Does not show practical implementation	MapReduce
Zhao et al. [131]	VM placement	Divide and conquer strategy with branch and bound algorithm (DCBB)	• Low execution time • Better convergence speed	• Yet to prove theoretically • Algorithm adaptation on DCBB is not clear	Amazon Elastic Compute Cloud (EC2)
Sui et al. [132]	VM Scheduling	Genetic algorithm based SVR_GA for classification, Differential evaluation based adaptive algorithm for local search (ESA_DE)	• Reduce energy • Low virtual migration	• Low scalability • Increase throughput	CloudSim
Li et al. [133]	Dynamic VM scheduling	GA based dynamic VM scheduling strategy	• Improve utilization • Better load balancing	• No significant results	CloudSim/ OpenStack
Feng et al. [134]	Predictive VM Scheduling	Revivification-based prediction (ERP) model and ERPA	• Reduced execution time	• Conservative time synchronization schema	Java
Karthikeyan and Soni [99]	VM Scheduling	GA, variable neighborhood search (VNS) and PSO based approach	• Utilization and Completion time	• Did not mentioned algorithm improvement	CloudSim

Reference	Problem addressed	Algorithm / Technique	Improvement / Achievement	Weakness / Limitation	Tool / Hypervisor
Kruekaew and Kimpan [100]	VM Scheduling	Enhanced ABC	• Makespan and degree of imbalance	• High recourse cost	Matlab
Naik, Singh [101]	VM migration	Fruit fly Hybridized Cuckoo Search (FHCS)	• Energy and resource leakage	• Did not considered deadline constraints	CloudSim
Rana, Abd Latiff [102]	VM Scheduling	M-WODE	• Makespan and Cost	• Migration cost ignored	CloudSim
Medara and Singh [103]	VM Scheduling	EASVMC	• Energy reduction and utilization	• Deadline constraint ignored	WorkflowSim
Ajmera and Tewari [104]	VM Scheduling	VMS-MCSA	• Energy	• Tested on synthetic workload	CloudSim
Chaudhury [105]	VM Scheduling	Particle Swarm optimization and Ant Colony Optimization approaches called (PSACO).	• Load Balancing, energy	• High computational cost	CloudSim
Alsadie [106]	VM Scheduling	Metaheuristic framework called MDVM	• Energy usage, makespan and cost	• High computational cost, homogeneous environment considered only	CloudSim
SS and HS [86]	VM Scheduling	GA based Technique	• Energy usage, utilization	• SLA Violation in VM migration	CloudSim
Sheng, Hu [107]	ML based VM scheduling prediction system	SchedRL	• Allocation time	• Increased computational time	Python

VM scheduling can be perceived as the allocation and placement of several VMs to a set of PMs. In this regard, Kumar and Raza [89] proposed an enhanced VM scheduling policy for VM allocation in cloud data centers based on particle swarm optimization (PSO). The suggested policy intelligently distributes the virtual machines among the fewest possible physical hosts, hence reducing resource costs. According to the findings, the strategy not only reduces the number of VMs allocated to the host machines but also improves performance and scalability.

There are common pitfalls in existing evolutionary algorithms, in defining the problem-specific parameters for constrained optimization problems and static in nature, which leads to premature crossover. Liu, Zhang [90] provide a metaheuristic approach using an adaptive penalty function for workflow scheduling to enhance time constraints. When compared to existing state-of-the-art algorithms, the presented algorithms perform admirably and produce reasonable results under constraints such as time and money.

In another progress, Zhou and Yao [91] developed a revolutionary scheduling method based on teaching and learning optimization (TLBO) to cut down on energy use. It divides the VM scheduling in two, one pool of the VMs is to keep in active mode to cater for the arrival of a dynamic workload. The second pool of VMs is kept in reserve and put in low energy saving mode or sleep mode. The reserve pool of VMs allocated and deallocated based on resource demand. In a different work, the authors of [92], presented a whale optimization algorithm (WOA) based cloud framework for multi-objective VM scheduling in data centers.

Qin, Jin [93] proposed a semi-sleep mode issue in virtual machine scheduling was considered, and a plan to decrease the average latency of resource requests was offered to help preserve power in data centers. In their proposed system, the authors introduced a cost function to optimize the semi-sleep parameter using Ant Colony Optimization (ACO) and was able to reduce the cost function of the system. In another study, Xu and Li [94] anticipated the problem of calculating the total execution time of processes on a virtual machine. They considered this problem as NP-hard and introduced a learning effect based waited for the model. Their model accurately estimates the total completion time and maximum lateness minimization. The proposed schedule-based rule exhibits better near-optimal results.

In another progress, Zhao, Liu [95] investigated an improved scheduling technique to reduce the high upfront cost of the systems. The proposed dynamic bin packing model used a divide and conquer strategy with a branch and bound algorithm (DCBB) for minimizing the virtual machines on the physical servers. The method is evaluated on three different real-time workloads and also on synthetic workloads. The experimental results show its superiority over comparative techniques for execution time and fast convergence rate.

By applying a machine learning technique for load balancing, Sui, Liu [96] established an intelligent technique for scheduling of VMs in the data centers. First, the prediction is done for incoming workloads on the servers by utilizing a hybridization of genetic algorithm with the combination of Support Vector Machine (SVM) named SVR_GA. Then, to improve the local search capability Differential Evaluation (DE) based adaptive algorithm (ESA_DE) is utilized to overcome the problem of load balancing. When compared to the benchmark algorithms the proposed method overtakes in terms of energy saving by minimizing the VM migration. An intelligent Genetic Algorithm (GA) based metaheuristic technique is proposed for dynamic virtual machine scheduling for optimum resource

allocation. In this work, both memory and CPU utilization is considered equally for VM migration in the scheduling process. The work claims improvement in load balancing and resource utilization, however, the results are not mentioned in the article [97]. Similarly, Feng, Yao [98] implemented a GA-based Revivification-based prediction (ERP) model to estimate the execution time of applications on VMs. Then, another method ERPA is used to minimize the execution times for parallel and distributed application running on the optimized set of VMs. The simulation results confirm better execution time for the selected VMs.

Karthikeyan and Soni [99] proposed a hybrid GA, variable neighbourhood search (VNS) and PSO to address the VM allocation problem, improving resource utilization and minimizing completion time. However, they did not mention how this algorithm improved the parameters. A similar work proposed an ABC-based scheduling algorithm, HABC, to reduce the average makespan time of task allocation and the degree of load imbalance in the VMs. The algorithm is designed to work in both homogeneous and heterogeneous systems [100]. The fruit fly is combined with Cuckoo search to overcome the deficiency of local optima entrapment, perform better in local search, and find the optimal solution for VM mapping in the cloud data centers. The proposed method works well compared to similar techniques to reduce energy and resource leakage [101].

Rana, Abd Latiff [102] combined WOA with DA to develop VM scheduling techniques in the cloud environment. This work uses WOA as a global optimizer to generate optimal solutions. In contrast, DA is employed to replace the substandard solutions generated by WOA and improve the searching speed in the local search space. Medara and Singh [103] presented a bridging solution between workflow scheduling and VM scheduling in the data center to reduce energy consumption and resource utilization. The method uses nature-inspired water wave optimization (WWO) algorithm to find the optimal solution for VM migration on the host machines. An artificial immune-based clonal selection algorithm is modified to cope with the ever-changing cloud environment for VM scheduling. The randomized mutation operator is introduced to handle the dynamic load on the VM while scheduling. As shown in the simulation graphs the presented method showed superior performance compare to benchmark methods for energy reduction [104].

In an identical work, Chaudhury [105] put forward a metaheuristic-based scheduling algorithm for VM scheduling combining PSO and ACO. The proposed method retains the historical details of the scheduling components in its searching process. It uses it to predict the incoming load on the cloud, reducing the load imbalance on the servers. Similarly, Alsadie [106] modified the NSGA-II metaheuristic algorithm to cope with the dynamic environment of cloud scheduling. The technique works on two levels; first, the algorithm finds the optimal mapping solutions for tasks to the suitable VMs; secondly, the optimal solutions are generated for VM allocation to the best-fitted host in the data centers. The method outperforms other similar techniques but works only in a homogeneous environment.

Because recent techniques do not consider NUMA architecture while designing VM scheduling, Sheng, Hu [107] proposed multi-NUMA VM scheduling techniques by applying a machine learning approach. The authors first converted the VM scheduling problem into combinatorial optimization and then used reinforcement learning to guide the schedule per sample data. As per the result, the proposed techniques efficiently reduce the task allocation time on the host node.

