Journal of Cloud Computing:
Advances, Systems and Applications

REVIEW

Open Access

A systematic review on effective energy utilization management strategies in cloud data centers

Suraj Singh Panwar*, M. M. S. Rauthan and Varun Barthwal

Abstract

Data centers are becoming considerably more significant and energy-intensive due to the exponential growth of cloud computing. Cloud computing allows people to access computer resources on demand. It provides amenities on the pay-as-you-go basis across the data center locations spread over the world. Consequently, cloud data centers consume a lot of electricity and leave a proportional carbon impact on the environment. There is a need to investigate efficient energy-saving approaches to reduce the massive energy usage in cloud servers. This review paper focuses on identifying the research done in the field of energy consumption (EC) using different techniques of machine learning, heuristics, metaheuristics, and statistical methods. Host CPU utilization prediction, underload/overload detection, virtual machine selection, migration, and placement have been performed to manage the resources and achieve efficient energy utilization. In this review, energy savings achieved by different techniques are compared. Many researchers have tried various methods to reduce energy usage and service level agreement violations (SLAV) in cloud data centers. By using the heuristic approach, researchers have saved 5.4% to 90% of energy with their proposed methods compared with the existing methods. Similarly, the metaheuristic approaches reduce energy consumption from 7.68% to 97%, the machine learning methods from 1.6% to 88.5%, and the statistical methods from 5.4% to 84% when compared to the benchmark approaches for a variety of settings and parameters. So, making energy use more efficient could cut down the air pollution, greenhouse gas (GHG) emissions, and even the amount of water needed to make power. The overall outcome of this review work is to understand different methods used by researchers to save energy in cloud data centers.

Keywords: Cloud computing, Resources, Data center, Virtual machine, Energy consumption, SLAV

Introduction

Cloud Computing has become a flexible, resourceful, efficient, and prevalent computational technology that offers users reliable, customized, and dynamic computing environments. Cloud applications are hosted on high-capacity systems and storage devices in multiple locations around the world. Rapid demand for cloud-based facilities essentially requires the development of massive data centers that consume excessive amounts of electricity. Optimization of energy can be proficient by uniting resources based on current utilization, well-organized network, and the thermal position of nodes and computing equipment. Because maximizing the utilization of physical servers is essential in lowering a data center's (DC) energy demand, virtual machines (VMs) have been effectively introduced in DCs to increase server resource utilization. A method for cost-effective VM migration based on fluctuating electricity prices cuts the energy costs of running a cloud service by a large amount.

Cloud computing is an extension of parallel computing, utility computing, cluster computing, and grid computing. It is distributed in nature, so a group of independent

*Correspondence: surajpanwar1977@gmail.com

Hemvati Nandan Bahuguna Garhwal University, Srinagar Garhwal, Uttarakhand, India

resources are spread in remote locations. Cloud computing is defined by NIST as "a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., storage, networks, servers, services, and applications) that can be rapidly provisioned and released with minimal management effort or service provider interaction" [1, 2].

The service models of cloud computing are Software as a Service (SaaS), Platform as a Platform (PaaS), and Infrastructure as a Service (IaaS). In SaaS, the client has access to cloud services via a web browser to maintain user interaction and data in the cloud. PaaS is a service that allows customers to use the platform and tools instead of purchasing and paying for software licences for platforms such as operating systems, databases, and intermediary applications.

IaaS means the necessary environment to facilitate cloud services. It contains the pool of hardware resources related to computing, storage, networking, etc. Based on the model of deployment, clouds are categorized into four types. The term "public cloud" refers to an infrastructure that allows the general public to store and access data from any location using a client device with an internet connection. Private Cloud: A private cloud or enterprise cloud is one where the facilities and infrastructure are available for the organization or partner's use only. A Hybrid Cloud: When a private cloud is combined with public cloud computing. Community Cloud: Resources are shared by multiple organizations that serve a particular community with common concerns [3, 4].

Today, research community's top priorities are energy conservation and effectiveness. The issue of excessive energy utilization arises as a result of unexpected and rapid changes in the environment around the globe [5]. The levels of carbon footprint and Green House Gases (GHG) in the environment have rapidly increased. The information and communication technology (ICT) industry has been identified as the primary emitter [6]. The rise of sophisticated and diverse data-intensive services and applications has exacerbated energy challenges. The intensity and constant growth of ICT energy demand have necessitated not only meeting energy requirements but also developing and implementing efficient energy-savings methods. According to a 2016 survey, the total global energy consumption and CO_2 emissions are expected to rise by 48% and 34%, respectively, between 2010 and 2040 [7]. Also, the Climate Action Group found that the world released 32 gigatonnes of CO2 in 2015 [8].

The paper is organized as follows: First, a brief introduction of cloud computing, motivation, virtualization, energy consumption, SLAV, VM consolidation, CloudSim, workload datasets, purpose, and classification of the survey have been explained. Further next section defines the discussion, analysis, objectives, limitations, and evaluation of existing related work for heuristic, metaheuristic, machine learning, and statistical techniques with tools, performance metrics, and comparisons with their benchmark algorithms related to energy consumption. In last section, result analysis, major challenges, suggestions, and future work are elaborated. Finally, the summary and conclusion of the review paper is summarised to improve energy efficiency in cloud data centers.

Motivation

The idea behind cloud computing is to provide on-demand quick access to cloud data centers and to administer the operations from a remote location. Cloud computing operates on a pay-as-you-go pricing model, allowing organizations to reduce operational costs and manage infrastructure more effectively. The motivation for conducting the survey, entitled 'Effective Energy Utilization Management Strategies in Cloud Data Centers' is to reduce power utilization in well-organized data centers with the help of VM consolidation. There are several proposed resource management approaches for several computing domains, but only a few addresses the issue of energy efficiency in addition to optimizing profit and service quality. Many magnificent studies have been devoted to confirming the consolidation achieved to an appreciable value, but it is still in its developing stage. Various survey papers on load balancing [9, 10], resource provisioning [11], resource scheduling [12, 13], resource allocation [14], and resource utilization [15] have been published. These surveys explored resource management classification and compared state-of-the-art algorithms based on many significant characteristics of cloud computing. But the classification and techniques related to effective energy utilization approaches have not been discussed in detail in the current study. As a result, there is a need for a complete and systematic assessment of existing energy-efficient strategies, as well as their limitations, to entice academics to work in this domain. This study provides an attempt to investigate the categorization of energy-efficient virtual machine consolidation thoroughly, which will be useful for future research in developing new energy-efficient algorithms or methodologies. The limitations of current approaches are emphasized to inspire future research work challenges and the development of algorithms. The following are the primary contributions of the review paper:

- Investigate and analyze the various existing energy-efficient methods in cloud data centers.
- Classification of VM management using heuristics, metaheuristics, machine learning, and statistical techniques.

- The most important parts of each classification are explained, and a summary of future research goals is also given.
- An overview of the tools and workload traces that can be used in the cloud environment to measure how well an algorithm works has been shown.

Overall, the goal is to ascertain how well computers use their resources and consume the least amount of energy possible while still meeting SLA limits for RAM, CPU, bandwidth, etc.

Virtualization

Virtualization technology manages massive data centers more efficiently by allowing several applications, software, and operating systems to run on a single host. It bridges the hardware resources and the operating system, dividing the cloud services into logical units called virtual machines (VMs) [16]. Virtualization solutions such as Xen, VMware, and KVM (Kernel-based VM) are used to construct virtual environments in cloud data centers [17]. Figure 1 displays the classification of energy management techniques.

