IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 16, 2019, accepted September 25, 2019, date of publication September 30, 2019,
date of current version October 10, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2944420

Load Balancing and Server Consolidation in Cloud Computing Environments: A Meta-Study

MOHAMMED ALA'ANZY[1] AND MOHAMED OTHMAN[1,2], (Senior Member, IEEE)
[1]Department of Communication Technology and Networks, Universiti Putra Malaysia (UPM), Serdang 43400, Malaysia
[2]Laboratory of Computational Science and Mathematical Physics, Institute of Mathematical Research (INSPEM), Universiti Putra Malaysia (UPM), Serdang 43400, Malaysia

Corresponding authors: Mohammed Ala'anzy (m.alanzy.cs@gmail.com) and Mohamed Othman (mothman@upm.edu.my)

This work was supported in part by the Malaysian Ministry of Education under Research Management Centre, Universiti Putra Malaysia, Putra Grant scheme with High Impact Factor, under Grant UPM/700-2/1/GPB/2017/9557900.

ABSTRACT The data-center is considered the heart of cloud computing. Recently, the growing demand for cloud computing services has caused a growing load on data centers. In terms of system behavior and workload, patterns of cloud computing are very dynamic; and that might serve to imbalance the load among data center resources. Eventually, some data-center resources could come to be over-loaded/under-loaded, which leads to an increase in energy consumption in addition to decreased functioning and wastage of resources. Just considering energy-efficiency (that can be attained efficiently by consolidate the servers) may not be enough for real applications because it may cause problems such as unbalanced load for each Physical Machine (PM). Therefore, this paper surveys published load balancing algorithms that achieved by server consolidation via a meta-analysis. Load balancing with server consolidation enriches the exploitation of resource utilization and can enhance Quality of Service (QoS) metrics, since data-centers and their applications are increasing exponentially. This meta-study, reviews the literature on load balancing and server consolidation and presents a ready reference taxonomy on the most efficient algorithms that achieve load balancing and server consolidation. This work attempts to present a taxonomy with a new classification for load balancing and server consolidation, such as migration overhead, hardware threshold, network traffic, and reliability.

INDEX TERMS Cloud computing, load balancing, server consolidation, energy efficiency, VM live migration.

I. INTRODUCTION

Cloud computing is a current computer technology for delivering services to customers based on demand. This technology eases access to information through various devices, for instance, Smart-phones, PDAs, PCs, and tablets. Nowadays, cloud computing is considered a worldwide trend, with many advantages in three models of cloud service, namely Infrastructure as a Service (IaaS), Software as a Service (SaaS), and Platform as a Service (PaaS). Many clients, industries, and so forth are migrating their data, data processing, information, etc. onto cloud computing platforms. The resources are spread all around the world for the rapid delivery of services to the users [1], [2]. Many challenges were encountered once cloud computing first emerged such as scaling, security, QoS management, resource scheduling, data-centre energy consumption, service availability, data lock-in, and competent load balancing [3], [4]. Therefore, load balancing of the cloud's servers and the cloud's energy consumption are the main concerns in cloud computing [5], [6]. Load balancing is the process of assigning and reassigning the load among available resources in order to maximize throughput, while reducing cost, response time, and energy consumption, improving resource utilization and performance [7], [8]. On other hand, server consolidation can play a vital role in enhancing most of the above-mentioned metrics, while preserving the Service Level Agreement (SLA) and achieving the satisfaction of the end users, which could be achieved by a suitable load balancing policy. Therefore, effective server consolidation and load balancing algorithms/mechanisms can boost the success of cloud computing environments. A lot of research has been done on load balancing and server

The associate editor coordinating the review of this manuscript and approving it for publication was Kaitai Liang.

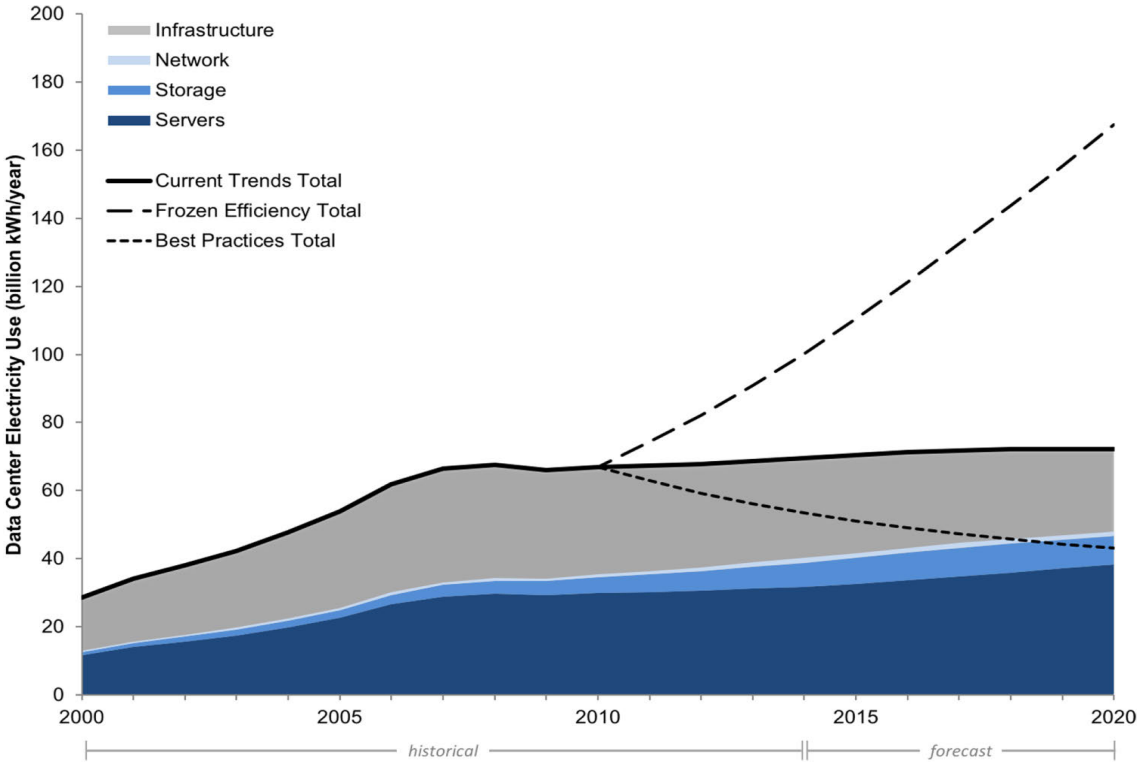


FIGURE 1. US data-centres current trends equipment [12].

TABLE 1. Physical and logical resources [10].

Physical resources	Logical resources
Storage	Bandwidth (BW)
Memory	Energy
CPU	Operating system
Workstations	Information security, protocols
Network elements	APIs
Sensors/actuators	Network loads, delays

consolidation, additionally, on task scheduling in the environment of cloud computing, however, even though cloud computing still faces many problems, load balancing is considered the main one. Cloud computing should have two goals: task scheduling and resource allocation; therefore, the result of these goals is [9]:

1) *High resource availability*
2) *Increasing resource utilization*
3) *Reduction in resource cost*
4) *Preserving the elasticity of cloud computing*
5) *Reduction of carbon emissions*
6) *Energy savings.*

Resources are collections of physical or virtual components of bounded availability within a computer structure. Any device connected is considered as a resource and also any internal component of the system is considered as resource, as listed in Table 1 [10].

In terms of energy and carbon emissions, the literature reveals that the world's data-centres consumed twice as much electricity in 2005 compared to 2000, nevertheless, the upward trend in energy consumption slowed remarkably from 2005-2010 which was due to the economic crisis, and since 2005 the industry has made more effort to improve the efficiency of data-centres and concurrently to spread virtualization technology that improves the exploitation of the data-centres [11], [12]. For instance, United States data-centres reported growth, where, 6000 data-centres consumed 61×10^9 kWh of energy in 2006, which represents 1.5% of all U.S. electricity consumption, costing $4.5 billion [13]. More recently, U.S. data-centres consumed around 70×10^9 kWh, and the energy consumption was about 2% of the total electricity consumption of the country. At the same time the data centres' workload exponentially increased [12], while the recent enhancements of the data-centres were carried out in recent years. Nevertheless, the growth of data-centre electricity in 2020 and beyond is uncertain, as illustrated in Fig. 1. The modelled trends chart indicates the past and the projected growth average of the electricity consumed for the years between 2000-2020. While the previous measures may not be enough for the data-centres in the future, if the industry does not address this issue by using an efficient optimization method, such as the successful stabilization of data-centre energy consumption, there will need to be innovations in the efficiency of the data-centres. Therefore, [14] predicts the energy consumption will reach 10300 TWh per year in 2030, based on 2010 efficiency levels. The major user of date centres is Google, and that company's facilities

represent less than 1 percent of all data-centre electricity use worldwide [11].

All of these enlargements in the energy consumption are projected. The standby (idle) and underutilized servers could be contributing significantly to energy wastage and carbon emissions. In [15] it is reported that standby servers emit 11 million tons per year of CO_2 and the total cost for standby servers is about $19 billion. Gartner research [16] reported the ratio of the unutilized servers as 18% in the huge data-centres, while the utilization of the x86 servers is even lower at 12%. These results confirmed that server utilization is in the range of between 10-30% [18]. As a result, efficient resource management can be utilized to reduce both operational costs and environmental effects (such as carbon emissions) while achieving system stability.

II. MOTIVATION

The rapid development of information technology and its variety of uses has led to the emergence of cloud computing after decades of evolution of computing facilities. Previous computing technology has had many challenges and drawbacks. Therefore, the next technology seeks to overcome or avoid those drawbacks by making new technology more extendable, advanced, and accommodatable with other technologies. It is clear that cloud computing is tied to many technologies. Examples are the Internet of Things (IoT) [19], [20], e-Health applications with cooperating Wireless Body Area Networks (WBAN) [21], big data management and Vehicular Ad-hoc Networks (VANET) [20]. The complex diversity of approaches to cloud computing and the burden of its energy needs make it challenging to narrow the whole field down to one comprehensive survey. Considering energy efficiency which can be gained from server consolidation is not enough for a real application, hence this will lead to problems such as unbalanced loads for each PM [24]. Therefore, we combined server consolidation and load balancing together to review integrated solutions. The current literature pays attention to a group of concerns regarding load balancing and server consolidation in cloud computing while we tried to identify research articles that tackle these two aspects in a more efficient manner. Each article is followed by summarizations, objectives, and testing environments, in addition to the commonly-used metrics to evaluate the techniques and a correlated taxonomy of load balancing and server consolidation. The provided taxonomy is restricted to this type of research (i.e. the meta-study), hence we have included the mutual issues of these two concepts (balancing and consolidation). Consequently, the taxonomy is necessary to provide an in-depth understanding of virtualization opportunities and challenges for future research.

III. CHALLENGES

There are several main challenges encountered in load balancing and server consolidation in the cloud computing environment, from the literature [17], [22], [23], [25]–[28]. In spite of addressing load balancing broadly in various aspects, we can say that load balancing is far from being solved perfectly, and we summarize the challenges below:

1) Virtual machine migration
   This challenge is related to two main issues, the time of migrating the service and its security. The resources should be provided once the user requests service. Meanwhile, Virtual Machines (VMs) have to migrate among servers, possibly on a remote server.
2) Cloud nodes are distributed geographically
   The algorithm of load balancing in this challenge should take into consideration the communication parameters. For instance, communication speeds, network bandwidth, cloud node distances, and the client to resource distance.
3) Centralized algorithm
   The challenge here is to avoid the single point of failure, hence an algorithm for load balancing should not be held by one node. A distributive or decentralized algorithm should be designed, because if the node that carry out the algorithm (i.e. controller node) broke down, the whole system will break down.
4) Algorithm simplicity
   The complex algorithm in terms of operation and implementation has a negative impact on the load balancing process and performance.
5) Small data-centres emerging in cloud computing
   Minimizing resources is the main focus of cloud computing, whereas small data-centres are low-cost compared with large data-centres and consume less energy. Regarding the distribution of cloud computing resources around the world, the designed algorithm should be able to achieve a satisfactory response time.
6) Energy consumption
   The designed algorithm should be able to decrease energy consumption. Therefore, the load balancing algorithm should pursue an energy-aware load scheduling methodology [29].

IV. THE PERFORMANCE METRICS MEASUREMENTS

The major load balancing metrics in cloud computing environments are as follows:

- Response time: the overall time needed by the system to serve a presented demand [31], [36], [101], [102].
- Performance: determines the system's efficiency after performing load balancing. Hence, it will check all the metrics if they are optimally satisfied or not [48], [54], [55].
- Makespan: specifies the greatest completion time or allocation time of the resources to the clients [51].
- Throughput: the rate of sending or receiving data by a node in a system, in a time unit. A high throughput is needed for better performance [31], [49], [101].
- Resource utilization: the degree of system-utilized resources. A greatest resource utilization is the desirable load balancing algorithm [46]–[48].

- Migration time: the required time to migrate a VM from PM to another. A short migration time will show better cloud system performance [46], [50].
- Scalability: the ability of the system to accomplish the load balancing algorithm, taking into consideration the number of machines or hosts [43], [45].
- The degree of imbalance: rates the imbalance among VMs [53].
- Fault tolerance: measures the ability of the algorithm to accomplish tasks consistently and appropriately even in during any arbitrary node collapse in the system [52].
- Energy consumption: the energy consumed by the nodes in the system. The server consolidation helps to reduce the number of active nodes along with load balancing, which will avoid overheating [62].
- Carbon emission: the amount of carbon released from the cloud resources [62], [91].