Table 7
Comparison of parameters used in Virtual Machine scheduling

Reference	Response Time	Makespan	Degree of Imbalance	Waiting Time	Execution Time	Energy	Performance	Latency	Execution cost	SLA	Bandwidth
Rathor et al.[79]	√								√		
Salimi et al. [80]			√				√				
Takouna et al. [85]						√					
Knauth and Fetzer [86]						√					
Pegkas et al. [87]	√									√	
Imai et al. [88]						√					
Bazarbayev et al. [89]											
Lago et al. [84]		√				√					
Hu et al. [81]						√					
Xia et al. [82]								√			
Von Laszewski et al. [83]			√			√					
Rao and Zhou [90]				√							
Ding et al. [91]						√			√		
Quang-Hung and Thoai [92]			√			√					
Xiao et al. [145]						√					
Kim et al. [94]	√										
Zhao et al. [95]						√				√	
Wu et al. [96]					√		√				
Ebrahimirad et al. [97]		√				√					
Saravanakumar and Arun [98]											
Kim et al. [100]											√
Miao and Chen [101]							√				
Kertesz et al. [102]			√				√				
Quesnel et al. [103]					√				√		
Adhikary et al. [104]											
ho Seo et al. [105]						√					
Li et al. [106]		√									
Wu et al. [107]						√				√	
Li et al. [108]					√		√				
Xu et al. [109]											
Xu et al. [111]						√					
Lago et al. [112]			√								
Al-Dulaimy et al. [113]						√					√
Xu et al. [114]						√			√		

Reference	Response Time	Makespan	Degree of Imbalance	Waiting Time	Execution Time	Energy	Performance	Latency	Execution cost	SLA	Bandwidth
Yu et al. [115]					√				√		
Qiu et al. [129]				√			√				
Kim et al. [100]		√				√					
Xu et al. [118]						√	√				
Xing et al. [117]						√					
Hu et al. [125]			√						√		
Kumar and Raza [126]							√				
Cho et al. [123]											
Gondhi and Sharma [124]						√					
Liu et al. [127]					√				√		
Wang et al. [128]		√				√					
Qin et al. [129]								√			
Xu and Li [130]		√			√					√	
Zhao et al. [131]					√						
Sui et al. [132]						√					
Li et al. [133]			√								
Feng et al. [134]					√						
Xu et al., [119]						√					
Wan, Dang [81]		√			√						
Qi, Chen [82]		√									
Saravanakumar, Geetha [83]						√					
Xu, Xu [84]							√		√		
Karthikeyan and Soni [99]						√		√			
Kruekaew and Kimpan [100]		√							√		
Naik, Singh [101]						√					
Rana, Abd Latiff [102]						√					
Medara and Singh [103]	√		√								
Ajmera and Tewari [104]			√			√					
Chaudhury [105]						√					
Alsadie [106]						√					
SS and HS [86]				√							
Sheng, Hu [107]		√									

6 Vm Scheduling In Mobile Edge Computing

6.1 Mobile Edge Computing

Mobile edge computing (MEC), commonly known as multi-access computing or multi-access edge computing is a distributed computing ecosystem that moves processing and data storage closer to the network's edge. It has been envisaged to delegate mobile devices from running heavy and power-hungry

algorithms. Among other things, MEC is used to offload traffic off the main network, allowing operators to save money while expanding network capacity [108]. In the context of the Internet of things (IoT), MEC enables seamless integration of IoT and 5G [5].

6.2 Scheduling In MEC

In multi-access edge computing, virtual machine scheduling is essential for task offloading and resource allocations. Dynamic resource allocation using Lyapunov optimization, a decision engine and deep-reinforcement learning. Priority scheduling is when tasks are scheduled based on their priority [109, 110]. The authors of [111, 112] proposed joint offloading and priority-based task scheduling. The goal has been to reduce task completion time and the cost of edge server VM use. The same approach has been used in [113], where the authors extended further the scope to include multi-users in a narrow-band IoT environment and solved the offloading using dynamic programming techniques. Cotask offloading and schedules have been investigated in [114]. The authors formulated the problem of cotask offloading as a nonlinear program and solved it using the deep dual learning method. Similarly, Choi, Yu [115] present a deadline-aware task offloading algorithm for mobile edge computing environments. The algorithm is based on classifying tasks according to their latency requirements and offloading them to the most appropriate edge server. The algorithm is designed to minimize the overall completion time of the tasks while satisfying the deadlines and maximizing resource utilization.

Zhu, Cai [116] proposed a new approach for offloading in mobile edge computing that utilizes an improved multi-objective immune cloning algorithm. The goal of the proposed method is to enhance the efficiency of offloading by optimizing multiple objectives, including maximizing computational performance and minimizing energy consumption. This new approach aims to improve the parameters of computational performance and energy efficiency in mobile edge computing offloading. Similarly, Li, Zhang [117] put forth a jointly non-cooperative game-based offloading and dynamic service migration approach in mobile edge computing. The approach uses game theory to optimize the performance of the system by making optimal offloading and migration decisions based on limited resources such as bandwidth and computation capacity. Naouri, Wu [118] put forward a novel framework for mobile-edge computing that optimizes task offloading. The authors aim to address the challenges in offloading tasks from mobile devices to edge servers. The framework employs optimization techniques to improve the offloading decision-making process, leading to better performance and reduced energy consumption. The results show that the proposed framework outperforms existing solutions in terms of efficiency and effectiveness.

In the same vein, Cui, Zhang [119] presented a new approach to task offloading scheduling for the application of mobile edge computing. The authors aim to improve the performance and efficiency of task offloading in mobile devices by proposing a new scheduling method. The approach considers various factors such as device resources, network conditions, and service requirements to make offloading decisions. The experimental results show that the proposed method outperforms existing solutions in terms of task completion time and energy consumption. Sheng, Hu [120] proposed a computation offloading strategy for mobile edge computing. The authors aim to optimize the offloading of computationally intensive tasks from mobile devices to edge servers. The proposed strategy takes into account various factors such as network conditions, device resources, and task requirements to make offloading decisions. The results show that the proposed strategy improves performance and reduces energy consumption compared to existing solutions. Hao, Pang [121] examined a formal concept analysis approach to virtual machine scheduling in mobile edge computing. The authors aim to address the challenge of resource allocation in mobile devices when offloading tasks to edge servers. The proposed approach uses formal concept analysis to model the scheduling problem and find optimal solutions for task offloading.

Deadline-aware scheduling is another scheduling problem in which tasks are scheduled based on the time at which the task should be completed. The work of Zhu, Shi [122] addressed the problem of scheduling multiple mobile devices under a varying number of MEC servers. Lakhan, Mohammed [123] devised an algorithm for scheduling fine-grained tasks in mobile edge computing environments. The algorithm takes into account both the deadlines of the tasks and the energy efficiency of the edge servers when scheduling the tasks. The algorithm aims to minimize the total energy consumption while satisfying the deadlines of the tasks and maximizing resource utilization. The authors evaluate the proposed algorithm using simulations and results show that the algorithm outperforms existing algorithms in terms of energy efficiency and meeting deadlines. Ali and Iqbal [124] put forward a task scheduling technique for offloading microservices-based applications in mobile cloud computing environments. The technique takes into account both the cost and energy efficiency when scheduling the tasks. The technique is designed to minimize the total cost while satisfying the energy efficiency and meeting the deadlines of the tasks. The authors evaluate the proposed technique using simulations and results show that the technique outperforms existing techniques in terms of cost and energy efficiency. In the same vein, Bali, Gupta [125] take into account the priority of the tasks when scheduling tasks to offload data at edge and cloud servers. The technique is designed to minimize the total completion time while satisfying the priority and meeting the deadlines of the tasks. The authors evaluate the proposed technique using simulations and results show that the technique outperforms existing techniques in terms of completion time and meeting the priority.

Yadav and Sharma [126] developed a method for improving the sustainability of mobile edge computing through the use of blockchain technology. The presented method uses blockchain to secure cooperative task scheduling in these environments. The method aims to enhance the security of task scheduling by utilizing the decentralized and immutable nature of blockchain. The results show the improvement in security and sustainability of task scheduling in mobile edge computing. The authors of Li, Zhou [127] proposed a solution to enhance the efficiency of mobile edge computing by collaborating between User Plane Functions (UPFs) and edge servers. Their proposed algorithm, UPF selection, takes into account the current load and computing capacities of both UPFs and edge servers for optimal resource utilization. The simulation results show that this approach leads to improved system performance compared to traditional methods. In conclusion, the authors state that collaboration between UPFs and edge servers can significantly improve mobile edge computing performance A different work is presented by Lou, Tang [128] on addressing the problem of scheduling dependent tasks in a mobile edge computing environment while considering the startup latency caused by limited bandwidth on edge servers. The authors propose a novel algorithm named SDTS (Startup-aware Dependent Task Scheduling), which selects the edge server with the earliest finish time for each dependent task. The selection process considers the downloading workload, computation workload, and processing capability of the edge servers. Additionally, the algorithm employs a cloud clone for each task to utilize the scalable computation resources in the cloud. The results of simulations using real-world datasets show that SDTS outperforms

existing baselines in terms of makespan. The authors plan to further study the dependent task scheduling problem in more dynamic edge computing networks in future work.

A scheduling and resource allocation technique for Mobile Edge Computing was proposed by Kuang, Xu [129] using the opposition-based Marine-Predator Algorithm. The method seeks to optimize the scheduling of multiple workflows and the allocation of resources in the mobile edge computing setting, balancing computation load and energy consumption. The opposition-based Marine-Predator Algorithm is a combination of the marine-inspired algorithms and predator-prey algorithms, which is designed to effectively address the multi-objective optimization problem in mobile edge computing systems. Jian, Bao [130] presented a new high-efficiency learning model for virtual machine placement in mobile edge computing. The model aims to optimize virtual machine placement in a way that improves the efficiency of the system, taking into consideration various factors such as computational resources, network constraints, and other relevant variables. The authors describe how the proposed model utilizes machine learning techniques to dynamically adjust the placement of virtual machines based on real-time system conditions, resulting in a more efficient and effective mobile edge computing environment. Similarly, Hao, Cao [131] proposed a new energy-conscious scheduling method for edge computing using clustering techniques. The aim is to balance energy consumption and performance in edge devices. The method involves grouping edge devices based on their energy consumption characteristics and scheduling tasks accordingly. The results indicate that the proposed solution offers a significant improvement in energy efficiency while preserving performance compared to existing approaches. Alfakih, Hassan [132] presented a multi-objective optimization technique for resource allocation in mobile edge computing using accelerated particle swarm optimization and dynamic programming. The authors aim to improve resource utilization in edge devices by considering multiple objectives such as energy consumption, processing time, and cost. The proposed method balances these objectives to find optimal solutions for resource allocation. The results show that the proposed technique outperforms existing methods in terms of efficiency and effectiveness.