Energy – efficient cloud computing

Cloud computing offers virtualized resources in cloud data centers for handling several requests for different tasks. A cloud data center's infrastructure often consists of thousands of huge computing hosts with fast processing resources that use a tremendous amount of energy. So, energy-efficient cloud computing is a step forward in analyzing, and implementing global energy reductions in a system providing quality of services while lowering costs [18]. We may conserve energy by consolidating hardware and minimising repetition. If necessary, services should be able to be virtualized and controlled within a data centre, as well as relocated to other locations. To support energy efficiency in the future, machine-readable accounting of the requirements and characteristics of applications, networks, servers, or even entire sites must be available [19]. Energy consumption in a cloud DC organization with m nodes and n switching elements is written as follows [20].

$$E_{Cloud} = m\left(E_{Memory} + E_{CPU} + E_{Disk} + E_{NIC} + E_{Mainboard}\right) + n\left(E_{Chassis} + E_{Ports} + E_{Linecards}\right) + \left(E_{StorageController} + E_{DiskArray} + E_{NASServer}\right) + E_{Others} \tag{1}$$

PUE (Power Usage Effectiveness) is a typical efficiency indicator for data center energy usage that describes how satisfactorily a data center utilizes energy. The PUE formula is well described by eq. (2), which says that it is the ratio of the total energy used in the building to the total energy used by IT equipment in a data center:

$$\text{PUE} = \frac{\textit{Total energy use in the facility}}{\textit{Total energy consumption of IT equipment}} \tag{2}$$

As measured at the meter, the electricity dedicated to the data center facility is included in the total facility energy which includes all loads, such as IT equipment, lighting systems, cooling systems, and power supply components. Total IT equipment includes all the energy used by storage, computing, networking, and other control devices like KVM switches, displays, workstations, and laptops, etc.

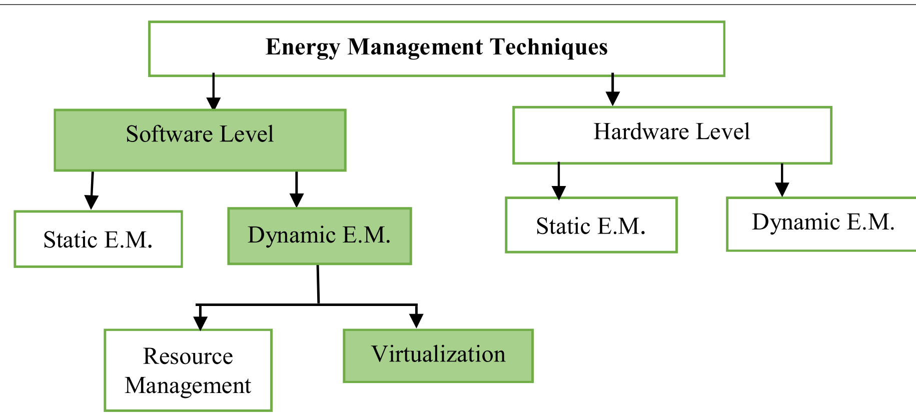

Fig. 1 Classifications of energy management techniques

Energy consumption and service level agreement

Cloud service providers develop an infrastructure where large numbers of high-end computers or servers are installed and interconnected. This hardware platform provides computing, storage, and different amenities to the customer via the internet. As a cloud service provider, the management of power consumption becomes a crucial task. Effective management of resources are required to optimize power utilization, quality of service, cost-effective, and maximize performance with accuracy. In addition to energy utilization and SLA violation, financial expenses and CO_2 emissions from data center cooling systems have a substantial impact on the environment [21].

The most significant challenges in cloud computing are task scheduling, resource utilization, load balancing [9], SLA, quality of service (QoS), scalability, disaster recovery, safety, fault tolerance, resource management, energy efficiency, virtual machine migration, and automated service provisioning [22]. This review work focuses on the previous study of energy efficiency or power consumption, which should be minimized. However, energy and SLAV are inversely associated, as illustrated in Fig. 2. There is a trade-off between energy consumption and performance (QoS). Performance is described in terms of SLA, which defines the standards and services with throughput, service time, delay time, and reaction time given by the deployed system. A simulation for the environment mentioned in [23] is performed using the LrMmt host overload detection method with various safety parameter values. The allocation strategy uses the tuning parameters to anticipate the CPU utilization by the host. For example, if the parameter is set to 1.2, the projected utilization is increased by 20%, providing the host a 20% safety buffer to enhance its consumption without violating SLAs. The results reveal that when this value drops, more VMs are packed into a host. Figure 2 shows that when the safety parameter falls, the EC drops and the number of SLA breaches grows. As a result, the parameters must be set to balance the SLAV and EC.

As a result, cloud providers must cope with the trade-off between energy-performance and reducing energy consumption while fulfilling QoS standards. Buyya et al. [23] showed that when the utilization threshold increases, energy usage is reduced but the percentage of SLAV is also increased. This is because a higher utilization threshold permits more aggressive VM consolidation but at the expense of an increased chance of SLAV. As a result, to save energy, aggressive VM consolidation may result in performance or QoS deterioration, resulting in SLAV. So, while reducing energy utilization, SLAV should also be considered to ensure high adherence to the SLA. To minimize EC and SLAV, the combined metric ESV that captures energy consumption and the level of SLAV is calculated for the performance parameter, as EC decreases with the increased level of SLAV.

VM consolidation

In a cloud data center, a central node routes customer applications to the appropriate servers. This facility is known as VM scheduling. To advance the quality of services and efficient management of power consumption, VM scheduling has been done in such a way that a minimum number of hosts are in a state of running. This method is also known as Dynamic Consolidation of Virtual Machine (DCVM) [23]. Predicting host utilization is an ongoing research effort, and a variety of solutions have been proposed. A single host can host more than one VM, and as per user request, VMs use hosts' resources. When the request of resource host is underutilized or

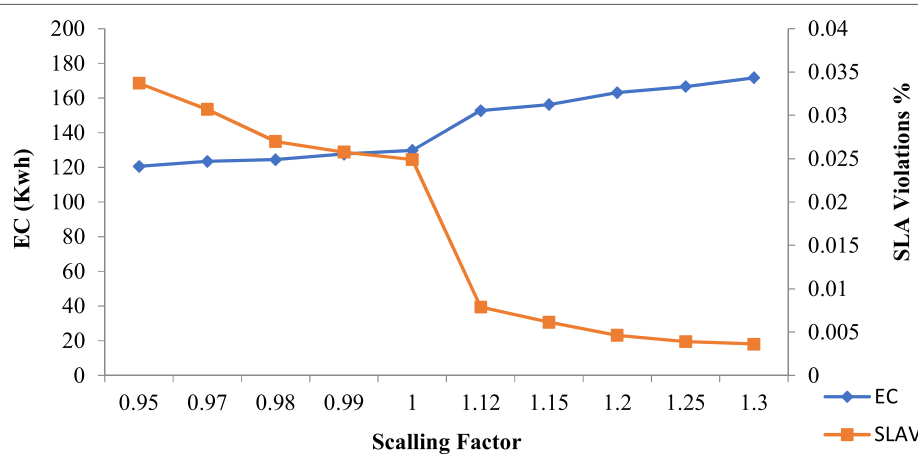

Fig. 2 The output of Energy consumption and SLAV by LrMmt policy

overutilized then VM has to be relocated. This action is known as VM migration and is a popular approach for controlling power consumption. Migration of virtual machines from underutilized and overloaded hosts is a difficult job. To shrink the quantity of VM migrations, appropriate VM selection, and VM placement methods must be developed. When a VM moves from a host that is too busy, both the source host and the new host use power without providing any services.

CloudSim

CloudSim [24, 25] is free, accessible software for simulating cloud computing services and frameworks. This simulator was designed by the CLOUDS (Cloud Computing and Distributed Systems) research laboratory at Melbourne University. Written entirely in Java, CloudSim is a toolkit used to prototype and imitate a cloud computing setting. It enables the modelling of virtualized environments, as well as their administration and on-demand resource management [11]. This simulator is also enhanced to allow for energy-aware models and power models to simulate service applications with variable workloads.