V. RESEARCH METHODOLOGY

To expand our comprehension of load balancing using server consolidation, a systematic literature review (SLR) was carried out, with the benchmark proposed by [39]; with a precise concentration on research associated with load balancing mechanisms. Hence, this research method originated from the field of medicine to provide a replicable research method with appropriate detail [40]–[42]. To guide the reader on why server consolidation within load balancing are necessary in cloud computing, we have chosen three research questions to tackle the vital concepts of load balancing in the cloud computing environment, as formalized in the following subsections.

A. QUESTION FORMALIZATION

The aim of the questions in this section was to clarify the essential issues and challenges along with the concept of load balancing and server consolidation in cloud computing, including performance, overloading, underloading, response time, QoS assessments, and system stability. This survey attempts to tackle the following research questions (RQs):

- **RQ1:** What is load balancing and server consolidation?
- **RQ2:** Why is it necessary along with the rapid enlargement of cloud computing?
  These two questions will demonstrate that the purpose of load balancing research has been addressed over time along with the enlargement in cloud usage.
- **RQ3:** Where should new researchers put their focus?
  This question aims to help researchers to dig deep.
- **RQ4:** How can the server consolidation attain better algorithms along with load balancing techniques?
  The objective of this question is to clarify the server consolidation and its relationship with load balancing and obtaining optimal algorithms, identifying challenges and techniques.

After defining and outlining, the needs for the research are identified (i.e., search query, research questions, selection criteria, data extraction, and quality assessment). The scope of the survey could lead to detailed ready answers for readers. This is for almost all papers that achieve load balancing along with server consolidation.

B. SURVEY PLAN AND ORGANIZATION

The research articles in this study were selected from highly reputed research journals and also chosen according to the quality assessment checklist presented in [30]–[33]. In particular, sources of research articles included IEEE, Elsevier, Springer, ACM, and Taylor and Francis, as these provided deep analysis. We started with title filtering then abstract filtering of the papers. If the abstract did not provide enough information, then the whole paper was read. Therefore, papers are included in this review based on a careful investigation of their content, as well as the papers' quality. This allows us to deliver a clear and exhaustive understanding of load balancing along with server consolidation techniques in cloud computing.

C. SEARCH QUERY

Paper review was conducted from the end of November 2017 up to September 2019. Boolean functions (OR, AND, NOT) were used, with defined strings by synonyms and alternative spellings to dig deep into hundreds of articles in this area [35]. A mix of keywords were used, as in the following query:

---

*("load-balancing"**AND**"cloud computing"**AND**("cloudsim"*
***OR** "cloud analyst" **OR** "real testbed" **OR** "simulation"))*
***OR** ("load balancing" **AND** "cloud computing")*
***OR** ("host consolidation" **AND** "cloud computing")*
***OR** ("server consolidation" **AND** "cloud computing")*
***OR** ("VM migration" **AND** "load balancing" **AND** "cloud computing")*
***OR** ("VM allocation" **AND** "load balancing" **AND** "cloud computing")*
***OR** ("VM placement" **AND** "load balancing" **AND** "cloud computing")*
***OR** ("virtualization" **AND** "cloud computing" **AND** "VM")*

After the first filtering, a re-filtering was conducted to obtain a set of papers more precisely related to the review scope, to ensure that there were no papers neglected in our review, as in the statements below:

---

*(("load Balance" **OR** "load balancing")*
***AND** ("migration" **OR** "live migration")*
***AND** ("consolidation" **OR** "server consolidation"))*

D. ELIGIBILITY CRITERIA

To be included in our survey inclusion list, articles were assessed according to the quality assessment checklist (QAC)

from [30], [33]. In this way, the list of papers in the survey met the scope of the review, since every article met the following criteria:

- Does the research paper achieve load balancing and server consolidation?
- Does the research paper obviously identify the methodology?
- Does the research methodology use available tools to re-implement (simulation or real system)?
- Is the study analysis accomplished properly?

If "**yes**", articles will be selected after meeting the following criteria:

- Every article that met the criteria listed in the keywords box will be selected first
- After filtering the article by reading the abstract, it will be listed in the final set
- Articles related to load balancing and server consolidation will be included.

Conditions protocol for the review:
Inclusions:

- An article which obviously defined how load balancing could be functional and supported in cloud computing beside the server consolidation.
- An article which is expanded on by practitioners or academics.
- An article which is available in the cloud computing domain.
- An article which is peer-reviewed.
- An article which is in English.

Exclusions:

- Duplicated articles if found
- An article which references journal articles only.
- An article which is not focused on load balancing in server consolidation of cloud computing.

E. DATA EXTRACTION AND QUALITY ASSESSMENT

Next, the data was extracted and those studies were re-capped for further analysis. A total of 921 studies were found in addition to 150 studies after a secondary search. Next, the researchers rechecked if any research articles conformed to the criteria or if any were neglected. Therefore, firstly the title phrase was read, and if the title was related to the study, abstracts and concepts which mirrored the articles' contributions were noted. Once it was found that the abstract was insufficient then the entire article was reviewed, taking into account the inclusion/exclusion criteria given above from [30]. A set of 38 research articles met our scope limitations and were identified as the primary research articles for review. Fig. 2 demonstrates the process used for picking out the papers for review. As practitioners and academics usually publish their findings in journals, conference papers were excluded.

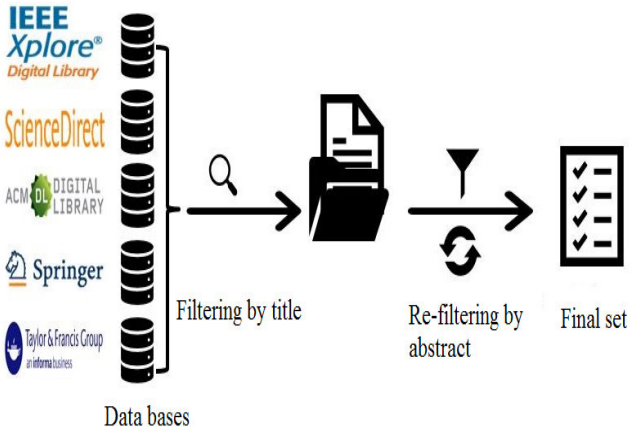


FIGURE 2. Searching steps.

VI. CONCEPTS BRIEFING REVIEW

Cloud computing is distributed worldwide, whereas the technology encountered many issues and challenges to be tackled, such as security, data-centre sprawl, performance monitoring, data lock-in, data-centre energy consumption, resource scheduling, scaling, SLA violation, etc. [3], [4]. Most concepts mentioned above are related to efficient load balancing and server consolidation. The following sections and subsections will demonstrate these two concepts in brief and their relevance in terms of metrics, challenges and the taxonomy.

A. LOAD BALANCING

Cloud computing loads are unsteady, based on users' requirements and the needs of resources. Load balancing is one of the main challenges in this area which cannot be neglected [5]. It is the process of assigning and reassigning the load among available resources in order to get better utilization to minimize the cost, energy consumption and response time [7], [8]. Load balancing organizes the workload in a perfect manner across all the resources to achieve competent resource utilization, user satisfaction, fair allocation of resources, expanding scalability, preventing over-provisioning and bottlenecks, etc. [33]. An overview of the load balancing model is demonstrated in Fig. 3. The presented model introduces some components of the data-centre such as physical components (servers) and the virtualized components (i.e VMs). We can see that tasks load balancer receives clients' demands and implements a load balancing algorithm for the tasks to allocate the demands among the VMs. The load balancer selects the suitable VM that should be allocated to the upcoming demand. The data-centre controller is responsible for task management. Hence, tasks had been submitted to the load balancer, which implements the load balancing algorithm to select the suitable VM to handle that task or set of tasks and then the balancer will preserve on PMs' balance all the time. The VM manager is responsible for VMs. The technology considered dominant in the cloud computing environment is virtualization that aims to distribute expensive hardware among VMs. A virtual machine is a software application

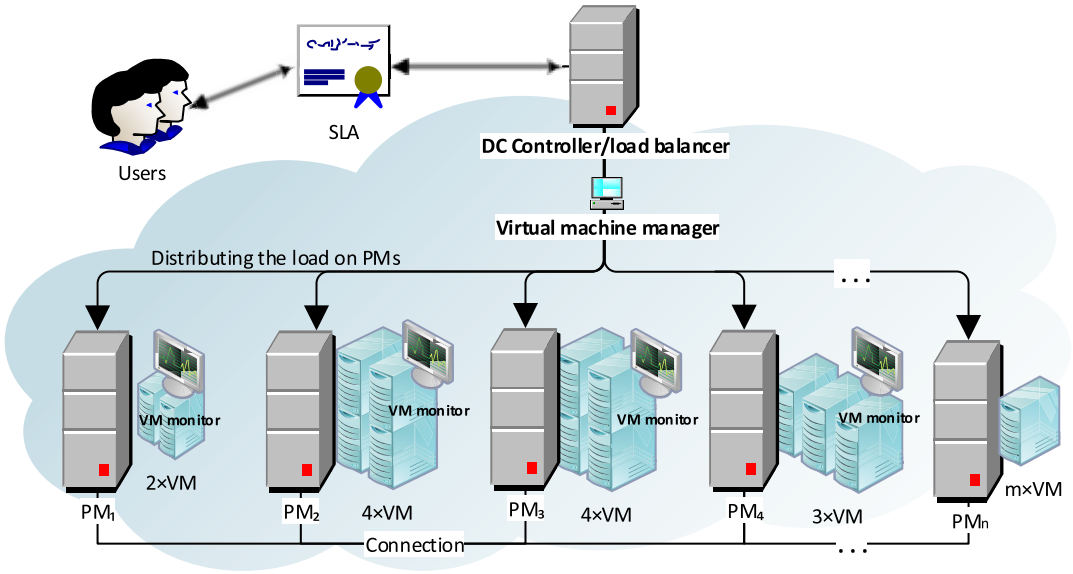


FIGURE 3. Load balancing overview.

which handles systems that allow applications to run. Cloud computing users are placed all around the world and randomly submit their demands to the VMs for processing. Thus, the assignment of the task is one of the most important concerns in the area of cloud computing, and should be taken into account to preserve the quality of service. When some VMs are idle, overloaded, or have few tasks to handle, then the QoS will be decreased which leads to user dissatisfaction, and the user will try to migrate their work to another service provider. The Virtual Machine Monitor (VMM) or "hypervisor" is used to manage and create the VMs [34]. VMM presents four procedures: provision (resume), suspension (storage), multiplexing, and live migration [37]. These procedures are essential for load balancing.

B. SERVER CONSOLIDATION

The services of data-centres are exponentially propagated. Cloud providers present their services by virtualized PMs in active virtual machines. That needs to be sold to the clients by offering a high performance and high data repository volume [103]. Meanwhile, the virtualization technology of data-centres is broadly employed to ease the management of PMs or "servers". However, this employment of PMs to VMs might affect the performance of the data-centres if carried out incorrectly. This leads to avoiding the data-centre's sprawl, energy consumption, and a large carbon footprint [15], on the other hand, this technology brings many benefits such as resource allocation, VM resizing, live migration, and server consolidation [104]. Server consolidation is widely employed to decrease the total energy consumed in data-centres as well as carbon emissions [105], [107], [108]. Furthermore, the wastage of resources is at the heart of the spread of cloud computing [106]. This leads to more energy wastage. Statistics reveal between 10% to 50% are the average server utilization levels [18]. The main attribute that made the server consolidation a prominent topic for researchers is the virtual machine live migration. The live migration is considered the best way to reduce energy consumption by reducing the number of active servers in the data-centre. Fig. 4 demonstrates a server consolidation overview by taking four servers as an example to implement VMs migration and turning off the unused servers. The virtual machine live migration has the ability to employ VMs to move among the servers with much better system downtime to avoid SLA violations, while preserving the QoS. In other words, server consolidation [32] is placing several VMs on a smaller number of PMs for enhancing resource utilization and reducing energy consumption while using a more attractive feature for the server consolidation technique (VM live migration). Hence, this feature allows a processing VM to be relocated from a PM to another without interrupting the service. The VM migration methods might differ based on parameter variations.

VII. SERVER CONSOLIDATION AND THE LOAD BALANCING WORKFLOW TAXONOMY

This section presents a taxonomy for the similar factors in load balancing and server consolidation in cloud computing. This meta-study taxonomy categorizes into static or dynamic, then based on the system model used (exact method, heuristic, and meta-heuristic), and after that based on the parameters considered to optimize the system which are hardware threshold, network traffic, migration overhead, and reliability. Refer to Fig. 5.

A. STATIC/DYNAMIC BALANCING

The static and dynamic approaches of server consolidation and load balancing have many noticeable differences. Table 2 shows the differences between them.

TABLE 2. Static and dynamic techniques.

Static technique	Dynamic technique
No migration	Migration/Live migration
Map VM-PM is not changed for a long time	Map VM-PM changes several times
Useful in consistent demand	Useful for both
Easier	More complicated
Resources over-provisioning	More balanced
No SLA violation or less violation	More SLA violation thus, many algorithms found to reduce it
Wasting for resources when PM under-utilization	In case of PM under-utilization the balancer will handle that PM
Workload is pre-predicted	Workload might be pre-predicted or post-predicted or both together
Stability in the BW among servers and VMs	More BW has to be allocated for VM migration
No ON-OFF for servers	Decreasing the life cycle for servers by on-off of the PM

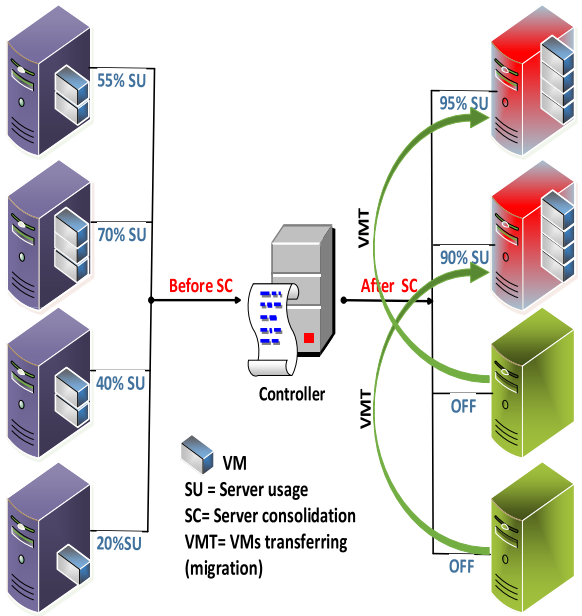


FIGURE 4. Server consolidation overview.