6.3 Comparing VM scheduling and MEC

Virtual Machine scheduling in cloud computing and in Multi-Access Edge Computing (MEC) are similar in that they both aim to allocate resources effectively and efficiently to multiple VMs running on a single physical host. However, there are some differences between the two which are found in the literature below.

6.3.1 Similarities

- Both focus on resource allocation: Both cloud and MEC aim to allocate physical resources, such as CPU, memory, and network bandwidth, to multiple virtual machines in a way that maximizes resource utilization and minimizes resource waste.
- Both use algorithms to schedule VMs: Both cloud and MEC use various scheduling algorithms to determine which VMs should run on which physical resources, based on factors such as priority, performance requirements, and resource availability.

6.3.2 Differences

- Scale: Cloud computing operates on a much larger scale compared to MEC, with data centers often serving thousands of users. In contrast, MEC operates at the edge of the network, closer to end-users, with fewer VMs and less overall computing power.
- Latency requirements: MEC is designed to provide low-latency services to users, whereas cloud computing is less concerned with latency. As a result, MEC often has more stringent requirements for VM scheduling and resource allocation, to meet its low-latency goals.
- Network connectivity: Cloud computing is typically located far from end-users, connected to them over a wide-area network (WAN). In contrast, MEC operates at the edge of the network, close to end-users, and is connected to them over a local area network (LAN). This difference affects the scheduling algorithms used, as well as the types of resources that are available for allocation.

6.3.3 Common Parameters In MEC

- Latency: The time taken for data to travel from the source to the destination. Low latency is critical in MEC to provide real-time services.
- Bandwidth: The amount of data that can be transmitted per unit of time. High bandwidth is necessary to support data-intensive applications.
- Computing resources: The amount of processing power, memory, and storage available at the edge. This affects the ability of MEC to support complex applications and services.
- Energy consumption: The amount of power required to run MEC services. This is a critical factor in mobile devices with limited battery life.
- Availability: The degree to which MEC services are available to users. This can be affected by network conditions, system failures, and other factors.
- Security: The measures in place to protect MEC services from unauthorized access, hacking, and other security threats.
- Cost: The economic cost of deploying and operating MEC infrastructure and services.
- Scalability: The ability of a MEC system to handle increasing amounts of data and devices over time.

6.4 Validity of the Research

The SLR analyzes (see section 3.3) the existing literature on VM scheduling and presents a taxonomy of approaches to solving virtual scheduling problems. It tries to put forward the most significant solutions in the field of scheduling technique optimization, to date. Although the authors have cautiously selected the most relevant articles in their selection and QAC processes from different reliable sources. Yet, there is a chance of threat to the validity of the job at hand; in the conduct, design and analysis phases. To avoid the biasness in the exclusion and inclusion process, the authors tried to search the maximum available literature. Even though, there is a possibility of oversight of some studies due to ambiguity in the literature, technical reports and theses. This survey's stringent

methodology serves as the study's proof of validity. (See sections 3.4 & 3.5). The dissemination of the analysis of this study will allow the researchers to effectively utilize the results.

Table 8
Online available cloud datasets

No.	Dataset/ Workload	Url Source
1	OpenCloud Hadoop workload	http://ftp.pdl.cmu.edu
2	Eucalyptus IaaS cloud workload	https://www.cs.ucsb.edu/~rich/workload/
3	Yahoo cluster traces	https://webscope.sandbox.yahoo.com
4	TU Delft Bitbrains traces	http://gwa.ewi.tudelft.nl/datasets/
5	Cloud Dataset	https://archive.ics.uci.edu
6	Public Cloud Dataset	https://www.quora.com
7	Public Cloud Dataset	https://www.kdnuggets.com/
8	SEA dataset	http://www.schonlau.net/intrusion.html
9	Greenberg Dataset	http://saul.cpsc.ucalgary.ca
10	CERIT-SC grid workload	http://jsspp.org/workload/
11	RUU Dataset	http://sneakers.cs.columbia.edu
12	Public Cloud Dataset	http://www.cloudbus.org/workloads.html
13	Purdue University dataset	https://purr.purdue.edu/publications/datasets
14	CIDD Dataset	http://www.di.unipi.it/~hkholidy/projects/cidd/
15	Cloud computing services	https://data.europa.eu/euodp/data/dataset/
16	Open Nebula	https://opennebula.org/documentation/archives/
17	Python Library Dataset	https://www.python.org/downloads/
18	Dura Cloud	https://wiki.duraspace.org/
19	Azure	https://azure.microsoft.com/en-us/resources/
20	Rackspace	https://www.rackspace.com/en-gb
21	Google Cloud Traces	https://cloud.google.com/public-datasets/

7 Future Issues And Opportunities

Despite the availability of a plethora of literature in the area of VM scheduling techniques, there remain several aspects that have not been addressed extensively and exhaustively. This is true in the case of problem formulation and the enhancement of techniques. Many authors have discussed the challenges and opportunities in this area with different aspects. Whereas, we emphasize the fundamental performance metrics and objectives of virtual machine scheduling, allocation, and deallocation of resources. Moreover, we offer our thoughts on where the state-of-the-art algorithms and methods could go and how they could be improved upon. The following sections provide further explanation.

7.1 Recourse mapping problem

In the scheduling problem, the mapping of a task to VMs, and VMs to PMs is treated as the formulation of the problem using several techniques. Notably, in the heterogeneous infrastructure, it becomes ubiquitous to examine the mapping of tasks to VMs. In general, the users are only interested to map their tasks efficiently and safely to PMs using VMs. However, the clearer distinction of the mapping at each level in the scheduling is crucial. Hence, the investigation for enhancement and development of tri-lateral scheduling techniques is an issue worth considering.

7.2 Energy-aware optimization

Although all the optimization techniques discussed in the paper are essential, however, some of the techniques were found contradictory to each other. Some of the techniques consolidate the VMs and increase physical resources when workloads increase. The other techniques de-consolidate VMs in the case of overheating and put extra constraints on the nodes. Therefore, combining these two optimization techniques seems a daunting task; to solve multi-objective problems. Existing techniques in VM scheduling use VM selection, VM placement, and VM migration methods. The selection of a method for designing a scheduling technique is crucial and needs a distinct understanding of the issue.

Moreover, in server-level scheduling, some traditional techniques are implemented to address the same problem. For example, Dynamic Voltage and Frequency Scaling (DVFS), individual level components-based scheduling - where the remaining nodes are switched off or put on sleep mode. On the network level, equipment like routers and switches are also taken into consideration, which makes all these processes more complex. At both levels, the scheduling

techniques mainly work on a static or fixed node in a controlled environment. Hence, more work is needed to explore and design efficient techniques, which can cater to both levels of the scheduling problem in a dynamic environment to support increased utilization and scalability of the recourses.

7.3 Multi-objective optimization

Almost half of the literature focuses on solving a single-objective-optimization problem as shown in Fig. 9. Generally, the works compare the research with some traditional, vague, and even obsolete techniques which seem to fall short, given the magnitude of the problems. Secondly, the majority of the mentioned works focus on more common objective functions such as; makespan, energy, response time, waiting time, execution time and load imbalance. The works either completely ignore or lay inadequate stress on other important objectives such as; availability, throughput, recovery time, fairness, SLA, utilization, and fault tolerance. Also, a major share of the literature works is done on simulation-based tools using dummy datasets rather than real hypervisors, e.g., CloudSim, Xen, Open Nebula, and KVM. These works tend to neglect the real traces in the real environment. So, it is a much-needed stance of research to instigate future researchers to come out with efficient techniques which can focus on the real cloud environment for solving multi-objective problems.

7.4 Heuristics and Mete-heuristics approach

Virtual machine (VM) scheduling is an NP-hard problem for which state-of-the-art algorithms are modified to find a good approximation to the ideal solution. That is to say, the resilience and acceptability of heuristic and meta-heuristic approaches to the scheduling problem are making their ground-breaking solutions to the problem. Many improved rule-based heuristics, e.g., First Come First Serve Minimum Completion-Time, Minimum Execution-Time, Min-min, and Max-min have been proposed to resolve the problematic issues of cloud scheduling. These algorithms produce results faster than meta-heuristics algorithms, in certain circumstances and achieve the optimal result through accuracy, completeness and speed. Furthermore, several modified and hybrid nature-inspired algorithms are proposed based on some of the modern algorithms such as GA, ACO, and PSO, which have shown significant achievement in resolving single-objective and multi-objective problems. These algorithms perform better in multi-dimensional space as compared to exact algorithms, and approximation algorithms. Still, there are more to be explored from the gems of the recently developed swarm-based mete-heuristics algorithms like the League Championship Algorithm [133], Cuckoo Search (CS) [134], Krill Herd (KH) [135], Whale Optimization Algorithm (WOA) [136, 137], Simulated Annealing (SA) [138], to name a few.

7.5 Mobile Edge Computing (MEC)

The future of MEC is expected to be characterized by increased integration with 5G networks, advanced edge AI capabilities and more efficient and secure data processing. MEC will play a crucial role in the growth of the Internet of Things (IoT) and Industry 4.0, by enabling the processing of large amounts of data generated by connected devices in real-time and providing the necessary control and feedback. The use of MEC will also drive the development of virtual and augmented reality experiences, providing low-latency processing and high-speed connectivity. Additionally, MEC will facilitate the distribution of computing resources across the network edge, enabling a more flexible and scalable solution for various computing needs. With its ability to handle sensitive data and prevent cyber-attacks, MEC is expected to provide a more secure computing environment in the future. Overall, MEC is poised to play a significant role in shaping the future of computing and communication technology.