Workload data

As CloudSim simulator is the preferred tool for research where the workload traces of data is used to test the algorithm. Many researchers are working on PlanetLab or Bitbrains data workloads, where a file associated with one VM denotes the CPU utilization of physical machines. Some workload traces include dynamic data such as CPU, RAM, disc, and network I/O values [26]. PlanetLab workload traces [27] with statistical features are given in Table 9. Bitbrains is a cloud service agency that focuses on managed hosting and enterprise business computation [28]. Bitbrains' dataset comprises resources that are used by 1750 VMs from a distributed cloud center. This dataset is published online in the Grid workloads archive [29]. It is divided into fastStorage and Rnd traces. The fastStorage contains 1250 VMs, and Rnd traces have 500 VMs. The fastStorage data is divided into one file per VM, with each file comprising 30 days of data collected every 5 minutes. Bitbrains workload traces with statistical features are given in Table 10 Apart from PlanetLab or Bitbrains, some other workload traces such as Google cluster traces [30, 31], Alibaba cluster [32], Azure trace [33], microservices cluster [34], etc. are also used by researchers. In May 2011, Google released a 29-day cluster trace – a history of every job request, scheduling choice, and resource use statistics for all tasks in a Google Borg computing cluster. The Alibaba group publishes the Alibaba cluster trace program. Their initiative assisted researchers, students, and others interested in the subject by providing cluster traces from the real-world. This allows a better understanding of the features of current internet data centers (IDCs).

Purpose and classification of survey

The Cloud data centers that host and store data are the backbone of cloud computing, which consists of networked computers, power supply, cables, and other components. Data centers that host cloud applications require a lot of energy for resources, leading to high operational costs and carbon release. As expected from a survey, total global energy utilization and carbon dioxide emissions are expected to rise by 48% and 34%, respectively, between 2010 and 2040 [7]. According to a McKinsey analysis [35], "the entire expected energy expense for cloud data centers in 2010 was $11.5 billion, and cost of energy doubles every five years in a typical data center". So, cloud data centers are becoming very expensive and harmful to the environment. The authors of [36] has conducted a systematic examination of the present status of software solution that helps in reduction of energy consumption in data centers and also stated the impact of data centers on the environment. In [37] the use of big data, cloud, and IoT leads to higher demands for hyperscale data centers (HDCs) for data storage and processing. The analysis of 60 regions done by the researchers has predicted the overall increase in the energy consumption of HDCs, carbon emissions and electricity costs, that focus the purpose of the survey.

The main challenge is to set up a balance between system performance and energy utilization [38]. In this detailed systematic survey, a balance between energy efficiency and performance using VM placement [39], VM selection, and migrations [40], has been analyzed for data storage and processing [41]. This paper analyses the approaches performed by various academicians, organizations, researchers in the field of energy consumption in cloud data centers during VM scheduling. Researchers have also compared their method with the benchmark method using different algorithms of heuristics, metaheuristics, machine learning, and statistical methods. Their results show an improvement in energy-saving and thus reduces power consumption. Fig. 3 shows the detailed classification of effective energy management strategies in cloud data centers categorized into four groups i.e., heuristics, metaheuristics, machine learning and statistical. For efficient energy utilization the stages of dynamic VM consolidation is shown in Fig. 4.

Related work

In cloud computing effective energy management strategies related work has been provided by researchers. Many researchers have applied different techniques for VM management and energy-efficient strategies to

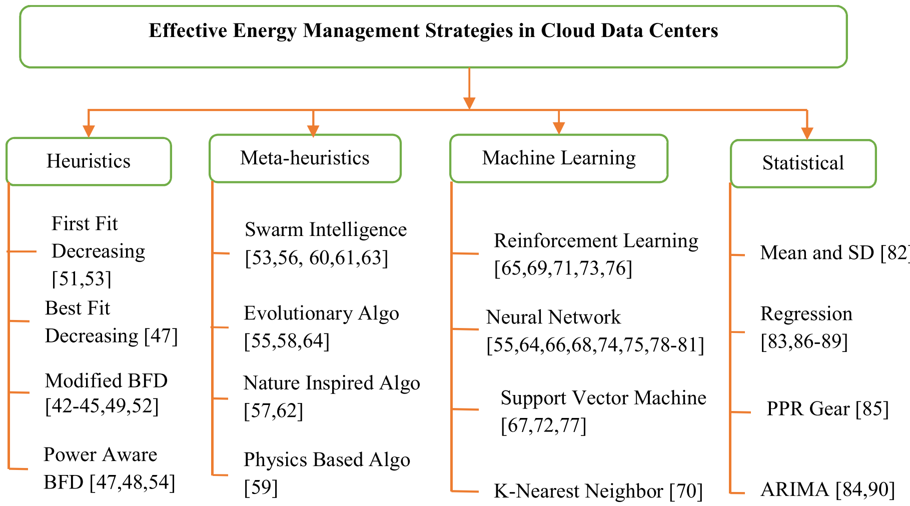

Fig. 3 Classification of Energy Management Strategies in Cloud

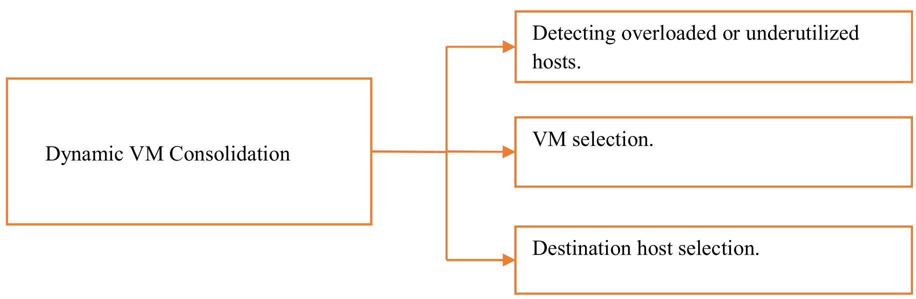

Fig. 4 Stages of Dynamic VM Consolidation

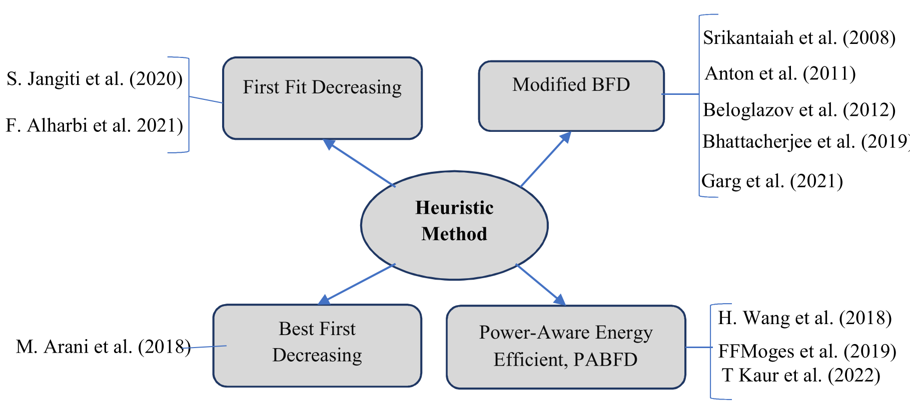

Fig. 5 Classification of existing research using a Heuristic method

reduce energy consumption in cloud data centers. Some have focused on heuristic methods as classified in Fig. 5, some on metaheuristics as classified in Fig. 6 and Fig. 7, some on machine learning as described in Fig. 8, and others on statistical methods categorized in Fig. 9. To balance the load and decrease energy usage, cloud data centers use live VM migration [42]. VMs are dynamically distributed among the hosts during a live migration to reduce the number of low utilization hosts and maximize the number of high utilization hosts. Although dynamic