B. METHODS FOR SERVER CONSOLIDATION/LOAD BALANCING

The most significant aspect for given a better result possibly is the method used to distribute the VMs to the servers in the cloud data-centre. This will determine the quality of service that can be delivered to the end user as well as the cloud data-centre efficiency. This paper demonstrated the significant methods and approaches used in load balancing and server consolidation algorithms. The exact method, heuristic and meta-heuristic methods are considered as more formal techniques to reach the optimal solution. Table 3 shows the references for the aforementioned methods and parameters included in this review, organised by publication year. More details about the references of the method will be presented in subsection VII-C. Some references used two or more considered parameters in their methods.

1) EXACT METHOD

The exact method is used to detect elegant solutions for a problem. Here it is used to select the optimum assignment for a VM to a server that must be done using two methods:

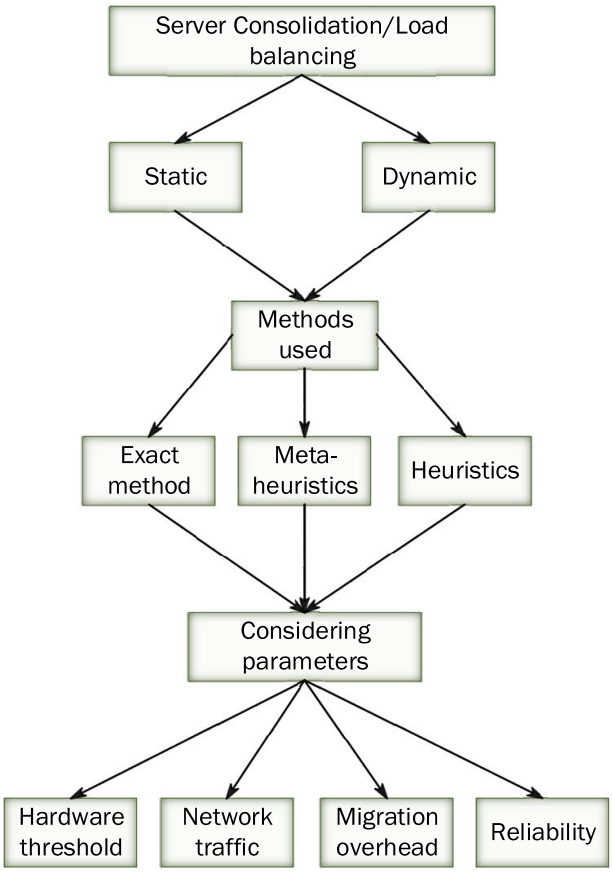


FIGURE 5. Server consolidation/load balancing taxonomy.

- A mathematical model approach
- Proposing an efficient algorithm to solve it.

This method achieves the most suitable system mapping based on specific problems occurring to solve it, where the problem is NP-hard generally [109]. This can be found in different approaches such as stochastic programming, linear programming, non-linear programming, dynamic, constraint, quadratic, and game theory. The Exact method is shown in [56], [68], [82], [93].

2) HEURISTICS METHOD

The heuristics method is used to solve a problem-dependent technique faster than classical methods once the exact method

TABLE 3. Parameters used in load balancing and server consolidation.

		Systematic criteria			
		Hardware threshold	Migration overhead	Network	Reliability
2018-2019	Heuristic	[57], [61], [64], [66], [75], [77]	[58], [59], [61], [72], [78]		[58], [71], [73], [78]
	Meta-heuristic	[63], [69], [70], [74]	[63], [65], [67], [76]	[60], [62]	[67]
	Exact method	[56]		[68]	
2016-2017	Heuristic	[80], [84], [86], [88]	[81], [85], [87]	[81]	
	Meta-heuristic	[83]	[79]		[79]
	Exact method	[82]			
2012-2015	Heuristic	[90], [91]	[92]		[92]
	Meta-heuristic		[89]		
	Exact method	[93]			

could not be achieved or it is difficult to find it. This method attempts to find a near-optimal optimized solution for a problem that gains experience from problem solving. Due to the servers' complexity (migration aspect) the heuristic optimization method reveals an acceptable solution that helped to increase system performance within a short time to solve NP-hard problems [32]. We can find this method in different approaches for bin-packing problems like (first fit decreasing (FFD), best fit decreasing (BFD), next fit, random fit, least full first, etc.). The heuristics algorithms in our systematic survey will be presented in details in subsection VII-C.

3) META-HEURISTICS METHOD

The meta-heuristics method is a problem-independent technique that can be applied to a broad range of problems. Moreover, the meta-heuristic is able to employ heuristics by guiding them over the search space in order to exploit its best capabilities to achieve better solutions. Even so, it is considered as an approximate optimization method. Thus, meta-heuristics take more time than heuristics to reach a solution [31]. We can find this method in different approaches like genetic algorithms (GA), ant colony optimization (ACO), practical swarm optimization (PSO), and hybrid optimization. Both heuristic and meta-heuristic methods aim to approach the optimal solution by various kinds of intuitions, simple solutions, or inspirations from nature and natural processes of evolution.

The following subsection gives more details for references using these methods to optimize their solutions, within the scope of our meta-study. Table 3 shows the classification of aforementioned methods.

C. PARAMETERS IN CONSIDERATION IN THE SYSTEMATIC SURVEY

1) HARDWARE THRESHOLDS

The most popular and interesting parameter given attention by researchers is the hardware threshold which deals with efficient utilization of servers' components and limitations for each cloud server e.g., memory, CPU, network, and disks. Thus, as much as an algorithm can exploit the available resources in a better manner, that can be paid back to cloud providers and the end users as well, yet an efficient resource utilization algorithm does not aim for a full utilization. Touching the threshold of every server is considered a drawback, because the server has a workload fluctuation. Therefore, an adaptive utilization threshold of the resource with a balanced workload could involve efficient exploitation of resources. Several studies have considered the hardware threshold to implement their algorithms. Server consolidation density is growing continuously at the same time, and the VMs memory and the I/O access could both have a real impact on the cloud applications' performance. The conventional approaches are unable to handle the system workload, as a result, the system performance will be degraded entirely. Therefore, [56] proposed an algorithm, namely the load-aware global resource affinity management framework (LG-RAM). The algorithm consists of three components: the shared resource load detector models, the VM resource access monitor, and the VM resource scheduler, to optimize VM consolidation performance on non-uniform memory access systems. As modern multi-core server architectures shift to non-uniform memory access, the complex interplay between data access affinity and shared resource overhead continues to pose challenges to consolidation efficiency. The [56] algorithm could outperform the state-of-the-art approaches, and optimized the memory and network I/O access affinities as well as avoiding overload on the shared resources.

[57] used a probability function to come up with an algorithm namely, self-adaptive consolidation (SAVE), for self-organizing data-centres. Based exclusively on local information, the algorithm optimizes the allocation and the migration of the VMs. A probabilistic method guarantees a suitable QoS level and averts accepting new VMs for the overloaded servers, while live VM migration is used to insure a smooth migration process. The SAVE algorithm achieved significant energy savings and increased the resource utilization, by testing the algorithm using simulation and a real testbed.

[63] proposed a smart elastic algorithm to schedule the VMs which relied on CPU and memory thresholds using a cooperative method. The smart elastic and adaptive worst fit decreasing method reduced the migration overhead. An adaptive threshold for VM migration presented by [75] was used to identify the upper and lower thresholds where the VMs were classified to three different modes based on the status of the VM resource utilization: the VM normal-load, VM overloaded, and VM underloaded. Afterwards the choice of target host was optimized based on the balance of resource usage

and transmission overhead. [84] presented a novel performance constrained framework, while the study automatically adjusts the threshold of the server based on the load intensity. This used the standard deviation of CPU utilization, and the higher rate deviation of a server could most likely reach the maximum limit of the server, which does not allow any space for workload fluctuations that will affect the SLA and the server performance. By using standard deviation, sample mean, and mean utilization, it is possible to increase the accuracy of the CPU utilization deviation.

Authors in [61], [74], [82] used the CPU and Memory utilization threshold as an objective function to balance the load or initiate migration among the PMs. They had success in attaining an efficient utilization level for the resources and reducing the energy consumption while keeping to the service level agreement. In [66], [69] authors introduced an optimization algorithm. [66] extended the first fit decreasing optimization algorithm which optimized the CPU utilization and identified the effective capacity of a PM. [69] used an inspired optimization algorithm, namely Locust, which mapped and consolidated based on the CPU threshold.

[86] presented a resource allocation which depends upon three-dimensional resources BW, CPU, and RAM. The resource scheduling process consists of three divisions: virtual resource allocation, virtual resource scheduling, and virtual resource optimization. Different objectives were achieved during these three phases.

In [64], [90] authors added to their method a monitor function, to achieve more reliability for resource load balancing and to reduce idle resources. [91] presented an optimization for the hardware threshold from the aspect of greenhouse gas and the impact of resource overutilization. The optimization tried to reduce overutilization of the resources that could cause more carbon emissions from the data-centres by introducing a new modified genetic algorithm.

2) NETWORK TRAFFIC

It is often forgotten that many researchers use resources such as CPU, memory, and storage utilization to design the cloud server problems without bearing in mind the network traffic of the data-centre components and the communications from VM to VM and server to server, which can handle and guarantee the best server to a VM. Moreover, the traffic flow of the data-centre network can be influenced by the network of the data-centre, for instance, the VMs network connection and the communication among them. Therefore, these impacts can cause degradation in the performance and the QoS. Furthermore, the network traffic has an obvious impact in the batch processing tasks and depended task, latency and a long time in communication among the nodes lead to increase the completion time, response time, and the makespan, hence, rises in the energy consumed in the data-centre. As a result, this latency in the network leads to increasing costs for the network traffic and might place the VM pairs on different racks, and the same issue might pertain in the case of servers' heavy traffic [38].

A study by [60] introduced a novel coordination approach between data-centre network topology and the communication topology, namely a virtual machine dynamic consolidation (VMDC) based on a multi-objective GA. In this approach, both objective functions and SLA constraints on communication traffic are formulated with topology awareness in switch link level. In order to improve the consolidation results based on VMDC topology awareness. This coordination is able to identify the network traffic causing bottleneck problems for different network topology types. The proposed algorithm has a real impact on reducing the energy consumption and the bottleneck communication as well as preserving the SLA constraint.

In [62], the authors improved overall network performance along with reducing carbon emissions. They introduce a hierarchical approach that consists of a two level approach to migrations depending on the packing algorithms and network communication. This was done in order to improve data-centre resource usage, especially highly utilized servers via the localised consolidation level, and to reduce the network communication latency via the network awareness level. In addition to hierarchical VM migrations which are considered as an extension approach to a community-based assignment, their analyses reveal that the initial assignment is decisive for high performance, while efficient energy consumption can be attained by the server consolidation. VM migration should be prevented if the migration within the tightly coupled communicating VMs will be done across many hops, which could have a negative impact on cloud latency.

An adaptive settable-complexity BW manager technique was introduced by [68]. They achieved live VM migration over the wireless by utilizing a multi-path TCP as the transport protocol over the 5G network. Thus, they achieved enhancement in response time, throughput, migration time, and energy consumption for the data-centre overall.

A study of communication latency of VM to VM was proposed by [81], who present a heuristic integrated solution for dynamic VM allocation and application autoscaling into a single algorithm namely, topology and application-aware dynamic VM management. Also, they take into consideration multi-VM applications, aiming to consolidate the applications on the same server, as a result, the latency among VMs will be reduced and the performance of applications will be increased as well as the VM live migration being optimized.

3) MIGRATION OVERHEAD

A vital technology for attaining server consolidation and efficient load balancing in cloud computing is VM migration. The most attractive migration is the VM live migration method which maintains the least downtime, and preserves the SLA for end users/service providers. Live VM migration's resources such as BW, RAM, and CPU should be available in both PMs to handle the migration [94]. Yet, too many live migrations happening at the same time could result in a collision [97]. Where the network BW will have more load

on the traffic flow due to the number of live migrations at a time, this could have an effect on the end users, hence in [98], [99] revealed that live migrations can consume significant BW for several seconds (500MB/s in 10s for a trivial web server). In [100], authors showed that more than 30% of the CPU requests will increase easily above the application, by the CPU overhead. Therefore, studying the migration overhead requirements, and then finding the optimum numbers of migrations will improve the cloud data-centre and performance entirely.

[58] presented a novel contribution to minimize the VM migrations by using an advanced prediction algorithm, namely an advanced prediction-based minimization of migration (APMM) algorithm. The prediction mechanism based on an existing policy namely, minimization of migration (MM), and they introduce a dynamic threshold mechanism instead of the static mechanism which sets a threshold at every time slot. The main improvement can we notice in this algorithm is the efficient usage of the workload since they use it as a history of the specific server that leads to efficiently estimate the upcoming fluctuation changes of servers' resource utilization. Thus, as large as the workload amount will be, the accuracy will be maintained. APMM shows an optimization in energy consumption as well as minimizing the VM migration.