8 Conclusion

The study presented a Systematic Literature Review (SLR) of VM scheduling techniques in cloud and mobile computing. The study follows a rigorous protocol to select the most relevant works from the literature for this study. The SLR analyzed 67 articles chosen out of 722 and presented the outcome for future researchers. The study answered three research questions as per collected data and the experience earned throughout the research. The first research question highlights the importance of VM scheduling and its possible contribution to the growth of cloud systems. The second question evaluates the performance of existing scheduling approaches in meeting the target of VM scheduling matrices. Finally, the third research question attempts to comprehend the role of VM scheduling in solving recent optimization problems and disseminates the challenges and future directions. Moreover, the SLR includes the most relevant articles addressing mobile edge computing (MEC) scheduling and analyzes the contemporary trends, similarities and differences with VM scheduling in a Cloud environment.

In addition, the study highlights the current scheduling techniques' strengths and weaknesses and classifies the possible solutions into three conventional methods: heuristics methods and metaheuristic methods. It also critically analyzes the most common performance metrics used in VM scheduling in MEC and Cloud computing. This study asserted that VM scheduling techniques in Cloud and MEC are indispensable as they let us introduce new paradigms in cloud scheduling. These developments significantly increase resource utilization, processing power, latency and network connectivity. The authors anticipate that this survey will help practitioners and academics select the most appropriate literature and utilize it as a reference point in their research to solve cloud scheduling problems.

Declarations

Authors' contributions

All authors contributed equally to the study conception, literature design, material preparation, data collection, analysis, preparation of drafts, read and approval of the final manuscript.

Funding

The authors declared that they had not received any fnancial support for this research.

Availability of data and materials

The data has been gathered from research papers and articles.

Ethics approval and consent to participate

Not applicable.

Competing interests

There are no competing interests involving the authors in the execution of this research work.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Manvi, S.S. and G.K. Shyam, *Resource management for Infrastructure as a Service (IaaS) in cloud computing: A survey.* Journal of Network and Computer Applications, 2014. **41**: p. 424-440.
2. Buyya, R. and R. Ranjan, *Special section: Federated resource management in grid and cloud computing systems.* Future Generation Computer Systems, 2010. **26**(8): p. 1189-1191.
3. Li, W., et al. *Multi-resource fair allocation with bounded number of tasks in cloud computing systems*. in *National Conference of Theoretical Computer Science*. 2017. Springer.
4. Khosravi, A., A. Nadjaran Toosi, and R. Buyya, *Online virtual machine migration for renewable energy usage maximization in geographically distributed cloud data centers.* Concurrency and Computation: Practice and Experience, 2017.
5. Qi, L., et al., *A QoS-aware virtual machine scheduling method for energy conservation in cloud-based cyber-physical systems.* World Wide Web, 2020. **23**: p. 1275-1297.
6. Parikh, S.M., N.M. Patel, and H.B. Prajapati, *Resource Management in Cloud Computing: Classification and Taxonomy.* arXiv preprint arXiv:1703.00374, 2017.
7. Wang, B., S. Jin, and B. Qin, *Batch arrival based performance evaluation of a VM scheduling strategy in cloud computing.* International Journal of Innovative Computing, Information and Control, 2018. **14**(2): p. 455-467.
8. Hou, L., et al., *Design and implementation of application programming interface for Internet of things cloud.* International Journal of Network Management, 2017. **27**(3).
9. Duan, J. and Y. Yang, *A Load Balancing and Multi-tenancy Oriented Data Center Virtualization Framework*. IEEE Transactions on Parallel and Distributed Systems, 2017.
10. Challita, S., F. Paraiso, and P. Merle. *Towards formal-based semantic interoperability in multi-clouds: the fclouds framework*. in *Cloud Computing (CLOUD), 2017 IEEE 10th International Conference on*. 2017. IEEE.
11. Aikat, J., et al., *Rethinking Security in the Era of Cloud Computing.* IEEE Security & Privacy, 2017. **15**(3): p. 60-69.
12. Uddin, M., et al., *Mobile agent based multi-layer security framework for cloud data centers.* Indian Journal of Science and Technology, 2015. **8**(12): p. 1.
13. Mousavi, S., A. Mosavi, and A.R. Varkonyi-Koczy. *A load balancing algorithm for resource allocation in cloud computing*. in *International Conference on Global Research and Education*. 2017. Springer.
14. Katyal, M. and A. Mishra, *A comparative study of load balancing algorithms in cloud computing environment.* arXiv preprint arXiv:1403.6918, 2014.
15. Rodriguez, M.A. and R. Buyya, *A taxonomy and survey on scheduling algorithms for scientific workflows in IaaS cloud computing environments.* Concurrency and Computation: Practice and Experience, 2017. **29**(8).
16. Rathore, N. and I. Chana, *Load Balancing and Job Migration Techniques in Grid: A Survey of Recent Trends.* Wireless Personal Communications, 2014. **79**(3): p. 2089-2125.
17. Abdulhamid, S.I.M., et al., *SCHEDULING TECHNIQUES IN ON-DEMAND GRID AS A SERVICE CLOUD: A REVIEW.* Journal of Theoretical & Applied Information Technology, 2014. **63**(1).
18. Beloglazov, A., J. Abawajy, and R. Buyya, *Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing.* Future generation computer systems, 2012. **28**(5): p. 755-768.
19. Uddin, M., et al., *Next-generation blockchain-enabled virtualized cloud security solutions: review and open challenges.* Electronics, 2021. **10**(20): p. 2493.
20. Ahmad, R.W., et al., *Virtual machine migration in cloud data centers: a review, taxonomy, and open research issues.* The Journal of Supercomputing, 2015. **71**(7): p. 2473-2515.
21. Qureshi, M.B., et al., *Encryption Techniques for Smart Systems Data Security Offloaded to the Cloud.* Symmetry, 2022. **14**(4): p. 695.
22. Ahmad, R.W., et al., *A survey on virtual machine migration and server consolidation frameworks for cloud data centers.* Journal of Network and Computer Applications, 2015. **52**: p. 11-25.
23. Li, Y., W. Li, and C. Jiang. *A survey of virtual machine system: Current technology and future trends*. in *Electronic Commerce and Security (ISECS), 2010 Third International Symposium on*. 2010. IEEE.
24. Zhan, Z.-H., et al., *Cloud computing resource scheduling and a survey of its evolutionary approaches.* ACM Computing Surveys (CSUR), 2015. **47**(4): p. 63.