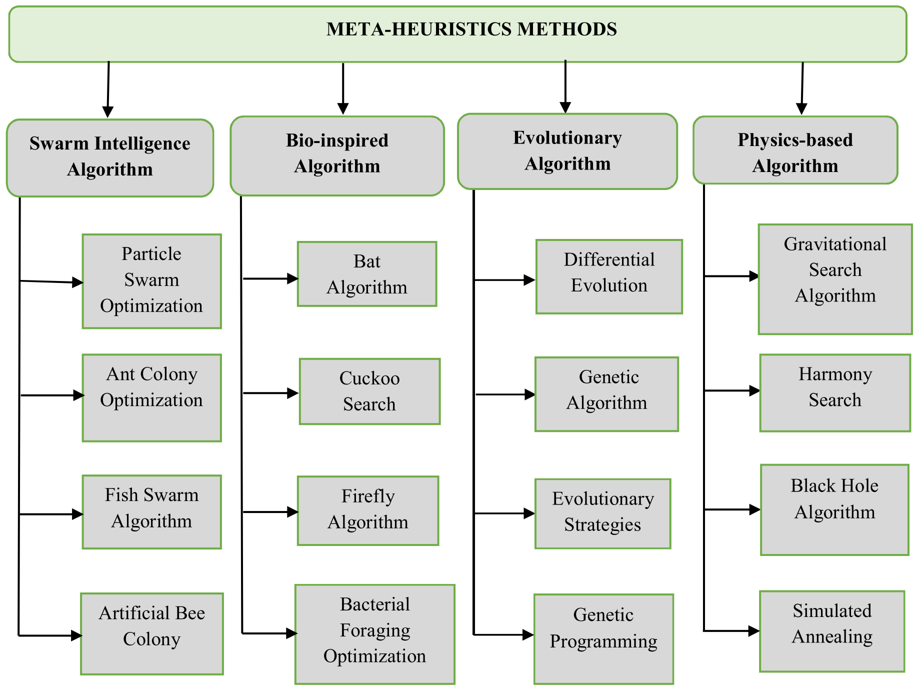

Fig. 6 Classification of Meta-Heuristic methods

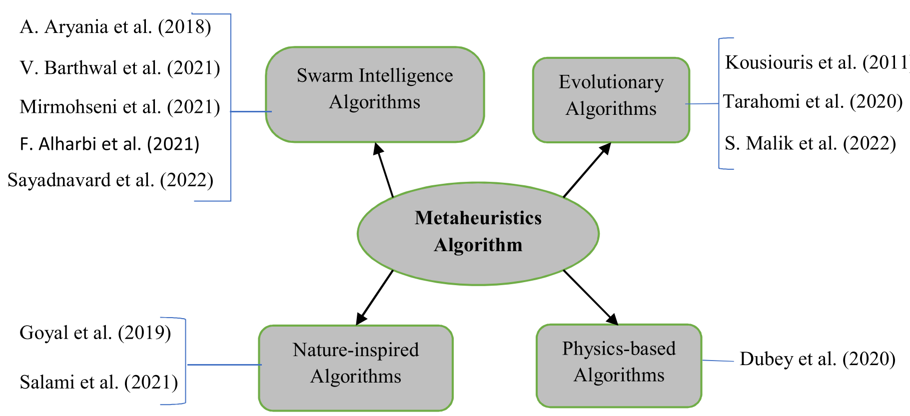

Fig. 7 Classification of existing research using the Metaheuristic method

VM consolidation can significantly reduce energy usage, live migration increases service level agreement violations. As a result, in order to decrease energy usage while satisfying service level agreements, cloud data centers require an effective dynamic VM consolidation solution. The dynamic VM consolidation procedure can often be divided into three parts [43].

Details of different techniques, algorithms, workload data, approaches, and the researchers' work are described below. In first section, the heuristic techniques and their different approaches are used to minimize energy usage. Many researchers work on the first fit decreasing (FFD), best fit decreasing (BFD), modified best fit decreasing, power-aware BFD, etc. Next, the metaheuristic techniques including swarm intelligence, evolutionary algorithm, nature-inspired algorithm, and physics-based algorithm are used to reduce energy consumption and to satisfy service level agreement. Machine learning techniques reinforcement learning, neural network (NN), support vector machine (SVM), and k-nearest neighbor (kNN) are elaborated in further section and finally statistical techniques using mean, standard deviation, regression, PPR gear, and ARIMA are explained.

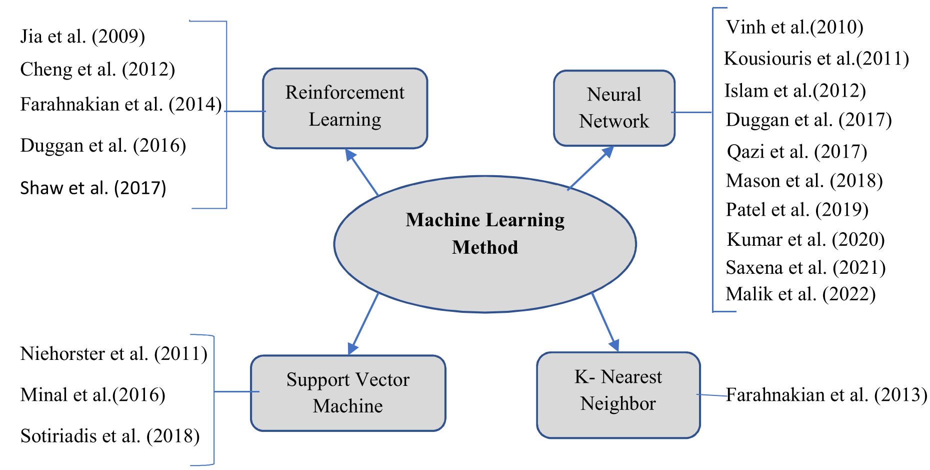

Fig. 8 Classification of existing research using a Machine learning method

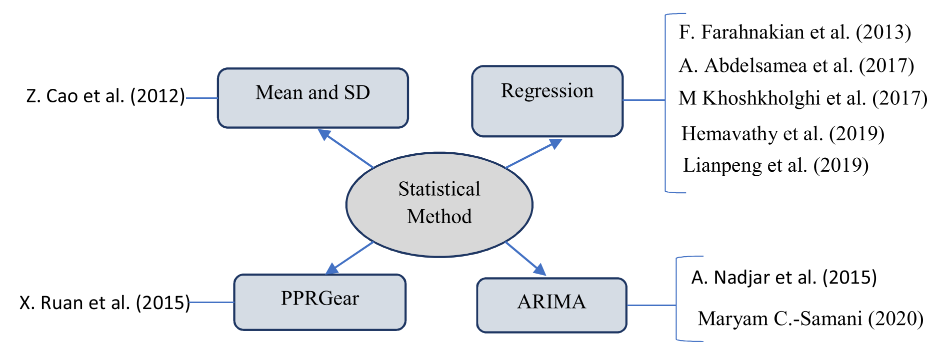

Fig. 9 Classification of existing research using Statistical method

Virtual machine management using heuristic techniques

A heuristic technique is a strategy for solving the problem, that is derived from the Greek term 'eurisko,' which means to search, find, or discover. It is about employing a practical technique that does not have to be perfect. Heuristic approaches reduce the time required to find a satisfactory answer. In cloud computing heuristic techniques are used for VM consolidation. In this approach, different researchers use FFD, BFD, MBFD, PABFD, and other algorithms for VM allocation, migration, and placement to reduce energy consumption.

Srikantaiah et al. 2008 [44], the virtual machine consolidation (VMC) problem was introduced as a bin packing problem. Researchers only examined two criteria: disc and CPU use. The analysis revealed that there is energy-performance compensation for consolidation, with the existence of optimal operating conditions. They constructed a cloud setting, collected data, and developed a bin packing issue using static random threshold values. Other resources, like memory and network, should also be measured, as they may be limiting resources for particular applications.

Beloglazov et al. (2010) [45] For VM consolidation with random data, researchers employed single threshold (ST), minimization of migration (MM), and bin packing strategies. The authors attempted to strike an ideal balance between energy savings and desired performance. They consolidated VMs based on current resource use, network topologies employed in VMs, and thermal status. An energy-aware resource scheduling system based on heuristics for VM allocation and live migration was suggested. The authors structured it as a bin packing issue and evaluated the effort using preset thresholds using the CloudSim toolbox. The results show that dynamic VM consolidation with adaptive thresholds outperforms static thresholds. Non-power-aware (NPA), dynamic voltage frequency scaling (DVFS), and ST methods were used to test the MM algorithm. Using energy savings, the MM algorithm outperformed ST, DVFS, and NPA by 23%, 66%, and 83%, respectively, with thresholds set at 30–70%, resulting in SLA breaches of 1.1%. The MM policy resulted in 6.7% SLA breaches and 43%, 74%, and 87% higher energy savings than the ST, DVFS, and NPA policies when the threshold value was kept at 50–90%.