A VM placement technique was proposed by [59] which takes into consideration the PM load (overloaded and underloaded) conditions using the heterogeneous data-centre. The algorithm, namely context adaptive self-managing VM load balancing scheme, presents a significant optimization of the energy consumption and an improvement in the overall data-centre performance. [65] proposed a game-based algorithm for VM consolidation to enhance servers' load in the data-centre while reducing the number of unnecessary VM migrations. PMs were grouped based on the number of VMs and the load is collected and predicted by gray theory, by putting the load energy as constraints, the proposed algorithm reduced the number of VMs that should be migrated. [72] presented a heuristic algorithm to find an optimal algorithm called shuffled leap frogging. The proposed algorithm outperformed on a PSO and showed a fast execution time compared with PSO and first-fit algorithms, due to reducing the number of VM migrations. [85] presented a Bayesian network algorithm for server consolidation. They classified the VM dynamic migration to nine nodes and he tried to make a connection between them by considering them as network nodes to solve some problems in the server consolidation and in predicting the suitable VM to migrate it. These nodes were linked with Bayesian networks, which helped to avoid inefficient VM migration as well as saving the energy consumption and efficiently preserving the QoS to predict the VMs overloaded/underloaded to be migrated. Also [76] presented a modified artificial bee colony with foraging behaviour for searching for overloaded and underloaded PMs which showed a better execution time.

In [87], authors introduced an instance migration cost model by using a worst-fit heuristic model. They analysed the current migration of a data-centre by using a real testbed and proposed the cost migration model, in addition to reducing the migration flow they achieved an oscillation-free consolidating service for the data-centre.

A multi-resource energy efficient model and a method of double threshold was introduced by [89] and the VM consolidation model was presented after the modified PSO algorithm in an efficient migration and consolidation algorithm with a minimum number of VM migrations. In the migration algorithm, the SLA violation has been avoided by the upper threshold in an adaptive manner based on certain requirements that avoid degradation in the system's performance.

The migration overhead might be stand-alone or combined with other parameters, because the consolidation depends on it. As in [61] and [63], their work is based on hardware threshold and the migration overhead. [61] classified the server resources based on resource ranking, while [63] scheduled the VMs based on the availability of the server resources where they achieved an optimal number of migrations. On the reliability side, the VM overhead has optimized as in [67], [73], [78], [79]. In [81] the network traffic was optimized by auto-scaling that integrates both application auto-scaling and dynamic VM allocation when they reduced the communication latency among VM2VM.

4) RELIABILITY

The reliability of hardware might have more attention from researchers than reliability of service, although both of them have a notable impact on the end user. Hardware reliability is a statement of the hardware's ability to handle its functions for some period of time [110], while hardware reliability problems can affect the following [96]:

1) System lifetime cycle (due to quick on-off) for the servers
2) Server temperature, increasing consequently (due to increased server utilization)
3) Hardware failure, which will cause [95]:
    a) Unavailability of the service
    b) SLA violation
    c) Performance degradation

On other hand, service reliability models are based mainly on the failure and availability history of service [110], yet, most cloud suppliers offer 99.99% availability for their servers [111]. That has a remarkable impact on the system quality as well as user experience. This should be taken into consideration as a parameter.

[71] proposed new metrics to implement scheduling, which is considered the first research article that depended on server failure, energy consumption, and cooling energy as performance metrics namely, failure-aware and energy efficient. These metrics could improve the system reliability and energy consumption as well as exploit the holistic operational attributes of the cloud data-centre involving server failures, computing infrastructure, and the cooling unit. Hence, they modelled the failure and the power profiles of the data-centre comprehensively.

[78] introduced models for VM migration and VM placement that deploy resources efficiently, taking into consideration the diversity of the clients' QoS demand, since they considered the constraints of clients' QoS demand (e.g. deadlines and budget) during the VM migration from server to server. By coming up with a novel heuristics-based energy aware model that achieves more stability for the data-centre and reduces the number of VMs to be migrated. [67] presented a mechanism for automating VMs migration by means of hybrid decision-making that decreases the migration probability with increase in the access probability while minimizing the downtime experience of users and boosting the PMs' balancing, thus increasing the users' satisfaction.

[73] studied the model of host overload threshold selection by using Markov decision processes of the virtual machine, hence his algorithm's success in finding the optimum overload threshold that enhances the resource utilization of the data-centre. Algorithms such as PSO were used and modified in [79]. Also, authors investigated the trade-off between the QoS of the cloud provider and energy consumption. This was to enhance data-centre service reliability and to improve the QoS while enhancing the energy consumed by the data-centre. Moreover, adaptive resource provisioning was presented in [92] using variable item size bin packing that enhances the number of active servers to support green computing. This is because the algorithm uses virtualization technology to distribute the resources dynamically while guaranteeing a stable data-centre workload.

VIII. SUMMARIZATION

The following Table 4 gives a brief summary of the parameters in consideration in section VII-C with more details such as the algorithm, methods and techniques, technique goals, and the testing environments. The table is sorted in descending order from 2019 to 2012.

IX. DISCUSSION

This section presents an analysis of the reviewed aspects of this meta-study of load balancing and server consolidation. Hence, they are the current issues in cloud computing, once we attain the maximum utilization of the resources in a competent manner that will lead to improving resource utilization, energy consumption, and reducing carbon emissions for efficient cloud computing servers, while maintaining the QoS and SLA violations. Load balancing and server consolidation should both cooperate to attain their targets and a suitable algorithm for both techniques should be considered by researchers. Furthermore, server consolidation is not enough for a real application which causes problems such as unbalanced loads for each PM and which needs an efficient method to attain this from different considered parameters and constraints such as the hardware thresholds, network traffic, reliability, and migration overhead.

This review briefly covered issues in load balancing with server consolidation (importance, challenges, metrics, a workflow taxonomy, statistical charts, and open issues).

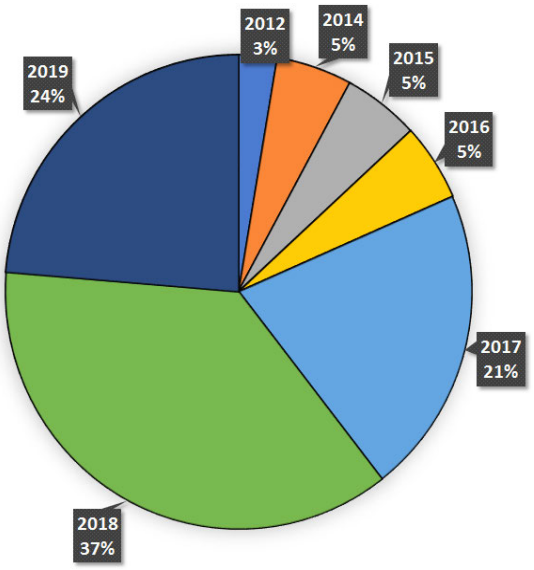


FIGURE 6. Articles distribution percentage over time.

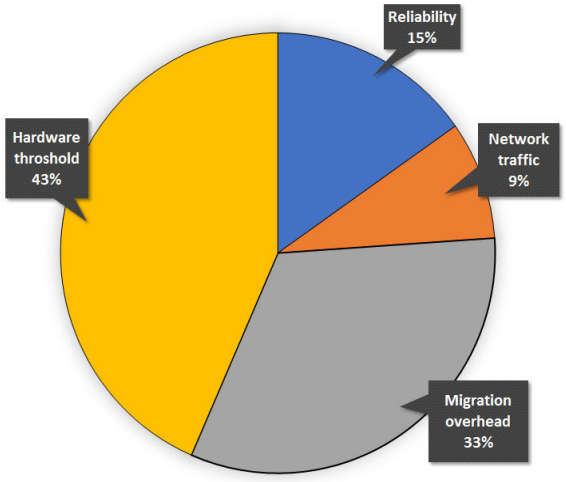


FIGURE 7. Percentage of parameters under consideration used in the review.

Different state-of-the-art methods were reviewed in this meta-study. The review classified and clarified the methods using deep analysis of over 38 research articles from a defined search query and selection of the papers via eligibility criteria. Also, the selection was aided by answers from an exploratory survey covering the years 2012-2019 for the surveyed papers. The percentage of the papers' distribution over time is shown in Fig. 6 which reveals that papers concentrating on load balancing by server consolidation have increased significantly since 2017.

The statistical chart in Fig. 7 shows the percentage of the parameters in the papers in this review and the largest number of considered parameters. The hardware threshold parameter is significantly discussed, and network traffic is the least commonly mentioned parameter in the review.

The presented taxonomy is layered based on the matched components between load balancing and server

TABLE 4. Summarization of the research techniques.

Year	Reference	Algorithm	Goal	Approach and technique	Testing environment
2019	Jianmin Qian *et al.* [56]	Load-aware global resource affinity management framework (LG-RAM)	• To optimize the memory and network I/O access affinities as well as averting overload on the shared resources • To optimize cloud workloads consolidation performance • To improve the performance on different platforms.	• Systematically characterizes the performance impacts of server consolidation on non-uniform memory access (NUMA) systems with various cloud applications • Proposing an algorithm (LG-RAM) that consists of three components: the shared resource load detector models, the VM resource access monitor, and the VM resource scheduler.	Real testbed
	Wenxia Guo *et al.* [57]	Self-adaptive consolidation (SAVE)	• To schedule all requests non-preemptively • Subjecting to constraints of PM capacities and running time interval spans • To minimize the total energy consumption for all PMs.	Makes decisions on the assignment and migration of VMs by probabilistic processes and is based exclusively on local information.	Simulation (CloudSim)/ Real testbed
	Srimoyee Bhattacherjee *et al.* [58]	Advanced prediction-based minimization of migration (APMM) algorithm	Reducing the data center energy consumption.	Uses a prediction mechanism that has been adopted on the existing Minimization of Migration (MM) policy for large history data set, followed by dynamic thresholding mechanism in place of static thresholds.	Simulation (CloudSim)
	Ashwin Kumar and B. Annappa [59]	Context adaptive self-managing VM load balancing scheme.	• Improving the performance • Saving the energy.	Context-aware adaptive heuristic-based solution for the virtual machine (VM) placement optimization in the heterogeneous cloud data centers.	Simulation (CloudSim)
	Guangyi Cao [60]	Topology-aware multi-objective virtual machine dynamic consolidation for cloud data-centre based on GA	• Introduce a novel approach for VM dynamic consolidation (VMDC) • The approach has the ability to identify the bottlenecks caused by communication traffic according to the topology awareness • Improve the energy consumed and the communication bottleneck while preserving the SLA constraints.	Multi-objective genetic algorithm.	Simulation (Network-CloudSim)
	Mahammad Mekala and P Viswanathan [61]	Energy-efficient virtual machine selection based on resource ranking	• To enhance the resource famine by reducing it. • Enhancing the energy consumption while preserving the QoS and the SLA.	An energy-efficient resource ranking and utilization factor-based virtual machine selection (ERVS).	Simulation (CloudSim)
	Sonja Filiposka *et al.* [62]	Multidimensional hierarchical VM migration management	• Maintaining high communication performance • Improving network performance • Reducing the carbon emissions • Improving the energy consumption of the HPC data-centre.	A two-level approach • Localised consolidation • Network awareness.	Simulation (CloudSim)
	Heba Nashaat *et al.* [63]	Smart elastic scheduling algorithm for VM migration	• Reducing the amount of transferred data during migration • Reducing number of migrations • Reducing performance degradation • Reducing SLA violations.	Cooperative approach • A Smart Elastic • An Adaptive Worst Fit Decreasing Virtual Machine Placement.	Simulation (CloudSim)
	Chao-Tung Yang *et al.* [64]	An energy-efficient cloud system with novel dynamic resource allocation methods	• Decreasing Idle resources • Saving the energy consumption • Monitoring the power distribution unit.	Dynamic resource allocation and Live VMs migration	OpenStack cloud platform
2018	Liangmin Guo *et al.* [65]	A Game Based	• To reduce the energy consumption • Balancing the resource load of servers without unnecessarily increasing the number of VM migrations.	• Load predicting based on gray theory • Grouping the online servers based on the number of VMs and the future load • Based on the grouping the destination server will select • Game-based virtual-machine consolidation will trigger from PMs source to PMs destination.	Simulation (CloudSim)
	Markus Hähnel *et al.* [66]	Extended First Fit Decreasing (FFD) algorithm	• To identify a computing node's effectiveness capacity • Performs superbly for a large number of services • To optimize the server CPU utilization.	Extending the Cutting Stock Problem for Consolidating Services with Stochastic workloads.	Real testbed
	Ronghui Cao *et al.* [67]	A hybrid decision-making mechanism for automating the migration of VMs	• Designed to increase user satisfaction • Reducing the down-time experience of users and increasing the server balance • Reducing the VMs migration probability with high access probability.	Extending the original VM migration by hybrid decision mechanism and designed a multi objective monitoring system.	OpenStack cloud platform
	Enzo baccarelli, *et al.* [68]	An adaptive settable complexity BW manager (SCBM) for the constrained minimization of the network energy	• Achieving the live migration of VMs over wireless (possibly, mobile) 5G FOGRAN connections • To utilize the Multi-Path TCP (MPTCP) as the transport protocol • To guarantee hard constraints on the downtime by proposing SCBM and overall migration time • Achieving rapid responsiveness to (possibly, unpredictable) fading and/or mobility induced abrupt changes of the state of the underlying MPTCP connection.	• Live Migration of VMs Over MultiPath • TCP/IP 5G Connections.	Simulation (Matlab)
	Heba Kurdi *et al.* [69]	A Locust-Inspired Scheduling Algorithm	• Aimed to support green cloud computing • Reducing the data-centre energy consumption • To improve resource utilization • To increase the number of active servers • To improve the response time and robustness as well as reduce the SLA violations.	Mapping and migration of the VMs based on the threshold.	Simulation (CloudSim)
	Domanal and Reddy [70]	A novel efficient cost optimized scheduling for Spot instances	• To enhance the response time and execution time • To maintain the load on the virtual machines • Actual and predicted spot prices perfectly correlated with error rate • To reduce the cost compared to on-demand instances.	• Modified Particle Swarm Optimization • While the prices predicted by using artificial Neural Network.	Simulation (CloudAnalyst simulator and Python based simulator)