25. Xu, F., et al., *Managing performance overhead of virtual machines in cloud computing: A survey, state of the art, and future directions.* Proceedings of the IEEE, 2014. **102**(1): p. 11-31.
26. Madni, S.H.H., M.S.A. Latiff, and Y. Coulibaly, *An appraisal of meta-heuristic resource allocation techniques for IaaS cloud.* Indian Journal of Science and Technology, 2016. **9**(4).
27. Xu, M., W. Tian, and R. Buyya, *A survey on load balancing algorithms for virtual machines placement in cloud computing.* Concurrency and Computation: Practice and Experience, 2017. **29**(12).
28. Kalra, M. and S. Singh, *A review of metaheuristic scheduling techniques in cloud computing.* Egyptian informatics journal, 2015. **16**(3): p. 275-295.
29. Madni, S.H.H., M.S.A. Latiff, and Y. Coulibaly, *Recent advancements in resource allocation techniques for cloud computing environment: a systematic review.* Cluster Computing, 2017. **20**(3): p. 2489-2533.
30. Keele, S., *Guidelines for performing systematic literature reviews in software engineering*, in *Technical report, Ver. 2.3 EBSE Technical Report. EBSE*. 2007, sn.
31. Kitchenham, B., *Procedures for performing systematic reviews.* Keele, UK, Keele University, 2004. **33**(2004): p. 1-26.
32. Charband, Y. and N.J. Navimipour, *Online knowledge sharing mechanisms: a systematic review of the state of the art literature and recommendations for future research.* Information Systems Frontiers, 2016. **18**(6): p. 1131-1151.
33. Navimipour, N.J. and Y. Charband, *Knowledge sharing mechanisms and techniques in project teams: Literature review, classification, and current trends.* Computers in Human Behavior, 2016. **62**: p. 730-742.
34. Milani, A.S. and N.J. Navimipour, *Load balancing mechanisms and techniques in the cloud environments: Systematic literature review and future trends.* Journal of Network and Computer Applications, 2016. **71**: p. 86-98.
35. Kitchenham, B., et al., *Systematic literature reviews in software engineering – A systematic literature review.* Information and Software Technology, 2009. **51**(1): p. 7-15.
36. Sharifi, M., H. Salimi, and M. Najafzadeh, *Power-efficient distributed scheduling of virtual machines using workload-aware consolidation techniques.* The Journal of Supercomputing, 2012. **61**(1): p. 46-66.
37. Prajapati, K.D., et al., *Comparison of virtual machine scheduling algorithms in cloud computing.* International Journal of Computer Applications, 2013. **83**(15).
38. Khan, M.A., et al., *Dynamic Virtual Machine Consolidation Algorithms for Energy-Efficient Cloud Resource Management: A Review*, in *Sustainable Cloud and Energy Services*. 2018, Springer. p. 135-165.
39. Bouterse, B. and H. Perros, *Dynamic VM allocation in a SaaS environment.* Annals of Telecommunications, 2017: p. 1-14.
40. Chauhan, N., N. Rakesh, and R. Matam, *Assessment on VM Placement and VM Selection Strategies*, in *Nature Inspired Computing*. 2018, Springer. p. 157-163.
41. Leelipushpam, P.G.J. and J. Sharmila. *Live VM migration techniques in cloud environment—a survey*. in *Information & Communication Technologies (ICT), 2013 IEEE Conference on*. 2013. IEEE.
42. Corradi, A., M. Fanelli, and L. Foschini, *VM consolidation: A real case based on OpenStack Cloud.* Future Generation Computer Systems, 2014. **32**: p. 118-127.
43. Patel, K.S. and A.K. Sarje. *VM provisioning method to improve the profit and SLA violation of cloud service providers*. in *Cloud Computing in Emerging Markets (CCEM), 2012 IEEE International Conference on*. 2012. IEEE.
44. Rahimikhanghah, A., et al., *Resource scheduling methods in cloud and fog computing environments: a systematic literature review.* Cluster Computing, 2021: p. 1-35.
45. Salimi, H., M. Najafzadeh, and M. Sharifi, *Advantages, Challenges and Optimizations of Virtual Machine Scheduling in Cloud Computing Environments.* International Journal of Computer Theory and Engineering, 2012. **4**(2): p. 189.
46. Hu, L., et al. *Magnet: A novel scheduling policy for power reduction in cluster with virtual machines*. in *2008 IEEE International Conference on Cluster Computing*. 2008. IEEE.
47. Xia, Y., et al. *Analysis and enhancement for interactive-oriented virtual machine scheduling*. in *2008 IEEE/IFIP International Conference on Embedded and Ubiquitous Computing*. 2008. IEEE.
48. Von Laszewski, G., et al. *Power-aware scheduling of virtual machines in dvfs-enabled clusters*. in *2009 IEEE International Conference on Cluster Computing and Workshops*. 2009. IEEE.
49. Lago, D.G., E.R. Madeira, and D. Medhi, *Energy-aware virtual machine scheduling on data centers with heterogeneous bandwidths.* IEEE Transactions on Parallel and Distributed Systems, 2017. **29**(1): p. 83-98.
50. Takouna, I., W. Dawoud, and C. Meinel. *Efficient virtual machine scheduling-policy for virtualized heterogeneous multicore systems*. in *Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA2011), Las Vegas, NV, USA*. 2011.
51. Knauth, T. and C. Fetzer. *Energy-aware scheduling for infrastructure clouds*. in *Cloud Computing Technology and Science (CloudCom), 2012 IEEE 4th International Conference on*. 2012. IEEE.
52. Pegkas, A., C. Alexakos, and S. Likothanassis. *Credit-based algorithm for Virtual Machines Scheduling*. in *2018 Innovations in Intelligent Systems and Applications (INISTA)*. 2018. IEEE.
53. Imai, S., S. Patterson, and C.A. Varela. *Uncertainty-aware elastic virtual machine scheduling for stream processing systems*. in *2018 18th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)*. 2018. IEEE.

54. Bazarbayev, S., et al. *Content-based scheduling of virtual machines (VMs) in the cloud*. in *Distributed Computing Systems (ICDCS), 2013 IEEE 33rd International Conference on*. 2013. IEEE.
55. Rao, J. and X. Zhou. *Towards fair and efficient SMP virtual machine scheduling*. in *ACM SIGPLAN Notices*. 2014. ACM.
56. Uddin, M., et al., *Power usage effectiveness metrics to measure efficiency and performance of data centers.* Applied mathematics & information sciences, 2014. **8**(5): p. 2207.
57. Quang-Hung, N. and N. Thoai. *Energy-Efficient VM Scheduling in IaaS Clouds*. in *International Conference on Future Data and Security Engineering*. 2015. Springer.
58. Xiao, P., et al., *Energy-efficiency enhanced virtual machine scheduling policy for mixed workloads in cloud environments.* Computers & Electrical Engineering, 2014. **40**(5): p. 1650-1665.
59. Kim, H., et al. *Task-aware virtual machine scheduling for I/O performance*. in *Proceedings of the 2009 ACM SIGPLAN/SIGOPS international conference on Virtual execution environments*. 2009. ACM.
60. Zhao, J., et al., *Using a vision cognitive algorithm to schedule virtual machines.* International Journal of Applied Mathematics and Computer Science, 2014. **24**(3): p. 535-550.
61. Wu, J., S. Lu, and H. Zheng. *On Maximum Elastic Scheduling of Virtual Machines for Cloud-based Data Center Networks*. in *2018 IEEE International Conference on Communications (ICC)*. 2018. IEEE.
62. Ebrahimirad, V., M. Goudarzi, and A. Rajabi, *Energy-aware scheduling for precedence-constrained parallel virtual machines in virtualized data centers.* Journal of Grid Computing, 2015. **13**(2): p. 233-253.
63. Saravanakumar, C. and C. Arun, *Efficient Idle Virtual Machine Management for Heterogeneous Cloud using Common Deployment Model.* KSII Transactions on Internet & Information Systems, 2016. **10**(4).
64. Xie, X., et al., *Design and implementation of process-aware predictive scheduling scheme for virtual machine.* The Journal of Supercomputing, 2014. **70**(3): p. 1577-1587.
65. Miao, T. and H. Chen, *FlexCore: Dynamic virtual machine scheduling using VCPU ballooning.* Tsinghua Science and Technology, 2015. **20**(1): p. 7-16.
66. Kertesz, A., J. Dombi, and A. Benyi, *A pliant-based virtual machine scheduling solution to improve the energy efficiency of iaas clouds.* Journal of Grid Computing, 2016. **14**(1): p. 41-53.
67. Quesnel, F., et al. *Advanced Validation of the DVMS Approach to Fully Distributed VM Scheduling*. in *2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications*. 2013. IEEE.
68. Adhikary, T., et al. *Energy-efficient scheduling algorithms for data center resources in cloud computing*. in *High Performance Computing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitous Computing (HPCC_EUC), 2013 IEEE 10th International Conference on*. 2013. IEEE.
69. Hu, J., et al. *A scheduling strategy on load balancing of virtual machine resources in cloud computing environment*. in *2010 3rd International symposium on parallel architectures, algorithms and programming*. 2010. IEEE.
70. Li, X., et al., *Holistic virtual machine scheduling in cloud datacenters towards minimizing total energy.* IEEE Transactions on Parallel and Distributed Systems, 2017.
71. Li, H., et al., *Energy-efficient and QoS-aware model based resource consolidation in cloud data centers.* Cluster Computing, 2017: p. 1-11.
72. Xu, X., et al., *VMSAGE: A virtual machine scheduling algorithm based on the gravitational effect for green Cloud computing.* Simulation Modelling Practice and Theory, 2018.
73. Yan-guang, C. and L. Ji-sheng, *Derivation and generalization of the urban gravitational model using fractal idea with an application to the spatial cross-correlation between Beijing and Tianjin.* , 2002. **21**(6): p. 742-752.
74. Al-Dulaimy, A., et al., *Type-Aware Virtual Machine Management for Energy Efficient Cloud Data Centers.* Sustainable Computing: Informatics and Systems, 2018.
75. Xu, H., et al., *Incentive-aware virtual machine scheduling in cloud computing.* The Journal of Supercomputing, 2018: p. 1-23.
76. Yu, C., L. Qin, and J. Zhou, *A lock-aware virtual machine scheduling scheme for synchronization performance.* The Journal of Supercomputing, 2019. **75**(1): p. 20-32.
77. Qiu, Y., et al., *Energy aware virtual machine scheduling in data centers.* Energies, 2019. **12**(4): p. 646.
78. Xing, G., et al., *Fair energy-efficient virtual machine scheduling for Internet of Things applications in cloud environment.* International Journal of Distributed Sensor Networks, 2017. **13**(2): p. 1550147717694890.
79. Xu, X., et al., *A balanced virtual machine scheduling method for energy-performance trade-offs in cyber-physical cloud systems.* Future Generation Computer Systems, 2017.
80. Xu, X., et al. *An Energy-Aware Virtual Machine Scheduling Method for Cloudlets in Wireless Metropolitan Area Networks*. in *2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*. 2018. IEEE.
81. Wan, B., et al., *Modeling analysis and cost-performance ratio optimization of virtual machine scheduling in cloud computing.* IEEE Transactions on Parallel and Distributed Systems, 2020. **31**(7): p. 1518-1532.
82. Qi, L., et al., *A QoS-aware virtual machine scheduling method for energy conservation in cloud-based cyber-physical systems.* World Wide Web, 2020. **23**(2): p. 1275-1297.