Anton et al. (2011) [46] provided a heuristic approach for resource distribution that is energy efficient. The policy allocated resources to consumer apps in an energy-efficient manner while ensuring QoS by utilizing energy-efficient mapping heuristics using the consolidation of virtual machines. For VM placement, an improved form of the best fit decreasing modified BFD (MBFD) technique was utilized, as well as three double-threshold VM selection policies, random choice policy (RCP), highest potential growth, and MM. CPU usage data were produced at random by utilizing fixed criteria. The results in the CloudSim toolbox showed that energy consumption was reduced by 77% and 53%, respectively, as compared to NPA and DVFS policies, with SLA breaches of 5.4%.

Beloglazov et al. (2012) [43] Researchers proposed a dynamic VM consolidation approach because fixed thresholds are not feasible in a dynamic cloud environment. The authors reported dynamic threshold values by statistically assessing four histories of CPU use. The reallocation was carried out utilizing a dynamic threshold method. The MBFD technique was utilized to place the VMs. SLA-aware metrics were also examined. The results obtained by running the algorithm on the CloudSim toolkit with a genuine PlanetLab trace demonstrated the validity of the suggested framework. But in the model only single-core CPUs were used, and only a single-core resource CPU was tested.

Arani et al. (2018) [47] by providing a VM placement strategy, researchers concentrated on reducing energy use (VMP-BFD). VMs were mapped to hosts using an approach centred on the best fit decreasing approach, which significantly decreased energy use and SLA violations. The developed algorithm employed the theory of learning automata, correlation coefficients, and the ensemble forecast technique for VM allocation to hosts. The method assigned a VM to a host whose VMs had the least association with the chosen VM for placement. Compared to other reference policies, the results of the simulations on the CloudSim platform showed a big improvement in lowering the energy use and the SLA violations.

Wang et al. (2018) [48] focused on energy-efficient dynamic virtual machine consolidation (DVMC) by introducing an approach for virtual machine placement called "Space-Aware Best Fit Decreasing" (SABFD). The authors also created a VM selection strategy called "High CPU Utilization-based Migration VM Selection" (HS). The suggested system was evaluated in several ways by utilizing the CloudSim toolkit and the Planet Lab workload. The results showed that DVMC designs with a range of SABFD and HS produced the better results.

F.F. Moges et al. (2019) [27] proposed the OpenStack Neat framework's VM placement method to address the issue of consolidation. They introduced VM placement methods that modify heuristics bin-packing to account for host energy efficiency. When linked to the reference algorithms PABFD and MBFD, the proposed algorithms improve energy proficiency. Depending on the host categories and workloads, the energy proficiency improvement over MBFD can be up to 67%. They also defined an innovative bin-packing method termed a "medium-fit" to avoid unnecessary SLAV and VM migrations. The MFPED (medium-fit power-efficient decreasing) offers a lower SLAV and VM migration rate compared to other VM placement methods. SLAV and VM relocations are reduced to 78% and 46%, respectively, when compared to MBFD, depending on the cloud scenario. They used CloudSim to test the suggested algorithms' performance in three different data-center situations: heterogeneous, homogeneous, and default. Data workloads that execute in cloud centers are derived from PlanetLab and Bitbrains cloud traces.

Bhattacherjee et al. (2019) [49] for large historical data sets, proposed prediction technique that was accepted and employed in the current strategy known as the minimization of migration and dynamic thresholding system instead of static thresholds. The MBFD algorithm is used in prediction-based minimization of migration (PMM) to place the VMs. Markov chain learning is applied to formulate the past data for upcoming forecasting deployments. CloudSim 3.0.3 has been used to run rigorous simulations, and the outcomes show a decrease in cloud data center energy utilization.

Xialin Liu et al. (2020) [50] proposed dynamic consolidation by using migration thrashing. It prioritizes VMs with high dimensions and remarkably decreases migration thrashing. The degree of relocations required maintaining service-level agreements (SLAs) by keeping VMs prone to relocation thrashing on the identical physical servers rather than migrating. Their method improves the relocation thrashing measured around 28%, the number of movements measured around 21%, and the SLAV measured around 19%. When the server is overloaded, their solution detects VMs with sufficient capacity by restricting that VMs with excessive capacity are not transferred. Imitations of a wide-ranging research setting employing a workload data set from numerous PlanetLab VMs were used to validate the suggested techniques.

Saikishor Jangiti et al. (2020) [51] DRR-FFD and DRR-BinFill are cutting-edge VMC algorithms based on the concepts of FFD (first-fit decreasing) and DRR (dominant residual resource) that organize VMs based on a single VM resource. Researchers proposed an energy-efficient architecture — EMC2 — for an IaaS cloud service provider. The vector bin-packing techniques VMNeAR-E and VMNeAR-D are proposed. In a python context, simulation tests were conducted utilizing a dataset acquired from the EnergyStar® API for diverse physical servers. The suggested VMNeAR-D heuristic saved up to 3.318% of energy on the average across 40 schedules.

Garg et al. (2021) [52] provided load-aware three-gear THReshold (LATHR) and the MBFD algorithm to reduce overall energy consumption even though they improved service quality in terms of SLA. It produces promising results when used with a dynamic workload and a flexible count of virtual machines (1–290) on each host. The results of the projected work were evaluated concerning service level agreements (SLAs), energy utilization, the number of relocations against various numbers of virtual machines (VMs), and instruction energy ratio (IER). The proposed technique reduces SLA defilements (26%, 55%, and 39%) as well as energy consumption (12%, 17%, and 6%) when related to interquartile range (IQR), median absolute deviation (MAD), and double threshold overload acknowledgement strategies, respectively.

Alharbi et al. (2021) [53] improved existing research that manages data center resources using two independent layers: applications allotted to VMs and VM placement to hosts; both are bin packing problems. This sequential double-layered bin packing (Consec2LBP) solves issues easily and restricts added solution quality development. This research proposes an integrated ant colony optimization strategy to deal with the layers simultaneously to overcome this issue. It converts two-layer resource management into an optimization problem known as integrated double-layer bin packing (Int2LBP). Then, to solve this optimization challenge, a combined FFD technique known as Int2LBP_FFD is derived. To improve the quality of the result, a combined ant colony system, Int2LBP_ACS, has been developed, where the result of Int2LBP_FFD is used as a preliminary solution. In simulations of nine scales of data centers based on GTC data logs, integrated double-layer Int2LBP_FFD outperforms sequential Consec2LBP_FFD. They've also shown that Int2LBP_ACS is better than Int2LBP_FFD concerning energy investments. The Int2LBP_ACS and Int2LBP_FFD algorithms provide scalability.

T Kaur et al. (2022) [54] The Power Aware Energy Efficient Virtual Machine Migration (PAEEVMM) Method has been developed to migrate virtual machines in data centres depending on the temperature threshold value. Based on temperature, this approach moves the heavily loaded virtual machine to the less loaded virtual machine. The simulation was run on CloudSim Plus, and the outcomes are assessed against first fit algorithms. The experiment demonstrates that the suggested approach performs better in terms of CPU and electricity usage.

A brief description of the above detailed literature review and algorithms developed using heuristic methods with different workload data is given in Table 1. Table 2, summarises the work, method, and comparison with their benchmark methods/ algorithm to evaluate energy consumption. Figure 10 depicts the percentage difference in energy reduction or energy savings in graphical form. The implementation of these algorithms has been tested using different settings. The authors have already talked about the host specification, virtual machine description, datasets, simulators, and other criteria for comparing the proposed method to their benchmark algorithm.