TABLE 4. *(Continued)* **Summarization of the research techniques.**

Year	Reference	Algorithm	Goal	Approach and technique	Testing
2018	Xiang Li, *et al.* [71]	Two failure-aware energy-efficient scheduling	• Design new metrics to schedule based on (energy, cooling energy, and system failure) • Reducing the energy consumed • Improving system reliability.	Holistic energy.	environment Simulation (CloudSim)
	G. Ganesh Kumar and P. Vivekanandan [72]	Shuffled leap frogging algorithm (SLFA)	• Improving overall execution time • To enhance the amount of migration and energy consumed, compared with PSO and first fit algorithms.	Heuristic based migration.	Simulation (CloudSim)
	Zhihua Li [73]	An adaptive overload threshold selection process	• Studying the host overload threshold selection model • The algorithm has the ability to find the optimum overload threshold, hence this will lead to enhance the energy efficiency and the resource utilization of the data-centres.	Markov decision processes of VM.	Simulation (CloudSim)
	Montassar Riahi and Saoussen Krichen [74]	An efficient framework for VM placement	• To minimize power consumption • To minimize total resource • To distribute the workload efficiently across the cloud servers • To predict the future workloads of the applications.	A multi-objective GA and Bernoulli simulation.	Real testbed
	Chunmao Jiang *et al.* [75]	Adaptive threshold migration (ATM)	• ATM determines the virtual machine to be migrated and optimizes the choice of target host based on the balance of resource usage and transmission overhead • ATM algorithm reduces power consumption.	Three-way decisions of virtual machine migration by modelling the energy consumption of the host from the perspective of resource usage.	Simulation (CloudSim)
	Zhihua Li *et al.* [76]	An energy-aware dynamic virtual machine consolidation (EC-VMC)	• Improving the energy consumption • Reducing VM migrations • As a result improving overall QoS of data-centre.	Optimized search based on artificial bee colony foraging behaviour.	Simulation (CloudSim)
	Milad Ranjbari and Javad Akbari Torkestani [77]	Learning automata overload detection (LAOD)	• Proposing a Power and SLA efficient resource allocation algorithm • Optimizing energy consumption • Reducing number of VM migrations • Reducing SLA violations.	• Using learning algorithms based on environmental signals • Random learning automata approach.	Simulation (CloudSim)
	Kamran and Babar Nazir [78]	QoS-aware VM placement	• Reducing SLA violations • Demonstrating rigorous reduction in power consumption • Reducing number of VM migrations.	A novel heuristics-based energy aware resource allocation to allocate the user's tasks in the form of cloudlets to the cloud resources.	Simulation (CloudSim)
2017	Hongjian Li *et al.* [79]	Energy-efficient and QoS-aware model	• Improving QoS and in the same time reduced energy consumption • Designed to investigate the trade-off between energy consumption and QoS.	Based on QoS trade-off with energy consumption constraints then combine it with particle swarm optimization (PSO) but sitting the energy consumption per QoS as an objective function for consolidation.	Simulation (CloudSim)
	Ashkan Paya and Dan Marinescu [80]	Energy-Aware Load Balancing and Application Scaling	• Identify optimal regimes for cloud servers • Preventing SLA violations.	• A server operating in the undesirable-low regime has been migrated and then switched to a sleep mode • On the other hand, the idle server switches it to a sleep mode too and reactivates the sleeping servers during a high load • Migrating the virtual machines from an overloaded server.	The simulation experiments conducted on Amazon cloud
	Tighe and Bauer [81]	Topology and Application Aware Dynamic VM Management	To reduce VM-to-VM communication latency; hence, the concentration is on trying to contain applications within the same racks.	A rule-based heuristic auto-scaling approach integrates both application auto-scaling and dynamic (VM) allocation into a single algorithm.	Simulation (DCSim)
	Konstantinos Tsakalozos *et al.* [82]	Live VM Migration Under Time-Constraints	• Improving the utilization of physical machinery • Reducing SLA violation • Reducing the cost by half.	A Real-time scheduling of live VM migrations in large share-nothing IaaS clouds, hence, it is empowered by the combined use of a network of Brokers and the MigrateFS file system, a scalable distributed network of brokers that oversees the progress of all ongoing migration operations within the context of a provider.	Real testbed/ Simulation (other)
	Rui Li *et al.* [83]	A Pareto-based Multi-Objective VM re-Balance solution (MOVMrB)	Optimizing the load balancing among different dimensional resources within an individual host machine (HM) as well as among various HMs.	Applied hybrid VM live migration to speed up VM placement. Find solutions that leverage the inter-HM and intra-HM loads and applies a multiple objective optimization strategy with hybrid VM live migration.	Simulation (CloudSim)
	Suresh and Sakthivel [84]	A novel performance constrained power management load balancing framework for cloud data-centres	• Optimizing power • Fewer SLA violations • Guaranteed performance • Optimizing the number of VMs • Optimizing the cost of setup time in data-centres.	An innovative self-adapting mechanism to address the mismatch between a server's energy-efficiency characteristics and the behaviour of server-class workloads.	Simulation (other)
	Zhihua Li *et al.* [85]	Bayesian network-based virtual machine consolidation	• Improving the QoS • Avoiding inefficient VM migration • Saving energy.	A developed Bayesian network-based estimation model (BNEM) for live VM migration, that takes into consideration 9 factors in the real data-centre as constraints.	Simulation (CloudSim)
	Wei Zhu *et al.* [86]	A three-dimensional virtual resource scheduling algorithm	• Reducing the power consumption • Minimizing SLA violation and saving the load balancing.	Bin packing problem based heuristics virtual resource allocation.	Simulation (CloudSim)

TABLE 4. *(Continued)* **Summarization of the research techniques.**

Year	Reference	Algorithm	Goal	Approach and technique	Testing environment
2016	Huining Yan *et al.* [87]	Cost-Efficient Consolidating Service for Aliyun's Cloud-Scale Computing	• Analyzing the current migration cost in "Aliyun" • Proposing the migration cost model • To achieve cost-efficiency • To achieve load balancing • To achieve oscillation-free consolidating service.	Instance migration cost model by using the Worst-Fit heuristic.	Real testbed
	Jukka Kommeri *et al.* [88]	A prototype system for load-based management	Improving the total energy efficiency when using the method together with resource overbooking and heterogeneous hardware.	Developed prototype system for load-based management of virtual machines for packing idle virtual machines into special park servers.	OpenStack Cluster (Real hardware) / Simulation (CloudSim)
2015	Hongjian Li *et al.* [89]	Energy-efficient migration and consolidation algorithm	• To improve energy efficiency • To reduce the number of active servers • To reduce VM migrations.	• A multi-resource energy efficient model • Designing a method of double threshold with multi-resource utilization to trigger the migration of VMs • The Modified Particle Swarm Optimization method is introduced into the consolidation of VMs.	Simulation (CloudSim)
	Gutierrez and Ramirez [90]	Agent-based	• Monitoring and balancing different workload types in a distributed manner • Balancing workloads across a set of commodities, heterogeneous hosts in a cooperative way.	A collaborative agent-based problem solving technique (heuristically).	Simulation (Others)
2014	Fereydoun *et al.* [91]	Carbon aware multi-level grouping genetic algorithm GA (MLGGA)	To reduce the carbon footprint of the cloud data-centres in a distributed cloud over the data-centre network as well as the energy consumption.	A new genetic algorithm (GA) is optimizing a new genetic algorithm to be a multi-level grouping that is implemented for multi-level bin packing problems (a new heuristic optimization algorithm).	Simulation (Matlab)
	Song *et al.* [92]	An Adaptive resource provisioning using online bin packing	• The algorithm excels in hot spot mitigation and load balancing • Reconciling the demands on both green computing and stability.	Variable Item Size Bin Packing (VISBP) that uses virtualization technology to allocate data-centre resources dynamically based on application demands and supporting green computing by optimizing the number of servers actively used.	Simulation (trace driven)/ Real testbed
2012	Lovász *et al.* [93]	Performance trade-offs of energy aware VM consolidation	Saving energy in the data-centre.	Applying two approaches: virtualization and consolidation for saving energy consumed in the data centre; nevertheless, these approaches are limited usually to a single application in a homogeneous server.	Mathematical/ Simulation (Desmo-J)

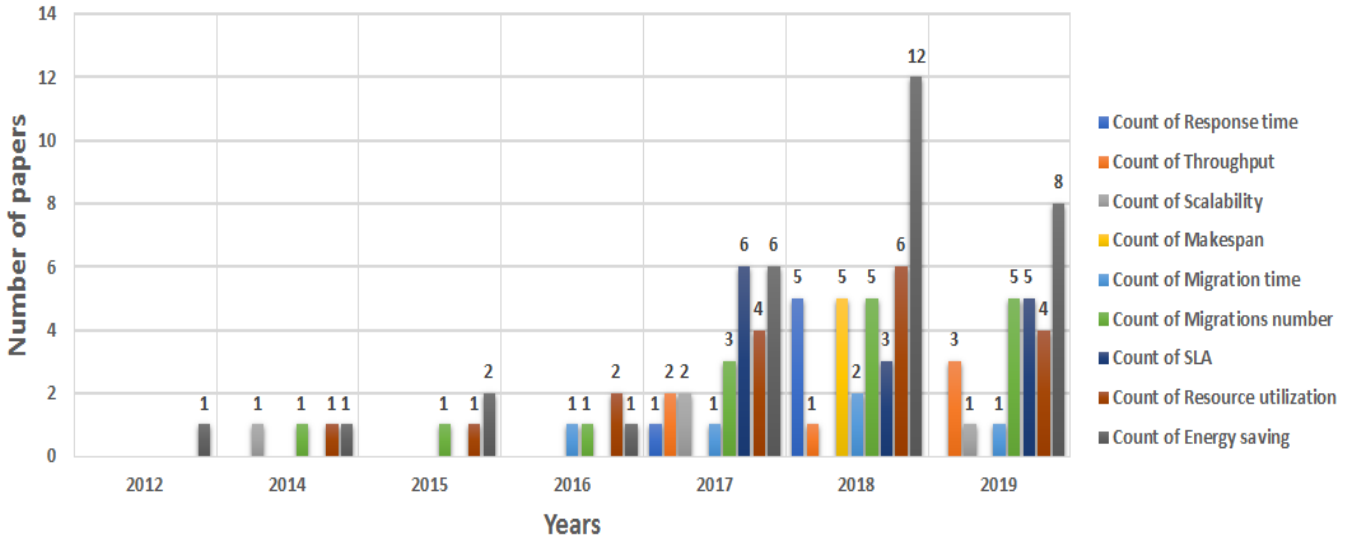


FIGURE 8. **Number of studies' metrics in reviewed techniques.**

consolidation, while paying the most attention to the parameters and classifications in the taxonomy. All of the parameters achieved their targets with some limitations.

Network traffic as well as the reliability of the cloud data-centre, needs more attention by researchers, hence migration, consolidation, and decentralized system depends on the network in general and any delay in the network may affect cloud systems drastically. It is considered the vital core in cloud system, and all the communication and latency to send or receive data depend on the network traffic. Creating fast paths to send tasks to data-centres using optimized network trees and merging new technologies such as 5G could provide a super-fast network flow.

The metrics measured in the reviewed papers are varied according to different research methods. Fig. 8 illustrates the highlighted metrics in most of the papers within our scope (i.e. load balancing and server consolidation). 9 popular metrics used by the authors in this area were selected, with a frequency count for each metric along with the number of papers that used them. This was to show the significant

TABLE 5. QoS assessment used in the reviewed techniques.

Reference	Response time	Throughput	Makespan	Resource utilization	Scalability	Migration time	Energy saving	SLA	Migrations number
[56]		✓							
[57]				✓		✓	✓		
[58]							✓	✓	✓
[59]					✓		✓	✓	✓
[60]		✓					✓	✓	
[61]				✓			✓	✓	✓
[62]		✓					✓		
[63]				✓			✓	✓	✓
[64]				✓			✓		
[65]				✓			✓		✓
[66]	✓		✓				✓		
[67]	✓			✓					✓
[68]	✓	✓				✓	✓		
[69]	✓			✓			✓	✓	
[70]	✓		✓	✓					
[71]			✓				✓		
[72]			✓				✓		✓
[73]				✓			✓		
[74]				✓			✓		
[75]							✓		
[76]			✓			✓	✓		✓
[77]							✓	✓	
[78]							✓	✓	✓
[79]	✓	✓					✓		✓
[80]				✓	✓		✓	✓	
[81]					✓		✓	✓	✓
[82]		✓				✓		✓	
[83]				✓					
[84]				✓			✓	✓	
[85]				✓			✓	✓	✓
[86]				✓			✓	✓	
[87]				✓		✓			✓
[88]				✓			✓		
[89]							✓		✓
[90]				✓					✓
[91]							✓		
[92]				✓	✓				✓
[93]							✓		

and insignificant metrics focused on by authors. Both energy and resource utilization are the top most highlighted measurements. Scalability is much less used in load balancing and server consolidation which is related to the system's ability to accomplish the load balancing algorithm within the number of servers. The QoS assessment for each paper in Fig. 8 is classified by every reference and what the author has measured for the review (see Table 5). Hence, we can notice the main aims for the most authors, are energy consumption and resource utilization. However, we did not find any research including or measuring all metrics for cloud computing quality of service that would have an effect on the end user. The major concern of cloud suppliers is SLA violation, which is caused by a limited number of server consolidation frameworks during the consolidation process [107], [112]–[115], and uncontrolled or unnecessary migrations which are triggered due to the static threshold [112]. The aforementioned migration problems lead to user dissatisfaction and SLA violation that has a notable impact on the cloud suppliers and cloud service cost as well.