83. Saravanakumar, C., et al., *An efficient technique for virtual machine clustering and communications using task-based scheduling in cloud computing.* Scientific Programming, 2021. **2021**.
84. Xu, H., et al., *Fault tolerance and quality of service aware virtual machine scheduling algorithm in cloud data centers.* The Journal of Supercomputing, 2022: p. 1-23.
85. Mukherjee, I. and P.K. Ray, *A review of optimization techniques in metal cutting processes.* Computers & Industrial Engineering, 2006. **50**(1-2): p. 15-34.
86. SS, V.C. and A. HS, *Nature inspired meta heuristic algorithms for optimization problems.* Computing, 2022. **104**(2): p. 251-269.
87. Cho, K.-M., et al., *A hybrid meta-heuristic algorithm for VM scheduling with load balancing in cloud computing.* Neural Computing and Applications, 2015. **26**(6): p. 1297-1309.
88. Gondhi, N.K. and A. Sharma. *Local Search Based Ant Colony Optimization for Scheduling in Cloud Computing*. in *Advances in Computing and Communication Engineering (ICACCE), 2015 Second International Conference on*. 2015. IEEE.
89. Kumar, D. and Z. Raza. *A PSO based VM resource scheduling model for cloud computing*. in *Computational Intelligence & Communication Technology (CICT), 2015 IEEE International Conference on*. 2015. IEEE.
90. Liu, L., et al., *Deadline-constrained coevolutionary genetic algorithm for scientific workflow scheduling in cloud computing*. Concurrency and Computation: Practice and Experience, 2017. **29**(5).
91. Zhou, J. and X. Yao, *Hybrid teaching–learning-based optimization of correlation-aware service composition in cloud manufacturing.* The International Journal of Advanced Manufacturing Technology, 2017. **91**(9): p. 3515-3533.
92. Rana, N. and M.S. Abd Latiff, *A cloud-based conceptual framework for multi-objective virtual machine scheduling using whale optimization algorithm.* International Journal of Innovative Computing, 2018. **8**(3).
93. Qin, B., S. Jin, and D. Zhao, *ENERGY-EFFICIENT VIRTUAL MACHINE SCHEDULING STRATEGY WITH SEMI-SLEEP MODE ON THE CLOUD PLATFORM.* INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING INFORMATION AND CONTROL, 2019. **15**(1): p. 337-349.
94. Xu, H. and X. Li, *Methods for virtual machine scheduling with uncertain execution times in cloud computing.* International Journal of Machine Learning and Cybernetics, 2019. **10**(2): p. 325-335.
95. Zhao, Y., et al., *Reducing the upfront cost of private clouds with clairvoyant virtual machine placement.* The Journal of Supercomputing, 2019. **75**(1): p. 340-369.
96. Sui, X., et al., *Virtual machine scheduling strategy based on machine learning algorithms for load balancing.* EURASIP Journal on Wireless Communications and Networking, 2019. **2019**(1): p. 160.
97. Li, J., et al. *Research on Dynamic Virtual Machine Scheduling Strategy Based on Improved Genetic Algorithm*. in *Journal of Physics: Conference Series*. 2019. IOP Publishing.
98. Feng, Y., et al., *An efficient virtual machine allocation algorithm for parallel and distributed simulation applications.* Concurrency and Computation Practice and Experience, 2019. **31**(17): p. 1-22.
99. Karthikeyan, P. and R. Soni, *A hybrid PSO optimised virtual machine scheduling algorithm in cloud computing.* International Journal of Business Information Systems, 2020. **34**(4): p. 536-559.
100. Kruekaew, B. and W. Kimpan, *Enhancing of artificial bee colony algorithm for virtual machine scheduling and load balancing problem in cloud computing.* International Journal of Computational Intelligence Systems, 2020. **13**(1): p. 496-510.
101. Naik, B.B., D. Singh, and A.B. Samaddar, *FHCS: Hybridised optimisation for virtual machine migration and task scheduling in cloud data center.* IET Communications, 2020. **14**(12): p. 1942-1948.
102. Rana, N., et al., *A hybrid whale optimization algorithm with differential evolution optimization for multi-objective virtual machine scheduling in cloud computing.* Engineering Optimization, 2021: p. 1-18.
103. Medara, R. and R.S. Singh, *Energy-aware workflow task scheduling in clouds with virtual machine consolidation using discrete water wave optimization.* Simulation Modelling Practice and Theory, 2021. **110**: p. 102323.
104. Ajmera, K. and T.K. Tewari, *VMS-MCSA: virtual machine scheduling using modified clonal selection algorithm.* Cluster Computing, 2021. **24**(4): p. 3531-3549.
105. Chaudhury, K.S., *A particle swarm and ant Colony optimization based load balancing and virtual machine scheduling algorithm for cloud computing environment.* Turkish Journal of Computer and Mathematics Education (TURCOMAT), 2021. **12**(11): p. 3885-3898.
106. Alsadie, D., *A metaheuristic framework for dynamic virtual machine allocation with optimized task scheduling in cloud data centers.* IEEE Access, 2021. **9**: p. 74218-74233.
107. Sheng, J., et al., *Learning to schedule multi-NUMA virtual machines via reinforcement learning.* Pattern Recognition, 2022. **121**: p. 108254.
108. Pham, Q.-V., et al., *A survey of multi-access edge computing in 5G and beyond: Fundamentals, technology integration, and state-of-the-art.* IEEE access, 2020. **8**: p. 116974-117017.
109. Alfakih, T., et al., *Task offloading and resource allocation for mobile edge computing by deep reinforcement learning based on SARSA.* IEEE Access, 2020. **8**: p. 54074-54084.
110. Wei, X., et al. *MVR: An architecture for computation offloading in mobile edge computing*. in *2017 IEEE international conference on edge computing (EDGE)*. 2017. IEEE.
111. Mao, Y., J. Zhang, and K.B. Letaief, *Dynamic computation offloading for mobile-edge computing with energy harvesting devices.* IEEE Journal on Selected Areas in Communications, 2016. **34**(12): p. 3590-3605.

112. Gao, L. and M. Moh. *Joint computation offloading and prioritized scheduling in mobile edge computing.* in *2018 International Conference on High Performance Computing & Simulation (HPCS)*. 2018. IEEE.
113. Lei, L., et al., *Joint computation offloading and multiuser scheduling using approximate dynamic programming in NB-IoT edge computing system.* IEEE Internet of Things Journal, 2019. **6**(3): p. 5345-5362.
114. Chiang, Y.-H., et al., *Deep-dual-learning-based cotask processing in multiaccess edge computing systems.* IEEE Internet of Things Journal, 2020. **7**(10): p. 9383-9398.
115. Choi, H., H. Yu, and E. Lee, *Latency-classification-based deadline-aware task offloading algorithm in mobile edge computing environments.* Applied Sciences, 2019. **9**(21): p. 4696.
116. Zhu, S.-f., J.-h. Cai, and E.-l. Sun, *Mobile edge computing offloading scheme based on improved multi-objective immune cloning algorithm.* Wireless Networks, 2023: p. 1-14.
117. Li, C., Q. Zhang, and Y. Luo, *A jointly non-cooperative game-based offloading and dynamic service migration approach in mobile edge computing.* Knowledge and Information Systems, 2023: p. 1-37.
118. Naouri, A., et al., *A novel framework for mobile-edge computing by optimizing task offloading.* IEEE Internet of Things Journal, 2021. **8**(16): p. 13065-13076.
119. Cui, Y., et al. *A new approach on task offloading scheduling for application of mobile edge computing*. in *2021 IEEE Wireless Communications and Networking Conference (WCNC)*. 2021. IEEE.
120. Sheng, J., et al., *Computation offloading strategy in mobile edge computing.* Information, 2019. **10**(6): p. 191.
121. Hao, F., et al., *Virtual machines scheduling in mobile edge computing: a formal concept analysis approach.* IEEE Transactions on Sustainable Computing, 2019. **5**(3): p. 319-328.
122. Zhu, T., et al., *Task scheduling in deadline-aware mobile edge computing systems.* IEEE Internet of Things Journal, 2018. **6**(3): p. 4854-4866.
123. Lakhan, A., et al., *Deadline aware and energy-efficient scheduling algorithm for fine-grained tasks in mobile edge computing.* International Journal of Web and Grid Services, 2022. **18**(2): p. 168-193.
124. Ali, A. and M.M. Iqbal, *A cost and energy efficient task scheduling technique to offload microservices based applications in mobile cloud computing.* IEEE Access, 2022. **10**: p. 46633-46651.
125. Bali, M.S., et al., *An effective Technique to Schedule priority aware tasks to offload data at edge and cloud servers.* Measurement: Sensors, 2023: p. 100670.
126. Yadav, A.M. and S. Sharma, *Cooperative task scheduling secured with blockchain in sustainable mobile edge computing.* Sustainable Computing: Informatics and Systems, 2023: p. 100843.
127. Li, Y., et al. *Collaborative Mobile Edge Computing Through UPF Selection*. in *Collaborative Computing: Networking, Applications and Worksharing: 18th EAI International Conference, CollaborateCom 2022, Hangzhou, China, October 15-16, 2022, Proceedings, Part II*. 2023. Springer.
128. Lou, J., et al., *Startup-aware Dependent Task Scheduling with Bandwidth Constraints in Edge Computing.* IEEE Transactions on Mobile Computing, 2023.
129. Kuang, F., Z. Xu, and M. Masdari, *Multi-workflow scheduling and resource provisioning in Mobile Edge Computing using opposition-based Marine-Predator Algorithm.* Pervasive and Mobile Computing, 2022. **87**: p. 101715.
130. Jian, C., L. Bao, and M. Zhang, *A high-efficiency learning model for virtual machine placement in mobile edge computing.* Cluster Computing, 2022. **25**(5): p. 3051-3066.
131. Hao, Y., et al., *Energy-aware scheduling in edge computing with a clustering method.* Future Generation Computer Systems, 2021. **117**: p. 259-272.
132. Alfakih, T., M.M. Hassan, and M. Al-Razgan, *Multi-objective accelerated particle swarm optimization with dynamic programing technique for resource allocation in mobile edge computing.* IEEE Access, 2021. **9**: p. 167503-167520.
133. Kashan, A.H., et al., *The League Championship Algorithm: Applications and Extensions*, in *Handbook of AI-based Metaheuristics*. 2021, CRC Press. p. 201-218.
134. Saif, M.A.N., S. Niranjan, and B.A.H. Murshed. *Multi-Objective Cuckoo Search Optimization Algorithm for Optimal Resource Allocation in Cloud Environment*. in *2022 3rd International Conference for Emerging Technology (INCET)*. 2022. IEEE.
135. Rahumath, A.S., M. Natarajan, and A.R. Malangai, *Resource Scalability and Security Using Entropy Based Adaptive Krill Herd Optimization for Auto Scaling in Cloud*. Wireless Personal Communications, 2021. **119**(1): p. 791-813.
136. Mirjalili, S. and A. Lewis, *The whale optimization algorithm.* Advances in engineering software, 2016. **95**: p. 51-67.
137. Rana, N., et al., *Whale optimization algorithm: a systematic review of contemporary applications, modifications and developments.* Neural Computing and Applications, 2020. **32**(20): p. 16245-16277.
138. Tanha, M., M. Hosseini Shirvani, and A.M. Rahmani, *A hybrid meta-heuristic task scheduling algorithm based on genetic and thermodynamic simulated annealing algorithms in cloud computing environments.* Neural Computing and Applications, 2021. **33**(24): p. 16951-16984.