Virtual machine management using metaheuristic methods

A metaheuristic is a problem-solving strategy based on a heuristic method that is independent of the problem's nature. A single-solution local search metaheuristic and a random search metaheuristic are the two types of metaheuristic methods. Metaheuristic approaches have been shown to produce near-optimal solutions in a reasonable amount of time and are problem-independent, allowing them to be used in a wide range of situations. It is advantageous in a cloud setting to locate a suboptimal solution quickly. Different metaheuristic techniques based on swarm intelligence, bio-inspired, physics-based, and evolutionary algorithms are used by researchers for VM consolidation to reduce energy consumption. This method was implemented for resource prediction, VM migration, VM placement, load balancing, etc.

Kousiouris et al. (2011) [55] worked on the analysis and performance of VM which depends on several parameters. They proposed the effects on VM performance prediction, persistent allocation proportions, VM co-placement, and instantaneous arrangement on the identical host. They applied a genetic algorithm (GA) to optimize an artificial neural network (ANN) and used linear regression to investigate degradation prediction.

Aryania et al. (2018) [56] proposed a technique using an ACS to resolve the VM consolidation (VMC) issue to reduce energy utilization in data centers. They took into account energy utilization through virtual machine migration. They presented an energy-aware VMC process based on an ACS to handle the VMC issue as a multi-objective optimization challenge. On the arbitrary workload in several circumstances, simulation findings showed that EVMC-ACS increased the number of sleeping hosts by 16% as related to ACS-VMC. Also, the suggested algorithm minimizes relocations by 89%, the

Table 1 Evaluation of Heuristic Methods for Cloud Data Center Resource Management

Author/Year	Algorithm/Method	Data set/ Workload	Tools/ Experiment Environment	Objective	Performance Metrics/Pros	Limitations
S. Srikantaiah et al. [44], 2008	Multi-Dimensional Bin Packing	Own data	Powermeter Xperf	CPU, Disk Utilization	Performance, Energy usage, and Resource utilization	RAM and Network usage are not considered
A. Beloglazav et al. [45], 2010	ST MM Bin Packing	Random data	CloudSim simulator	VM Consolidation	decreases operational expenses while maintaining required QoS, Savings on Energy	Utilization Threshold, Multiple Resources not considered
A. Beloglazov et al. [46], 2012	HeuristicBFD MBFD	PlanetLab	CloudSim	CPU Utilization	Energy Efficiency, QoS, Decreases CO_2 footprint, Cost Saving	Limited Scalability, Slow Optimization, only Simulation
A. Beloglazov et al. [43], 2012	DVMC THR MBFD	PlanetLab	CloudSim	Dynamic VM Consolidation, VM Migration, IaaS env	SLAV, VM Migration, ESV metric, Decreases CO_2 footprint	Complex workload, only simulation
M. Arani et al. [47], 2018	VMP-BFD	PlanetLab	CloudSim	VM Allocation, Learning Automata theory	Energy Consumption, SLAV, Migration Count	Required NN for better efficiency.
H. Wang et al. [48], 2018	SABFD HS	PlanetLab	CloudSim	VM Placement VM Migration, DVMC, Host overload detection	Energy Efficient, Suppressing SLAV, SLATAH, PDM, ESV, VM Migration, Host Shutdown	Real-world challenges, only simulation performance
F F Moges et al. [27], 2019	MFPED	Planet Lab Bitbrains	CloudSim	VM Placement, VM Migration	Energy Consumption, SLAV, VM Migrations Count	N/W devices and traffic effects are not considered
S. Bhattacherjee et al. [49], 2019	PMM-MBFD	PlanetLab Parallel Archive	CloudSim	VM Placement, Prediction based migration	Energy Consumption, Dynamic Thresholding	RAM and N/W uses are not considered, multi-objective optimization should be required
X. Liu et al. [50], 2020	DCMMT	PlanetLab	CloudSim	VM Migration Thrashing, Dynamic Consolidation	VM Migration, SLA Violation, Thrashing Index, SLATAH, PDM	No real-world cloud platform, required workload statistical properties
S. Jangiti et al. [51], (2020) and	EMC2, VMNeAR-D, VMNeAR-E	dataset extracted from EnergyStar® API	Python environment	multi-resource fairness, virtual machine consolidation	Energy-efficient, VMP, time complexity, IaaS	VM swapping into smaller PMs in case of very low resource occupancy in a huge PM
V. Garg et al. [52], 2021	LATHR MBFD	HPG4, 100 hosts, 290 VM	Matlab	VM Migration, Overload Detection Policy	QoS, Energy Consumption, IER, No. of Migration, SLAV	Limited workload, real-world implementation
F. Alharbi et al. [53], 2021	Int2LBP_FFD Int2LBP_ACS	GTC data logs	Java	Resource Management	QoS, Energy Saving, Integrated approach, Scalable	Dynamic decision required, Public CDC, Runtime VMC
T Kaur et al. [54], 2022	PAEEVMM	Dynamic data by user	CloudSim Plus	Temperature Threshold	CPU utilization, Power Usage	Load Balancing, Multiple Resources

Table 2 Comparison of benchmark concerning Energy Consumption for Table 1

S.No.	Reference	Year	Algorithm/Method	Benchmark Algorithm	Energy reduced in %
1	[44]	2008	MDBP	Optimal	5.4
2	[45]	2010	MM	ST	23
3	[46]	2012	MBFD	DVFS	53
4	[43]	2012	THR-MMT	DVFS	87
5	[47]	2018	VMP-BFD	PABFD	10.27
6	[48]	2018	SABFD	PABFD	72
7	[27]	2019	MFPED	MBFD	67
8	[49]	2019	PMM	MM	37
9	[51]	2020	VMNeAR-D	DRR-Binfill	3.318
10	[52]	2021	MBFD	THR	06
11	[53]	2021	Int2LBP_FFD	Consec2LBP_FFD	90

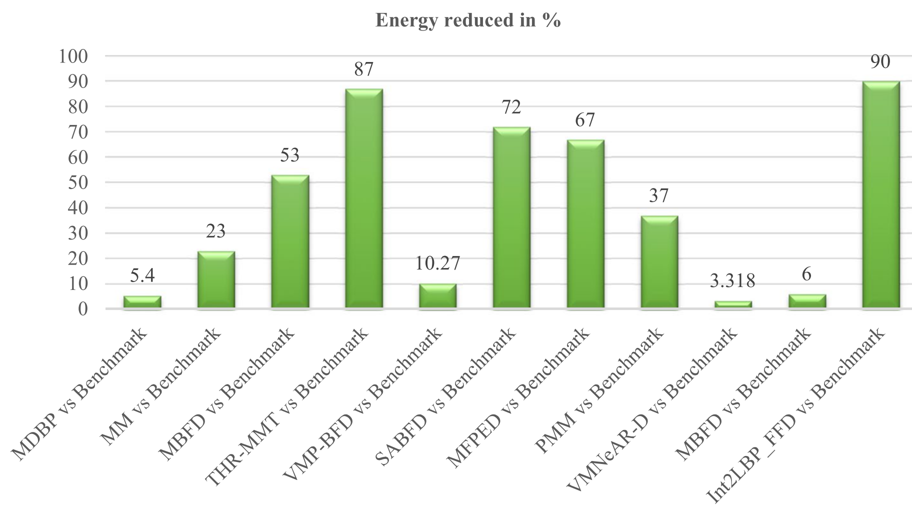

Fig. 10 Energy reduction using Heuristic techniques vs Benchmark

power consumption during a migration by 91%, SLA violations by 79%, and overall energy consumption by 25% relative to ACS-VMC.

Goyal et al. (2019) [57] worked on PSO and CSA algorithms. The goal of optimizing energy utilization in the cloud is also addressed in the article. CloudSim simulators and common programming languages were utilized in their suggested work. Several performance measures, such as energy efficiency, response time, and execution time, were used to judge how well the work performs.

M Tarahomi et al. (2020) [58] approached micro-genetic method for choosing the right physical host for a virtual machine. Their simulations reveal that the micro-genetic method enhances power consumption relations. The suggested approach was tested using CloudSim and their result was related to the reference algorithms (genetic and PABFD VM provisioning algorithms) in various scenarios with the datasets of 10 working days. According to experimental results by the CloudSim framework, the micro-genetic system reduced power consumption.