The presented load balancing algorithms have several limitations e.g., resources, energy wastage, inadequate monitoring frequencies, and fixed thresholds in some algorithms. Therefore, there is an immense scope for improvements. More effective and adaptive load balancing algorithms should be developed to exploit performance, resource utilization, and energy conservation to deliver quality services to the users with the least cost. Some metrics were neglected in [19], [62], [69], [91], such as carbon emission, CADCloud sun sensitivity, cloud wind, green cloud rate, average heuristics performance, and rebalance time. This is because this paper has focused on the most popular metrics as in Table 5.

X. FUTURE RESEARCH DIRECTIONS

In spite of the plentiful literature available in this area, this study has highlighted certain aspects which have potential for further exploration.

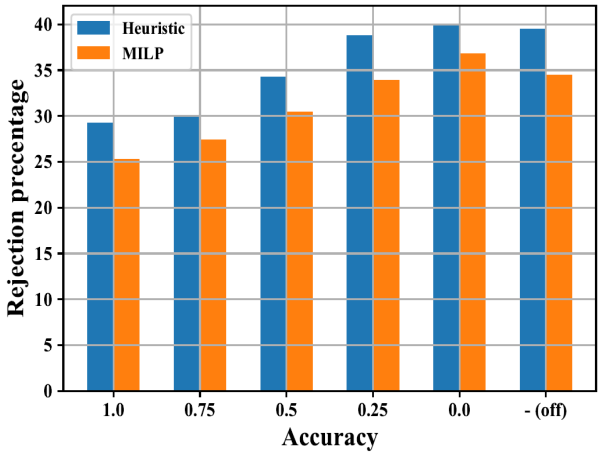


FIGURE 9. Accepted accuracy [116].

A. PREDICTION ACCURACY OF INCOMING WORKLOADS

Efficient workload prediction improves resource management and users' satisfaction to make improved usage of resources [124], but, is this true? What is the accuracy percentage of the workload prediction? How much is the prediction accuracy needed to improve decisions instead of causing harm?

In order to improve the system, these questions should be taken into consideration before proposing the prediction algorithms to be useful practically. If prediction accuracy is low, or prediction overhead is high, efficiency decreases. When constraints e.g. on deadlines, are loose, the improvements are less significant [116]. The prediction techniques and configurations should be set carefully to fit with any cloud problem. Because of the variety of cloud platforms and cloud fluctuations, prediction has advantages and disadvantages that can affect cloud resources [123]. As long as prediction is a probability, the need arises to use statistical methods such as the critical region (also called a rejection region) to tell you if your theory is probably true. Fig. 9 reveals the percentage of accuracy to achieve the desired improvement in the system. For instance, heuristic and mixed integer linear programming (MILP) algorithms show the accuracy must be at least 50% to make a sensible improvement. In a scenario when the predictor is off, accuracy is near to the average level of 0.25, which means it is not making any reasonable improvements. Often if the rejection rate goes up the service performance of the hybrid clouds goes down [117]. In this scenario, the prediction overhead is neglected. Nevertheless, if the prediction overhead was excessive, even excellent accuracy will cause degradation of efficiency of the cloud datacentre. Therefore, adopting this concept will open new areas of research and present a real integrated solution.

B. SECURITY-AWARE MIGRATION THREATS

Nowadays, data-centres face information security threats such as VM migration images attacks, hypervisor attacks, advanced persistent threats (APT), DoS/DDoS attacks,

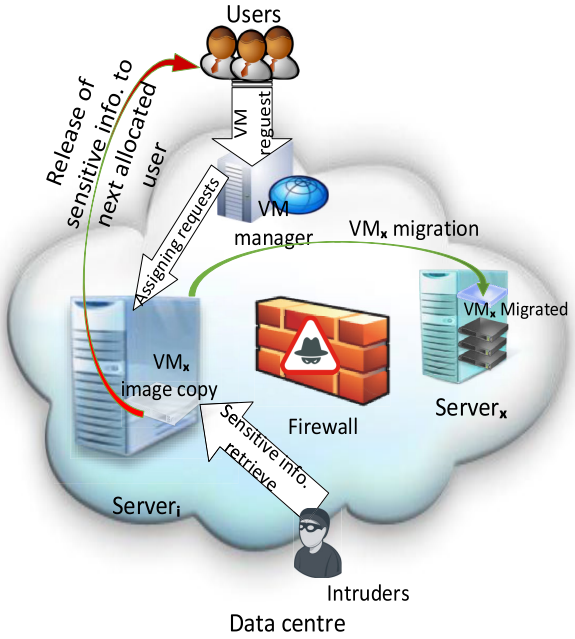


FIGURE 10. VM image threats.

etc. [118]–[120], which causes and increase in energy consumption and carbon footprints, even though the server consolidation neglected a sensitive aspect which is security during VMs migration within servers. The targeted server should implement an authentication method to check the VM identity before placing it in, so as to protect the system performance against intruders or unauthorized access to the server. Furthermore, VM images may not be deleted from the previous server [119], [121], [122] which could be released to the next assigned user or could retrieve/recover sensitive information by intruders as shown in the scenario in Fig. 10. Investigations are required to implement a consolidation method in such a way to protect the VMs images. This type of consolidation method is crucial, and the traditional migration should place confidential and personal information on top priority. Security countermeasures must be performed in a self-adaptive way. The predetermined security mechanism will be insufficient to tackle evolving threat vectors.

C. ADAPTIVE SERVER THRESHOLD AND CONSIDERING NEW METRICS

Some algorithms presented adaptive server threshold research which has a real impact in reducing the SLA violation. However, they did not accomplish 0% SLA violation. One of the solutions could be to utilize QoS-enabled to trigger the migration, and might reduce unnecessary migrations by designing new parameters such as SLA violation-free, power consumption, temperature, user workload pattern, and user scalability. This solution can control the migration decisions since it helps to predict the workload pattern that leads to SLA violation and unbalanced workload. To our knowledge, [71] is the first research that used computing energy, server failures, and cooling energy in a holistic approach along with a

novel performance metric for allocation decisions to assess the algorithms comprehensively. However, the presented algorithm and metrics need more investigation, looking into additional cooling components for instance water cooling devices. Also, the effect of server temperature on a system's reliability needs to be studied.

On the other hand, balancing via the migration processes (i.e. consolidation) uses up huge resources to handle it, such as CPU, BW, and memory that are taken from both hosts (i.e. the hosted server and the targeted server) [44]. Thus, as it rigorously affects their application, SLA, and QoS during the migration process could cause a system bottle-neck if the server exceeds the threshold. As a result, this will cause performance degradation in the whole system. Thus, the two-fold approach should be taken into consideration for a new algorithm.

D. NATURE-INSPIRED TECHNIQUES

An interesting issue for future investigation is energy efficiency techniques. There is a need for consistency or stability, for most of the analyses of these techniques' results. Whereas, nature inspired algorithms are heuristic generally. Thus, balance issues of these algorithms regarding development environments is required for further exploration. Furthermore, the success of accomplishing optimal energy efficiency of nature inspired algorithms depends on designing environments such as the set of parameters, operators, encoding scheme, and so on. While the majority of the algorithms are implemented in simulation tools and are assessed on various workload for the real problems, theoretical analysis must be designed first, then the simulation and implementation.

Real case scenarios are needed to solve real energy consumption issues to provide an optimal solution and performance to bridge the gaps and reduce the available limitations of the proposed algorithms and models.

E. DWINDLING RESOURCE REQUIREMENTS

Cloud computing profit can be maximized via limited utilizing of resources such as storage, CPU, BW, RAM, etc. By adapting the dwindling resource requirements in [125], [127] for job scheduling and revenue-driven service provisioning approach in [126] used in mobile cloud architecture to attain a maximum utilization of the cloud servers since that will result in increase in the cloud provider profit along with less migration. This is considered as an economic policy due to make the service provider the top priority, while placing the VMs preemptively in the hosts and assuming a limited resource available in a data-centre. Cooperation is needed with theoretical proofs to bridge the gap between practice and theory.

XI. CONCLUSION

This meta-study has reviewed the existing literature on load balancing using server consolidation. Various methods have been discussed using a thematic taxonomy which reflects the similarities factors between them. The methods analysed have a notable impact on reducing overall energy consumption besides efficient resource management in cloud data-centres. The review summarizes the load balancing methods from more than a thousand studies. Furthermore, this review aimed to examine the background of each method reviewed, as well as the wider content of the papers and the challenges presented by various methods. After this, a thematic taxonomy synthesised the similarities between load balancing and server consolidation and reviewed them from four points of view: hardware threshold, migration overhead, network traffic, and reliability. Finally, some descriptive statistics were provided, to enable the reader to make sense of the different methods in detail.

In the future, this meta-analysis will be expanded to include aspects of load balancing with/without server consolidation, including task load balancing for independent tasks and job scheduling with a more comprehensive taxonomy.

ACKNOWLEDGMENT

The authors would like to thank everyone who provided support to improve the content of this paper.

REFERENCES

[1] K. Dasgupta, B. Mandal, P. Dutta, J. K. Mandal, and S. Dam, "A genetic algorithm (GA) based load balancing strategy for cloud computing," *Procedia Technol.*, vol. 10, pp. 340–347, Jan. 2013.
[2] A. Apostu, F. Puican, G. Ularu, G. Suciu, and G. Todoran, "Study on advantages and disadvantages of Cloud Computing—The advantages of telemetry applications in the cloud," *Recent Adv. Appl. Comput. Sci. Digit. Serv.*, vol. 200, no. 1, pp. 118–123, 2013.
[3] K. Rajwinder and P. Luthra, "Load balancing in cloud computing," in *Proc. Int. Conf. Recent Trends Inf., Telecommun. Comput. (ITC)*, 2012, pp. 1–8.
[4] R. R. Malladi, "An approach to load balancing in cloud computing," *Int. J. Innov. Res. Sci., Eng. Technol.*, vol. 4, no. 5, pp. 3769–3777, 2015.
[5] Y. Jadeja and K. Modi, "Cloud computing—Concepts, architecture and challenges," in *Proc. Int. Conf. Comput. Electron. Elect. Technol. (ICCEET)*, Mar. 2012, pp. 877–880.
[6] C. Preist and P. Shabajee, "Energy use in the media cloud: Behaviour change, or technofix?" in *Proc. IEEE 2nd Int. Conf. Cloud Comput. Technol. Sci.*, Nov./Dec. 2010, pp. 581–586.
[7] P. Singh, P. Baaga, and S. Gupta, "Assorted load balancing algorithms in cloud computing: A survey," *Int. J. Comput. Appl.*, vol. 143, no. 7, pp. 34–40, 2016.
[8] S. Goyal and M. K. Verma, "Load balancing techniques in cloud computing environment—A review," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 6, no. 4, pp. 583–588, Apr. 2016.
[9] I. N. Ivanisenko and T. A. Radivilova, "Survey of major load balancing algorithms in distributed system," in *Proc. Inf. Technol. Innov. Bus. Conf. (ITIB)*, 2015, pp. 89–92.
[10] S. S. Manvi and G. K. Shyam, "Resource management for Infrastructure as a Service (IaaS) in Cloud computing: A survey," *J. Netw. Comput. Appl.*, vol. 41, pp. 424–440, May 2014.
[11] J. Koomey, "Growth in data center electricity use 2005 to 2010," *New York Times*, vol. 9, pp. 1–24, Aug. 2011.
[12] A. Shehabi, S. J. Smith, E. Masanet, and J. Koomey, "Data center growth in the United States: Decoupling the demand for services from electricity use," *Environ. Res. Lett.*, vol. 13, no. 12, 2018, Art. no. 124030.
[13] S. V. Vrbsky, M. Lei, K. Smith, and J. Byrd, "Data replication and power consumption in data grids," in *Proc. IEEE 2nd Int. Conf. Cloud Comput. Technol. Sci.*, Nov./Dec. 2010, pp. 288–295.
[14] C. Preist and P. Shabajee, "Energy use in the media cloud: Behaviour change, or technofix?" in *Proc. IEEE 2nd Int. Conf. Cloud Comput. Technol. Sci.*, Jan. 2010, pp. 581–586.
[15] M. Blackburn and A. Hawkins. *Unused Server Survey Results Analysis*. Accessed: Dec. 6, 2013. [Online]. Available: https://www:thegreengrid:org=media=WhitePapers=Unused%20Server%20Study_WP_101910_v1:ashx?lang=en