Figures

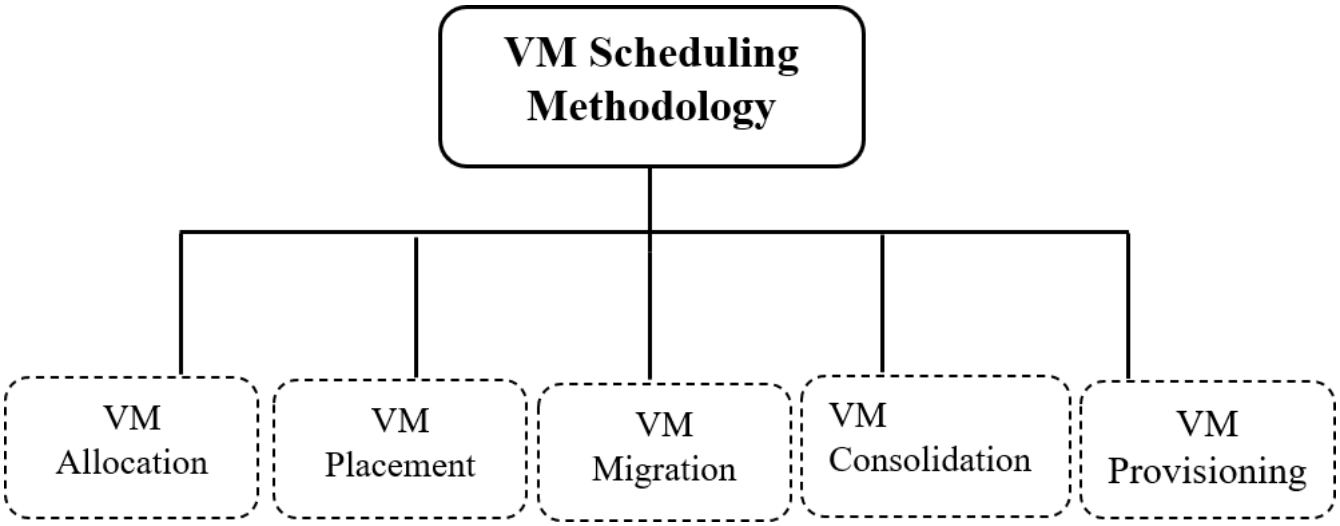

Figure 2

Classification of virtual machine management techniques

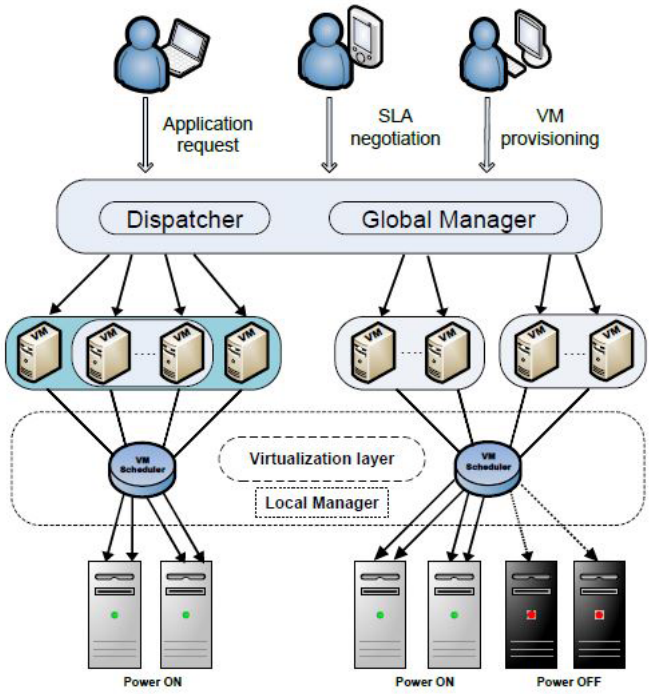

Figure 3

Virtual machine scheduling overview

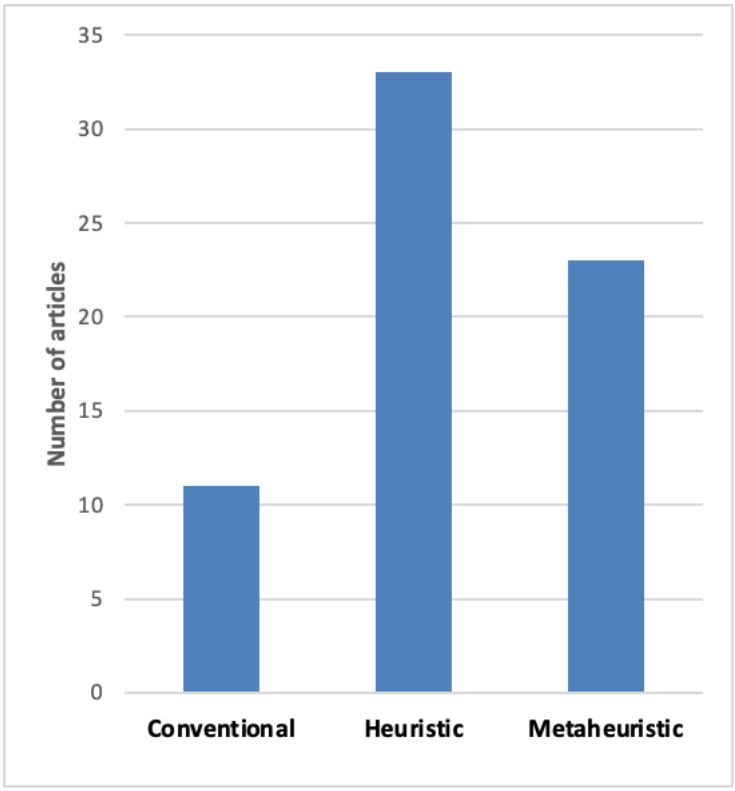

Figure 4

Distribution of articles based on the approach used in VM scheduling

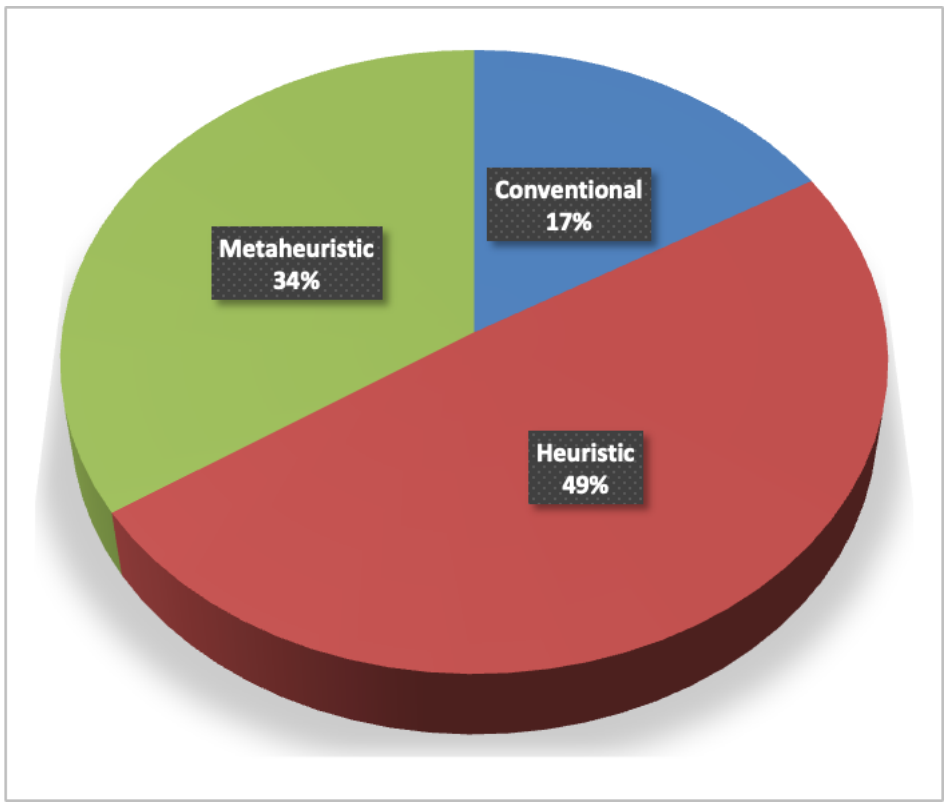

Figure 5

Percentage-wise distribution of approaches used in VM scheduling