Dubey et al. (2020) [59] suggested a virtual machine placement approach that reduces the makespan while reducing power consumption. The proposed technique was tested in the simulator CloudSim toolkit, and the findings proved that it exceeded typical work utilizing FCFS, Round-Robin, EERACC, and Random algorithms. The result shows that the recommended technique beats the other four mentioned methods regarding energy and power usage, server utilization, and makespan.

Barthwal et al. (2021) [60] proposed AntPu ACO metaheuristic predicted utilization for dynamically placing VMs in the cloud data center to minimize SLAV and energy utilization (EU). In CloudSim, a simulated environment is created, and the PlanetLab dataset is chosen because of its real-world properties. The CPU usage of VMs in five-minute intervals is shown in this data set. To assess the results, extensive simulations were run, showing that the proposed approach offers a significant improvement in energy utilization and SLA compared with other methods. AntPu improves performance by satisfying SLA, QoS, EC, VM migration, and PM overloading constraints.

Mirmohseni et al. (2021) [61] combined the outcomes of the particle swarm genetic optimization (PSGO) process. The findings were improved and a viable solution for load balancing operations was introduced by combining the advantages of these two algorithms. Instead of arbitrarily assigning the beginning population or data set in the GA, the most acceptable outcome is obtained by giving the starting population in their proposed approach, load balancing PSGO Improve Resource Allocation (LBPSGORA). The LBPSGORA method is compared to GA, PSO, and a hybrid GA-PSO approach. This method outperformed similar methods in terms of execution cost, load balancing, and time to completion. With task changes, the hybrid GA-PSO approach performs similarly to the suggested method. The LBPSGORA technique is 7.32% more effective in makespan and 6.87% more effective in execution cost compared to the hybrid GA-PSO. LBPSGORA outperformed the hybrid GA-PSO by 8.42%, GA by 10.61%, and PSO by 11.71% in terms of load matching.

Alharbi et al. (2021) [53] improved existing research that manages data center resources using two independent layers: applications allotted to VMs and VM placement to hosts; both are bin packing problems. This sequential double-layered bin packing (Consec2LBP) makes easier the issue solving and restricts added solution quality development. This research proposes an integrated ant colony optimization strategy to deal with the layers simultaneously to overcome this issue. It converts two-layer resource management into an optimization problem known as integrated double-layer bin packing (Int2LBP). Then, to solve this optimization challenge, a combined FFD technique known as Int2LBP_FFD was developed. To improve the quality of the result, the combined ant colony system Int2LBP_ACS is refined further using the Int2LBP_FFD result as a preliminary solution. In simulations of data centers based on GTC data logs, Int2LBP_FFD outperforms Consec2LBP_FFD. They've also shown that Int2LBP_ACS is better than Int2LBP_FFD concerning energy investments. The Int2LBP_ACS and Int2LBP_FFD algorithms provide scalability.

Salami et al. (2021) [62] offer a virtual machine placement problem (VMPP) based on the cuckoo search (CS) algorithm. New cost and perturbation metrics have been created to increase the algorithm's performance. Two well-known benchmark datasets were used to evaluate the suggested technique. The main objective is to organize virtual machines into actual machines to minimize the number of devices required. It beat the reordered grouping genetic algorithm and the FFD, BFD, and multiCSA, an older CS approach.

M. H. Sayadnavard et al. (2022) [63] approached a technique for dynamic VMC, which included a prediction model based on DTMC, a VM selection algorithm, and e-MOABC-based VM placement. Using this model in conjunction with the dependability model of PMs results in a more exact classification of PMs depending on their condition. Then, a multi-objective VM placement approach is proposed using the e-dominance-based multi-objective artificial bee colony algorithm to find the optimum VMs to PMs mapping, which can efficiently manage overall energy consumption, resource usage, and system performance to meet SLA and QoS requirements. By completing a performance assessment study with the CloudSim toolkit and PlanetLab workload traces, the proposed system is proved to be effective. The suggested technique greatly decreases energy usage while avoiding excessive VM migrations, according to a competitive analysis of the experimental findings. The investigation of various parameters reveals that the suggested approach outperforms other algorithms. MOABC-VMC decreases energy consumption by 11.35% and 35.25%, respectively, when compared to RE-VMC and LR-MMT.

S. Malik et al. (2022) [64], proposed Evolutionary Algorithms and Machine Learning Methods to Predict Resource Utilization in cloud data centers. The primary goal was to resolve the over-and under-provisioning problems. Over-provisioning of resources results in higher expenses and increased energy use. However, under-provisioning results in SLA violations and a decline in quality of service (QoS). The research focuses on functional link neural networks (FLNN) using hybrid Genetic Algorithms (GA) and Particle Swarm Optimization (PSO) for multi-resource usage prediction. The suggested model produces improved accuracy when compared to conventional procedures, according to experimental results using data from Google Cluster Traces. This study's primary objective was to examine how well neural networks predicted multi-resource allocation. The proposed model predicts using FLNN and trains the network weights using a hybrid GA-PSO.

To manage a large number of users, resources must be dynamically scaled for effective usage, low energy consumption, low cost, and higher quality of service (QoS).

A brief report of the above detailed literature review and algorithms mentioned using metaheuristic methods with different workload data is given in Table 3. Table 4, summarises researchers work, methods, and comparison with their benchmark algorithm to evaluate energy consumption. Figure 11 depicts the percentage difference in energy reduction or energy savings in graphical form. The implementation of these algorithms has been tested with different settings. About the host specification, virtual machine characteristics, workload datasets, simulators or tools, and other measures for comparing the proposed method to their benchmark algorithm has already been discussed earlier.

Virtual machine management using machine learning techniques

Machine learning technique are approaches and set of technologies that use AI concepts. Machine learning enables researchers to use data to train a system on how to solve a problem using machine learning algorithms and improve over time. Machine learning is frequently classified by how an algorithm learns to improve its prediction accuracy. Supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning are the four fundamental methodologies. In a cloud computing environment reinforcement learning, neural network, k-nearest neighbor, and support vector machine algorithm are used by researchers to consume less amount of energy in cloud environment.

Jia et al. (2009) [65] have proposed a reinforcement learning method called VCONF, which automates the VM configuration process by addressing the system's scalability and adaptability problems. By learning from repetitions with the environment, virtual machine configuration (VCONF) generates policies for the autoconfiguration of VMs. This method achieves the best cloud setup while improving adaptability and scalability also. Experimental results demonstrated the system's optimality in controlled problems, as well as its scalability and adaptability in a broader system. VCONF could be changed to a good configuration in seven steps and showed a 20% to 100% increase in throughput over simple RL approaches.

Vinh et al. (2010) [66] developed an energy-aware algorithm that uses a neural network (NN) to forecast upcoming load requirements built on previous data and reduces the number of hosts by shutting them down or restarting them as needed. Their research objective is to moderate the energy used in data centers. When the system load increases or decreases, the system turns on or off some hosts.

Niehorster et al. (2011) [67] have presented an approach for the provisioning of virtual machines using support vector machines (SVM). They created a self-configurable and self-optimized multi-agent system capable of learning its behaviour and estimating its cost. The system acquires performance models for various applications and develops a behaviour model, after which SVM is used to organize the data in the knowledge base.

Kousiouris et al. (2011) [55] depend on several parameters on VM performance prediction, persistent allocation proportions, VM co-placement, and instantaneous arrangement on the identical host. They used a genetic algorithm (GA) to improve an ANN and linear regression to study how well it could predict degradation.

Islam et al. (2012) [68] constructed a model for predicting future CPU resource requirements using the linear regression method. The input data set used historical data obtained by performing the Transaction Processing Performance Council (TPC), a typical client-server benchmark. To train the algorithm for prediction, the CPU utilization percentages of all VMs are used. They also used a neural network in the cloud for resource allocation and management. The neural network was trained with the back-propagation process, and experimental outcomes showed that NN-approximate predictions have a lower proportion error than LR-based predictions.