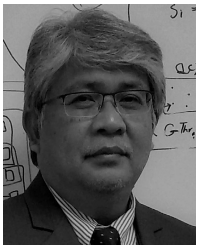
[16] B. Snyder. (2010). *Server Virtualization has Stalled*. [Online]. Available: http://www.infoworld.com/print/146901
[17] R. Palta and R. Jeet, "Load balancing in the cloud computing using virtual machine migration: A review," *Int. J. Appl. Innov. Eng. Manage.*, vol. 3, no. 5, pp. 437–441, 2014.
[18] L. A. Barroso and U. Hölzle, "The case for energy-proportional computing," *IEEE Comput.*, vol. 40, no. 12, pp. 33–37, Dec. 2007.
[19] S. Distefano, G. Merlino, and A. Puliafito, "A utility paradigm for IoT: The sensing cloud," *Pervasive Mobile Comput.*, vol. 20, pp. 127–144, Jul. 2015.
[20] A. Botta, W. De Donato, V. Persico, and A. Pescapé, "Integration of Cloud computing and Internet of Things: A survey," *Future Gener. Comput. Syst.*, vol. 56, pp. 684–700, 2016.
[21] O. Diallo, J. J. P. C. Rodrigues, M. Sene, and J. Niu, "Real-time query processing optimization for cloud-based wireless body area networks," *Inf. Sci.*, vol. 284, pp. 84–94, Nov. 2014.
[22] K. Al Nuaimi, N. Mohamed, M. Al Nuaimi, and J. Al-Jaroodi, "A survey of load balancing in cloud computing: Challenges and algorithms," in *Proc. IEEE 2nd Symp. Netw. Cloud Comput. Appl.*, Dec. 2012, pp. 137–142.
[23] V. R. Kanakala and V. K. Reddy, "Performance analysis of load balancing techniques in cloud computing environment," *TELKOMNIKA Indonesian J. Electr. Eng.*, vol. 13, no. 3, pp. 1–6, 2015.
[24] W. Tian, M. He, W. Guo, W. Huang, X. Shi, M. Shang, A. N. Toosi, and R. Buyya, "On minimizing total energy consumption in the scheduling of virtual machine reservations," *J. Netw. Comput. Appl.*, vol. 113, pp. 64–74, Jul. 2018.
[25] A. Khiyaita, H. El Bakkali, M. Zbakh, and D. El Kettani, "Load balancing cloud computing: State of art," in *Proc. Nat. Days Netw. Secur. Syst.*, 2012, pp. 106–109.
[26] S. Ray and A. Das, "Execution analysis of load balancing algorithms in cloud computing environment," *Int. J. Cloud Comput., Services Archit.*, vol. 2, no. 5, pp. 1–13, 2012.
[27] A. K. Sidhu and S. Kinger, "Analysis of load balancing techniques in cloud computing," *Int. J. Comput. Technol.*, vol. 4, no. 2, pp. 737–741, 2013.
[28] E. J. Ghomi, A. M. Rahmani, and N. N. Qader, "Load-balancing algorithms in cloud computing: A survey," *J. Netw. Comput. Appl.*, vol. 88, pp. 50–71, Jun. 2017.
[29] N. Vasić, M. Barisits, D. Kosti, and V. Salzgeber, "Making cluster applications energy-aware," in *Proc. ACM 1st Workshop Autom. Control Datacenters Clouds (ACDC)*, New York, NY, USA, 2009, pp. 37–42.
[30] B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering—A systematic literature review," *Inf. Softw. Technol.*, vol. 51, no. 1, pp. 7–15, 2009.
[31] A. Thakur and M. S. Goraya, "A taxonomic survey on load balancing in cloud," *J. Netw. Comput. Appl.*, vol. 98, pp. 43–57, Nov. 2017.
[32] A. Varasteh and M. Goudarzi, "Server consolidation techniques in virtualized data centers: A survey," *IEEE Syst. J.*, vol. 11, no. 2, pp. 772–783, Jun. 2017.
[33] A. S. Milani and N. J. Navimipour, "Load balancing mechanisms and techniques in the cloud environments: Systematic literature review and future trends," *J. Netw. Comput. Appl.*, vol. 71, pp. 86–98, Aug. 2016.
[34] H. Gupta and K. Sahu, "Honey bee behavior based load balancing of tasks in cloud computing," *Int. J. Sci. Res.*, vol. 3, no. 6, pp. 842–846, Jun. 2014.
[35] Z. Soltani and N. J. Navimipour, "Customer relationship management mechanisms: A systematic review of the state of the art literature and recommendations for future research," *Comput. Hum. Behav.*, vol. 61, pp. 667–688, Aug. 2016.
[36] E. Y. Daraghmi and S.-M. Yuan, "A small world based overlay network for improving dynamic load-balancing," *J. Syst. Softw.*, vol. 107, pp. 187–203, Sep. 2015.
[37] K. Hwang, J. Dongarra, and G. C. Fox, *Distributed and Cloud Computing: From Parallel Processing to the Internet of Things*. San Mateo, CA, USA: Morgan Kaufmann, 2013.
[38] X. Meng, V. Pappas, and L. Zhang, "Improving the scalability of data center networks with traffic-aware virtual machine placement," in *Proc. IEEE INFOCOM*, Mar. 2010, pp. 1–9.
[39] B. Kitchenham, "Procedures for performing systematic reviews," Dept. Comput. Sci., Keele Univ., Keele, U.K., Tech. Rep. TR/SE-0401, 2014, vol. 33, pp. 1–26.
[40] E. Kupiainen, M. V. Mäntylä, and J. Itkonen, "Using metrics in agile and lean software development—A systematic literature review of industrial studies," *Inf. Softw. Technol.*, vol. 62, pp. 143–163, Jun. 2015.
[41] Y. Charband and N. J. Navimipour, "Erratum to: Online knowledge sharing mechanisms: A systematic review of the state of the art literature and recommendations for future research," *Inf. Syst. Frontiers*, vol. 21, no. 4, pp. 1–21, 2017.
[42] N. J. Navimipour and Y. Charband, "Knowledge sharing mechanisms and techniques in project teams: Literature review, classification, and current trends," *Comput. Hum. Behav.*, vol. 62, pp. 730–742, Sep. 2016.
[43] Y. Lu, Q. Xie, G. Kliot, A. Geller, J. R. Larus, and A. Greenberg, "Join-Idle-Queue: A novel load balancing algorithm for dynamically scalable Web services," *Perform. Eval.*, vol. 68, no. 11, pp. 1056–1071, Nov. 2011.
[44] U. Deshpande, U. Kulkarni, and K. Gopalan, "Inter-rack live migration of multiple virtual machines," in *Proc. 6th Int. Workshop Virtualization Technol. Distrib. Comput. Date (VTDC)*, 2012, pp. 19–26.
[45] A. Nakai, E. Madeira, and L. E. Buzato, "On the use of resource reservation for Web services load balancing," *J. Netw. Syst. Manage.*, vol. 23, no. 3, pp. 502–538, 2014.
[46] Y. Fang, F. Wang, and J. Ge, "A task scheduling algorithm based on load balancing in cloud computing," in *Proc. Int. Conf. Web Inf. Syst. Mining*, Oct. 2010, pp. 271–277.
[47] S. Banerjee, M. Adhikari, S. Kar, and U. Biswas, "Development and analysis of a new cloudlet allocation strategy for QoS improvement in cloud," *Arabian J. Sci. Eng.*, vol. 40, no. 5, pp. 1409–1425, 2015.
[48] Z. Wang, H. Chen, Y. Fu, D. Liu, and Y. Ban, "Workload balancing and adaptive resource management for the swift storage system on cloud," *Future Gener. Comput. Syst.*, vol. 51, pp. 120–131, Oct. 2015.
[49] S. T. Maguluri, R. Srikant, and L. Ying, "Stochastic models of load balancing and scheduling in cloud computing clusters," in *Proc. IEEE INFOCOM*, Mar. 2012, pp. 702–710.
[50] F. Ramezani, J. Lu, and F. K. Hussain, "Task-based system load balancing in cloud computing using particle swarm optimization," *Int. J. Parallel Program.*, vol. 42, no. 5, pp. 739–754, 2013.
[51] S. M. Abdulhamid, M. S. A. Latiff, and I. Idris, "Tasks scheduling technique using league championship algorithm for makespan minimization in IaaS cloud," 2015, *arXiv:1510.03173*. [Online]. Available: https://arxiv.org/abs/1510.03173
[52] W. Voorsluys, J. Broberg, S. Venugopal, and R. Buyya, "Cost of virtual machine live migration in clouds: A performance evaluation," in *Proc. IEEE Int. Conf. Cloud Comput.* Berlin, Germany: Springer, Dec. 2009, pp. 254–265.
[53] M. Abdullahi, M. A. Ngadi, and S. M. Abdulhamid, "Symbiotic organism search optimization based task scheduling in cloud computing environment," *Future Gener. Comput. Syst.*, vol. 56, pp. 640–650, Mar. 2016.
[54] Z. Zhang and X. Zhang, "A load balancing mechanism based on ant colony and complex network theory in open cloud computing federation," in *Proc. 2nd Int. Conf. Ind. Mechatronics Automat.*, May 2010, vol. 2, no. 1, pp. 240–243.
[55] K.-M. Cho, P.-W. Tsai, C.-W. Tsai, and C.-S. Yang, "A hybrid meta-heuristic algorithm for VM scheduling with load balancing in cloud computing," *Neural Comput. Appl.*, vol. 26, no. 6, pp. 1297–1309, 2015.
[56] J. Qian, J. Li, R. Ma, L. Lin, and H. Guan, "LG-RAM: Load-aware global resource affinity management for virtualized multicore systems," *J. Syst. Archit.*, vol. 98, pp. 114–125, Sep. 2019.
[57] W. Guo, P. Kuang, Y. Jiang, X. Xu, and W. Tian, "SAVE: Self-adaptive consolidation of virtual machines for energy efficiency of CPU-intensive applications in the cloud," *J. Supercomput.*, pp. 1–25, Jun. 2019. doi: 10.1007/s11227-019-02927-1.
[58] S. Bhattacherjee, R. Das, S. Khatua, and S. Roy, "Energy-efficient migration techniques for cloud environment: A step toward green computing," *J. Supercomput.*, pp. 1–29, Mar. 2019. doi: 10.1007/s11227-019-02801-0.
[59] A. K. Kulkarni and B. Annappa, "Context aware VM placement optimization technique for heterogeneous IaaS cloud," *IEEE Access*, vol. 7, pp. 89702–89713, 2019.
[60] G. Cao, "Topology-aware multi-objective virtual machine dynamic consolidation for cloud datacenter," *Sustain. Comput. Inform. Syst.*, vol. 21, pp. 179–188, Mar. 2019.
[61] M. S. Mekala and P. Viswanathan, "Energy-efficient virtual machine selection based on resource ranking and utilization factor approach in cloud computing for IoT," *Comput. Elect. Eng.*, vol. 73, pp. 227–244, Jan. 2019.
[62] S. Filiposka, A. Mishev, and K. Gilly, "Multidimensional hierarchical VM migration management for HPC cloud environments," *J. Supercomput.*, vol. 75, pp. 5324–5346, Aug. 2019.