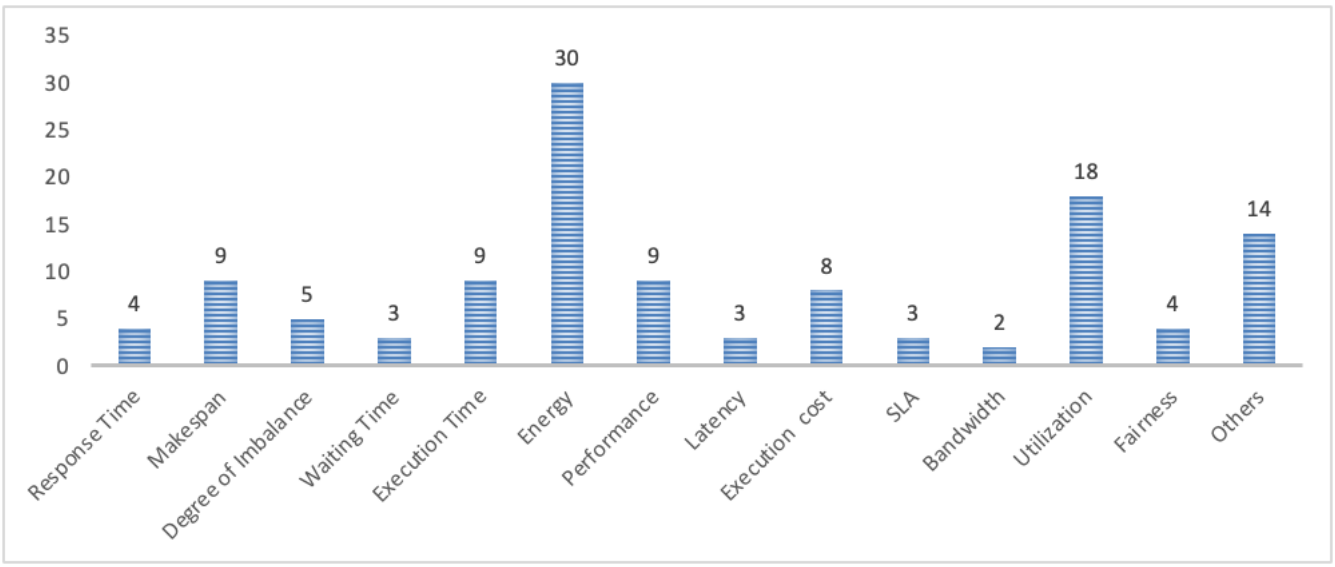

Figure 6

No. of parameters used in the reviewed literature

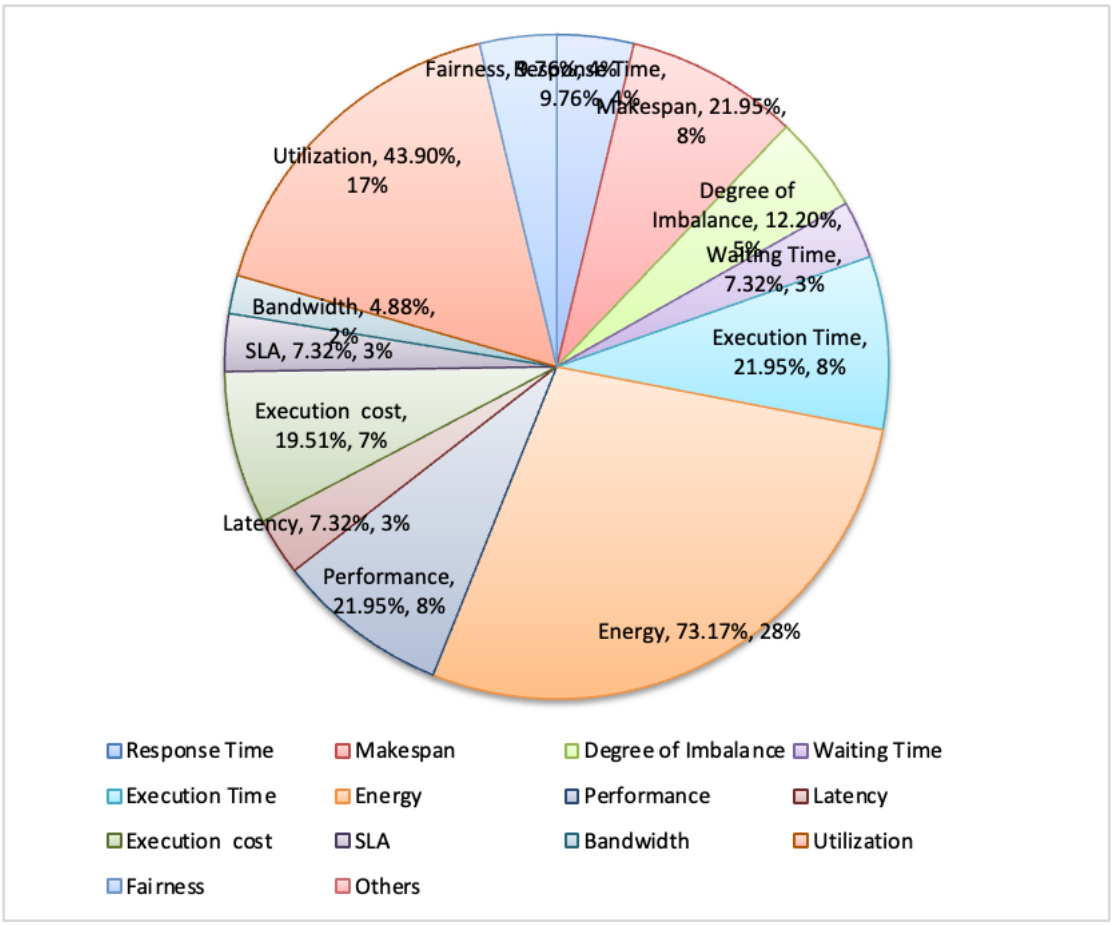

Figure 7

Percentage of virtual machine scheduling metrics in the reviewed literature

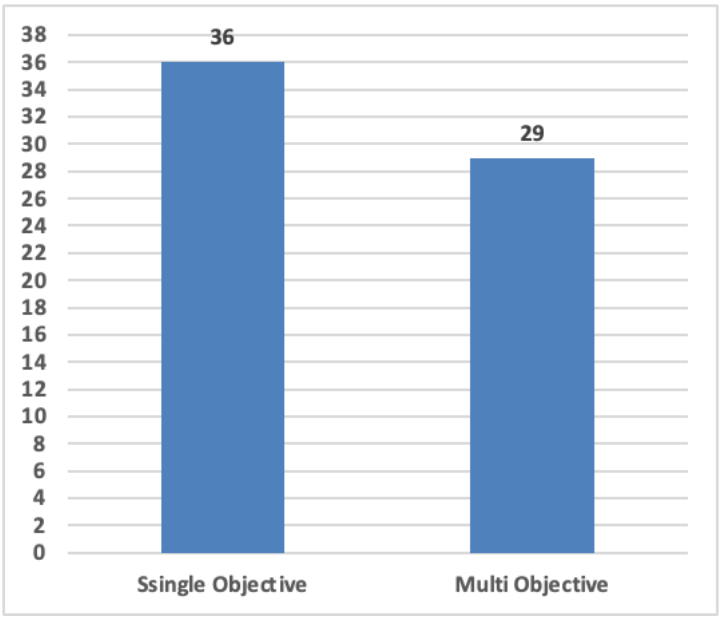

Figure 8

Comparison of single-objective and multi-objective literature in numbers

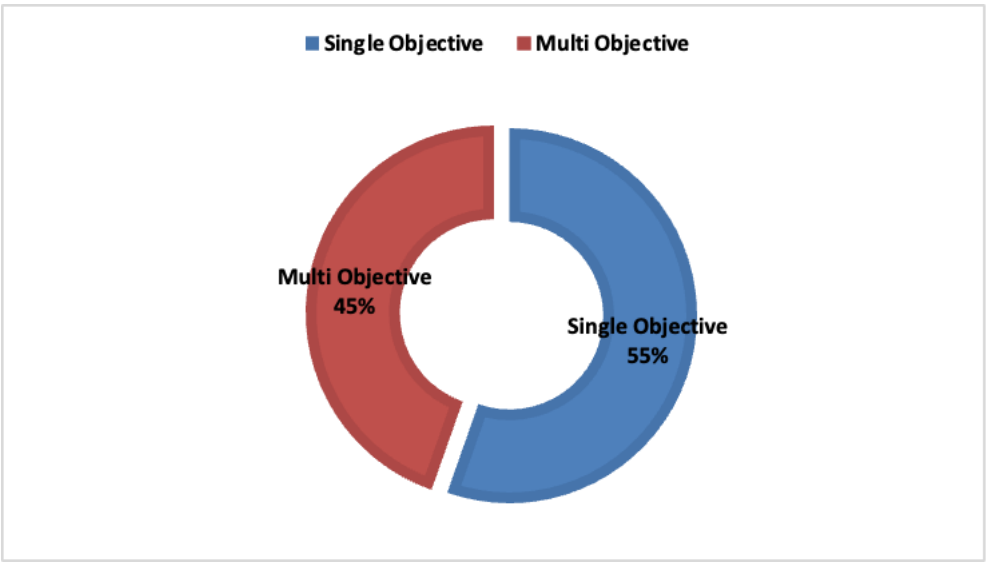

Figure 9

Percentage-based contrast between studies with one aim and those with many aims