[63] H. Nashaat, N. Ashry, and R. Rizk, "Smart elastic scheduling algorithm for virtual machine migration in cloud computing," *J. Supercomput.*, vol. 75, pp. 3842–3865, Jul. 2019.
[64] C.-T. Yang, S.-T. Chen, J.-C. Liu, Y.-W. Chan, C.-C. Chen, and V. K. Verma, "An energy-efficient cloud system with novel dynamic resource allocation methods," *J. Supercomput.*, vol. 75, no. 8, pp. 4408–4429, 2019.
[65] L. Guo, G. Hu, Y. Dong, Y. Luo, and Y. Zhu, "A game based consolidation method of virtual machines in cloud data centers with energy and load constraints," *IEEE Access*, vol. 6, pp. 4664–4676, 2018.
[66] M. Hähnel, J. Martinovic, G. Scheithauer, A. Fischer, A. Schill, and W. Dargie, "Extending the cutting stock problem for consolidating services with stochastic workloads," *IEEE Trans. Parallel Distrib. Syst.*, vol. 29, no. 11, pp. 2478–2488, Nov. 2018.
[67] R. Cao, Z. Tang, K. Li, and K. Li, "HMGOWM: A hybrid decision mechanism for automating migration of virtual machines," *IEEE Trans. Services Comput.*, to be published.
[68] E. Baccarelli, M. Scarpiniti, and A. Momenzadeh, "Fog-supported delay-constrained energy-saving live migration of VMs over multipath TCP/IP 5G connections," *IEEE Access*, vol. 6, pp. 42327–42354, 2018.
[69] H. A. Kurdi, S. M. Alismail, and M. M. Hassan, "LACE: A locust-inspired scheduling algorithm to reduce energy consumption in cloud datacenters," *IEEE Access*, vol. 6, pp. 35435–35448, 2018.
[70] S. G. Domanal and G. R. M. Reddy, "An efficient cost optimized scheduling for spot instances in heterogeneous cloud environment," *Future Gener. Comput. Syst.*, vol. 84, pp. 11–21, Jul. 2018.
[71] X. Li, X. Jiang, P. Garraghan, and Z. Wu, "Holistic energy and failure aware workload scheduling in Cloud datacenters," *Future Gener. Comput. Syst.*, vol. 78, pp. 887–900, Jan. 2018.
[72] G. G. Kumar and P. Vivekanandan, "Energy efficient scheduling for cloud data centers using heuristic based migration," *Cluster Comput.*, pp. 1–8, Feb. 2018. doi: 10.1007/s10586-018-2235-7.
[73] Z. Li, "An adaptive overload threshold selection process using Markov decision processes of virtual machine in cloud data center," *Cluster Comput.*, pp. 1–13, Mar. 2018. doi: 10.1007/s10586-018-2408-4.
[74] M. Riahi and S. Krichen, "A multi-objective decision support framework for virtual machine placement in cloud data centers: A real case study," *J. Supercomput.*, vol. 74, no. 7, pp. 2984–3015, 2018.
[75] C. Jiang, J. Wu, and Z. Li, "Adaptive thresholds determination for saving cloud energy using three-way decisions," *Cluster Comput.*, pp. 1–8, Feb. 2018.
[76] Z. Li, C. Yan, L. Yu, and X. Yu, "Energy-aware and multi-resource overload probability constraint-based virtual machine dynamic consolidation method," *Future Gener. Comput. Syst.*, vol. 80, pp. 139–156, Mar. 2018.
[77] M. Ranjbari and J. A. Torkestani, "A learning automata-based algorithm for energy and SLA efficient consolidation of virtual machines in cloud data centers," *J. Parallel Distrib. Comput.*, vol. 113, pp. 55–62, Mar. 2018.
[78] Kamran, and B. Nazir, "QoS-aware VM placement and migration for hybrid cloud infrastructure," *J. Supercomput.*, vol. 74, no. 9, pp. 4623–4646, 2018. doi: 10.1007/s11227-017-2071-1.
[79] H. Li, G. Zhu, Y. Zhao, Y. Dai, and W. Tian, "Energy-efficient and QoS-aware model based resource consolidation in cloud data centers," *Cluster Comput.*, vol. 20, no. 3, pp. 2793–2803, 2017.
[80] A. Paya and D. C. Marinescu, "Energy-aware load balancing and application scaling for the cloud ecosystem," *IEEE Trans. Cloud Comput.*, vol. 5, no. 1, pp. 15–27, Jan./Mar. 2017.
[81] M. Tighe and M. Bauer, "Topology and application aware dynamic VM management in the cloud," *J. Grid Comput.*, vol. 15, no. 2, pp. 273–294, 2017.
[82] K. Tsakalozos, V. Verroios, M. Roussopoulos, and A. Delis, "Live VM migration under time-constraints in share-nothing IaaS-clouds," *IEEE Trans. Parallel Distrib. Syst.*, vol. 28, no. 8, pp. 2285–2298, Aug. 2017.
[83] R. Li, Q. Zheng, X. Li, and Z. Yan, "Multi-objective optimization for rebalancing virtual machine placement," *Future Gener. Comput. Syst.*, to be published.
[84] S. Suresh and S. Sakthivel, "A novel performance constrained power management framework for cloud computing using an adaptive node scaling approach," *Comput. Elect. Eng.*, vol. 60, pp. 30–44, May 2017.
[85] Z. Li, C. Yan, X. Yu, and N. Yu, "Bayesian network-based virtual machines consolidation method," *Future Gener. Comput. Syst.*, vol. 69, no. 3, pp. 75–87, Apr. 2017.
[86] W. Zhu, Y. Zhuang, and L. Zhang, "A three-dimensional virtual resource scheduling method for energy saving in cloud computing," *Future Gener. Comput. Syst.*, vol. 69, pp. 66–74, Apr. 2017.
[87] H. Yan, H. Wang, X. Li, Y. Wang, D. Li, Y. Zhang, Y. Xie, Z. Liu, W. Cao, and F. Yu, "Cost-efficient consolidating service for Aliyun's cloud-scale computing," *IEEE Trans. Services Comput.*, vol. 12, no. 1, pp. 117–130, Jan./Feb. 2019.
[88] J. Kommeri, T. Niemi, and J. K. Nurminen, "Energy efficiency of dynamic management of virtual cluster with heterogeneous hardware," *J. Supercomput.*, vol. 73, no. 5, pp. 1978–2000, 2016.
[89] H. Li, G. Zhu, C. Cui, H. Tang, Y. Dou, and C. He, "Energy-efficient migration and consolidation algorithm of virtual machines in data centers for cloud computing," *Computing*, vol. 98, no. 3, pp. 303–317, Mar. 2016.
[90] J. O. Gutierrez-Garcia and A. Ramirez-Nafarrate, "Agent-based load balancing in cloud data centers," *Cluster Comput.*, vol. 18, no. 3, pp. 1041–1062, Sep. 2015.
[91] F. F. Moghaddam, R. F. Moghaddam, and M. Cheriet, "Carbon-aware distributed cloud: Multi-level grouping genetic algorithm," *Cluster Comput.*, vol. 18, no. 1, pp. 477–491, 2014.
[92] W. Song, Z. Xiao, Q. Chen, and H. Luo, "Adaptive resource provisioning for the cloud using online bin packing," *IEEE Trans. Comput.*, vol. 63, no. 11, pp. 2647–2660, Nov. 2014.
[93] G. Lovász, F. Niedermeier, and H. de Meer, "Performance tradeoffs of energy-aware virtual machine consolidation," *Cluster Comput.*, vol. 16, no. 3, pp. 481–496, 2012.
[94] J. Hall, J. Hartline, A. R. Karlin, J. Saia, and J. Wilkes, "On algorithms for efficient data migration," in *Proc. 12th Annu. ACM-SIAM Symp. Discrete Algorithms*, 2001, pp. 620–629.
[95] D. Oppenheimer, A. Ganapathi, and D. A. Patterson, "Why do Internet services fail, and what can be done about it?" in *Proc. USENIX Symp. Internet Technol. Syst.*, 2003, pp. 1–16.
[96] W. Deng, F. Liu, H. Jin, X. Liao, and H. Liu, "Reliability-aware server consolidation for balancing energy-lifetime tradeoff in virtualized cloud datacenters," *Int. J. Commun. Syst.*, vol. 27, no. 4, pp. 623–642, Apr. 2014.
[97] S. Takahashi, H. Nakada, A. Takefusa, T. Kudoh, M. Shigeno, and A. Yoshise, "Virtual machine packing algorithms for lower power consumption," in *Proc. SC Companion, High Perform. Comput., Netw. Storage Anal.*, 2012, pp. 1519–1520.
[98] A. Stage and T. Setzer, "Network-aware migration control and scheduling of differentiated virtual machine workloads," in *Proc. ICSE Workshop Softw. Eng. Challenges Cloud Comput.*, 2009, pp. 9–14.
[99] C. Clark, K. Fraser, S. Hand, J. G. Hansen, E. Jul, C. Limpach, I. Pratt, and A. Warfield, "Live migration of virtual machines," in *Proc. 2nd Conf. Symp. Netw. Syst. Design Implement.*, vol. 2, 2005, pp. 273–286.
[100] S. Akoush, R. Sohan, A. Rice, A. W. Moore, and A. Hopper, "Predicting the performance of virtual machine migration," in *Proc. IEEE Int. Symp. Modeling, Anal. Simulation Comput. Telecommun. Syst.*, Aug. 2010, pp. 37–46.
[101] L. D. D. Babu and P. V. Krishna, "Honey bee behavior inspired load balancing of tasks in cloud computing environments," *Appl. Soft Comput.*, vol. 13, no. 5, pp. 2292–2303, May 2013.
[102] G. Xu, J. Pang, and X. Fu, "A load balancing model based on cloud partitioning for the public cloud," *Tsinghua Sci. Technol.*, vol. 18, no. 1, pp. 34–39, 2013.
[103] A. Beloglazov, J. Abawajy, and R. Buyya, "Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing," *Future Generat. Comput. Syst.*, vol. 28, no. 5, pp. 755–768, 2012.
[104] W. Wang, H. Chen, and X. Chen, "An availability-aware virtual machine placement approach for dynamic scaling of cloud applications," in *Proc. IEEE 9th Int. Conf. UIC/ATC*, Sep. 2012, pp. 509–516.
[105] W. Vogels, "Beyond server consolidation," *ACM Queue*, vol. 6, no. 1, pp. 20–26, Jan./Feb. 2008.
[106] C. B. Pop, I. Anghel, T. Cioara, I. Salomie, and I. Vartic, "A swarm-inspired data center consolidation methodology," in *Proc. 2nd Int. Conf. Web Intell., Mining Semantics (WIMS)*, 2012, Art. no. 41.
[107] T. C. Ferreto, M. A. S. Netto, R. N. Calheiros, and C. A. F. De Rose, "Server consolidation with migration control for virtualized data centers," *Future Gener. Comput. Syst.*, vol. 27, no. 8, pp. 1027–1034, 2011.
[108] T. Wood, G. Tarasuk-Levin, P. Shenoy, P. Desnoyers, E. Cecchet, and M. D. Corner, "Memory buddies: Exploiting page sharing for smart colocation in virtualized data centers," in *Proc. ACM SIGPLAN/SIGOPS Int. Conf. Virtual Execution Environ. (VEE)*, 2009, pp. 31–40.

[109] M. Dorigo and G. Di Caro, "Ant colony optimization: A new meta-heuristic," in *Proc. Congr. Evol. Comput. (CEC)*, vol. 2, Jul. 1999, pp. 1470–1477.
[110] S. R. Welke, B. W. Johnson, and J. H. Aylor, "Reliability modeling of hardware/software systems," *IEEE Trans. Rel.*, vol. 44, no. 3, pp. 413–418, Sep. 1995.
[111] S. K. Habib, S. Ries, and M. Muhlhauser, "Cloud computing landscape and research challenges regarding trust and reputation," in *Proc. 7th Int. Conf. Ubiquitous Intell. Comput., 7th Int. Conf. Auton. Trusted Comput*, 2010, pp. 410–415.
[112] A. Beloglazov and R. Buyya, "Adaptive threshold-based approach for energy-efficient consolidation of virtual machines in cloud data centers," in *Proc. MGC@ Middleware*, 2010, Art. no. 4.
[113] A. Beloglazov and R. Buyya, "Energy efficient allocation of virtual machines in cloud data centers," in *Proc. 10th IEEE/ACM Int. Conf. Cluster, Cloud Grid Comput.*, May 2010, pp. 577–578.
[114] A. Beloglazov and R. Buyya, "Energy efficient resource management in virtualized cloud data centers," in *Proc. 10th IEEE/ACM Int. Conf. Cluster, Cloud Grid Comput.*, May 2010, pp. 826–831.
[115] A. Beloglazov and R. Buyya, "Managing overloaded hosts for dynamic consolidation of virtual machines in cloud data centers under quality of service constraints," *IEEE Trans. Parallel Distrib. Syst.*, vol. 24, no. 7, pp. 1366–1379, Jul. 2013.
[116] M. Niknafs, I. Ukhov, P. Eles, and Z. Peng, "Runtime resource management with workload prediction," in *Proc. 56th ACM Annu. Design Autom. Conf.*, 2019, pp. 1–6.
[117] Y. Cao, L. Lu, J. Yu, S. Qian, Y. Zhu, and M. Li, "Online cost-rejection rate scheduling for resource requests in hybrid clouds," *Parallel Comput.*, vol. 81, pp. 85–103, Dec. 2019.
[118] A. Zimba, H. Chen, and Z. Wang, "Bayesian network based weighted APT attack paths modeling in cloud computing," *Future Gener. Comput. Syst.*, vol. 96, pp. 525–537, Jul. 2019.
[119] R. Kumar and R. Goyal, "On cloud security requirements, threats, vulnerabilities and countermeasures: A survey," *Comput. Sci. Rev.*, vol. 33, pp. 1–48, Aug. 2019.
[120] S. Singh, P. Sharma, S. Y. Moon, and J. H. Park, "EH-GC: An efficient and secure architecture of energy harvesting green cloud infrastructure," *Sustainability*, vol. 9, no. 4, p. 673, 2017.
[121] J. Wei, X. Zhang, G. Ammons, V. Bala, and P. Ning, "Managing security of virtual machine images in a cloud environment," in *Proc. ACM Workshop Cloud Comput. Secur. (CCSW)*, 2009, pp. 91–96.
[122] A. Singh and K. Chatterjee, "Cloud security issues and challenges: A survey," *J. Netw. Comput. Appl.*, vol. 79, pp. 88–115, Feb. 2017.
[123] D. F. Kirchoff, M. Xavier, J. Mastella, and C. A. F. De Rose, "A preliminary study of machine learning workload prediction techniques for cloud applications," in *Proc. IEEE 27th Euromicro Int. Conf. Parallel, Distrib. Netw.-Based Process. (PDP)*, Feb. 2019, pp. 222–227.
[124] S. Seneviratne, S. Witharana, and A. N. Toosi, "Adapting the machine learning grid prediction models for forecasting of resources on the clouds," in *Proc. Adv. Sci. Eng. Technol. Int. Conf. (ASET)*, 2019, pp. 1–6.
[125] S. Albagli-Kim, H. Shachnai, and T. Tamir, "Scheduling jobs with dwindling resource requirements in clouds," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, Apr./May 2014, pp. 601–609.
[126] H. Wu, S. Deng, W. Li, J. Yin, Q. Yang, Z. Wu, and A. Y. Zomaya, "Revenue-driven service provisioning for resource sharing in mobile cloud computing," in *Proc. Int. Conf. Service-Oriented Comput.*, in Lecture Notes in Computer Science. Cham, Switzerland: Springer, 2017, pp. 625–640.
[127] S. Albagli-Kim, B. Schieber, H. Shachnai, and T. Tamir, "Real-time k-bounded preemptive scheduling," in *Proc. 18th Workshop Algorithm Eng. Exp. (ALENEX)*, 2015, pp. 127–137.

MOHAMMED ALA'ANZY received the master's degree in computer science from University Putra Malaysia, Serdang, Malaysia, in 2017, where he is currently pursuing the Ph.D. degree with the Faculty of Computer Science and Information Technology.

He has authored several journal articles. His current research interests include cloud computing, green computing, load balancing, and task scheduling.



MOHAMED OTHMAN received the Ph.D. degree (Hons.) from the National University of Malaysia. He was a Deputy Director of the Information Development and Communication Center, where he was in charge of the UMPNet network campus, uSport Wireless Communication Project, and the UPM Data Center. He is currently a Professor in computer science with the Department of Communication Technology and Networks, Universiti Putra Malaysia (UPM). He is also an Associate Researcher and a Coordinator of high-speed machines with the Laboratory of Computational Science and Informatics, Institute of Mathematical Science, UPM. He is also a Visiting Professor with South Kazakhstan State University, Shymkent, and L. N. Gumilyov Eurasian National University, Astana, Kazakhstan. He has filed six Malaysian, one Japanese, one South Korean, and three U.S. patents. He has published more than 300 International journals and 330 proceeding articles. His main research interests include computer networks, parallel and distributed computing, high-speed interconnection networks, network design and management (network security, and wireless and traffic monitoring), consensus in the IoT, and mathematical models in scientific computing. He is a Life Member of the Malaysian National Computer Confederation and the Malaysian Mathematical Society. In 2017, he received an Honorary Professorship from SILKWAY International University (formerly known as South Kazakhstan Pedagogical University), Shymkent, Kazakhstan, and the Best Ph.D. Thesis by Sime Darby Malaysia and the Malaysian Mathematical Science Society, in 2000.