Cheng et al. (2012) [69] proposed a unified reinforcement learning technique for autonomously configuring virtual machines and their applications and adjusting the VM resources efficiently and providing quality service assurance. They came up with a good plan for running their research on Xen VMs using different workloads.

Farahnakian et al. (2013) [70] introduced a dynamic consolidation of virtual machines (DCVM) where the active number of hosts are minimized based on present and historical use. The k-nearest neighbour (KNN) method is used to forecast each host's CPU utilization. To optimize dynamic VM consolidation, their prediction technique focuses on identifying overloading and underloading of hosts. The results indicated that their system consumes the least amount of energy while maintaining the SLA.

Farahnakian et al. (2014) [71] suggested a Reinforcement learning (RL) technique for dynamic consolidation of VM that uses a learning agent to find out the host's power strategy. The agent selects the host to make it active or sleep. The RL learning agent optimizes the active host by learning system behavior. Experiments with PlanetLab workload traces show that their model lowers the cost of using energy, improves performance, and cuts down on SLA violations.

Table 3 Evaluation of Meta-Heuristic Algorithms for Cloud Data Center Resource Management

Author/ Year	Algorithm/Method	Data set/ Workload	Tools/ Experiment Environment	Objective	Performance Metrics/ Pros	Limitations
G. Kousiouris et al. [55], 2011	GA, ANN, LR	6 Matlab Benchmark tests	MatlabR2007b	VM performance, VM Analysis	Scheduling decisions, co-placement of VMs	No real-world application, 5% margin of error, premature convergence.
A. Aryania et al. [56], 2018	EVMC- ACS	Random Workload	Java	VM Consolidation, VM Migration	SLAV, Energy Consumption, migrations, sleeping PMs	No real workload, EC during VM migration not considered.
Goyal et al. [57], 2019	PSO and CSA	Cloudlets jobs	CloudSim	Resource migration, Utilization	Energy Consumption, Response time, and Execution time	Not for Hybrid Energy efficient model, SLAV not addressed.
M. Tarahomi et al. [58], 2020	Micro-GA	PlanetLab	CloudSim	VM Placement	Power Consumption, SLAV, VM migration, host shutdown	Only simulation, need real data center environment, required OpenStack-based cloud data center.
Dubey et al. [59], 2020	SA	Xen server	CloudSim	VM Placement, Resource Utilization	Power Consumption, Makespan, Mapping VM to PM	Static approach, Dynamic VM problem, Actual load during run time not considered
V. Barthwal et al. [60], 2021	AntPu ACO MH.	PlanetLab	CloudSim	VM Placement, multi-objective optimization	Energy Consumption, SLA Violations, PM Overloading, VM migration	Memory, disk, and B/W usage were not considered to predict the PMs utilization more accurately.
S M Mirmohseni et al. [61], 2021	PSGO LBPSGORA	Own data	Matlab, CloudSim	Load Balancing,	Cost, Energy Consumption, Resource management	Complex, real environment.
F. Alharbi et al. [53], 2021	Int2LBP_FFDInt2LBP_ACS	GTC data logs	Java	Resource Management	QoS, Energy Saving	Only static decision, Public CDC, Runtime VMC
Salami et al. [62], 2021	CSA	Benchmark datasets	MatlabR2018b	VM Placement, new cost, and perturbation functions are introduced	Power Consumption, Execution time, cost/fitness computation, servers required for VMP	Disk and bandwidth usage not considered, required VM placement with more resources, hybridizing new CS with other metaheuristics
M H Sayadnavard et al. [63] (2022)	MOABC-VMC	PlanetLab	CloudSim	DVMC, Prediction model, VMP	SLATAH, PDM, SLAV, EC, VM migrations, ESV.	Resource overcommitment and B/W resource constraints, static cloud environment
S. Malik et al. [64], 2022	GA, PSO	Google cluster traces	Simulation	Multi Resource utilization	Prediction of Resources, Accuracy, Resource Utilization	Predicting disk utilization, cost-effectiveness, and network, Multi-variate resource utilization datasets

Table 4 Comparison of benchmark concerning Energy Consumption for Table 3

S.No.	Reference	Year	Algorithm/Method	Benchmark Algorithm	Energy reduced in %
1	[56]	2018	EVMC- ACS	ACS-VMC	25
2	[58]	2020	Micro-GA	GA	7.68
3	[60]	2021	AntPu	PABFD	24.59
4	[61]	2021	LBPSGORA	PSO	11.71
5	[53]	2021	Int2LBP_ACS	Consec2LBP_FFD	97
6	[63]	2022	MOABC-VMC	LR-MMT	35.25

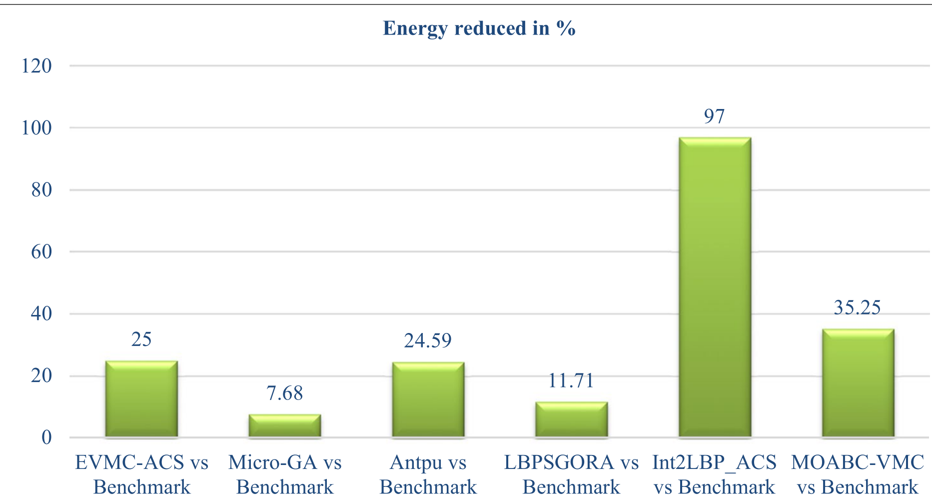

Fig. 11 Energy reduction in Metaheuristic techniques vs Benchmark

Minal et al. (2016) [72] Configure live VM migration using a support vector regression (SVR) model to forecast dirty pages using time series analysis. The service interruption time and migration duration were used to assess the performance of the live migration. They also created an ARIMA-based model, and findings show that SVR outperforms ARIMA in predicting dirty pages. Total pages transferred and migration time are the two most critical performance criteria for live migration in their proposed system.

Duggan et al. (2016) [73] developed a network-aware live migration technique that monitors bandwidth usage and takes appropriate action when there is network congestion based on experience Their structure functions as a decision support system, enabling a mediator to schedule VM migrations by determining the best time to do so. The amount of bandwidth available in the data center influences the migration process. According to their research findings, an agent in a cloud data center can learn available bandwidth during peak network capacity and schedule the migration of VMs from underutilized Hosts at the appropriate time using available bandwidth. They used the local regression approach to determine which hosts were overloaded. The Learning agent selects the best VM for migrating from an overloaded host while balancing migration and energy consumption. The findings of the research point to an autonomous VM selection method that can account for VM migration count and energy cost.

Duggan et al. (2017) [74] To create reliable predictions using time series data, researchers employed a recurrent neural network (RNN) to forecast future values of CPU consumption. They looked into the network's accuracy for prediction with a deep effect. Experiments have shown that it is possible to get a very accurate estimate of CPU usage for dynamic data sets that change.

Qazi et al. (2017) [75] provided a real-time resource consumption prediction classification that takes actual resource usage and sends it to multiple buffers built on time and resource type. A system with real CPU utilization traces from a cloud data center with 120 servers used the autoregressive neural network method on data blocks where the data did not track a Gaussian distribution. The experimental findings suggest that AR-NN outperforms ARIMA for a given data set.

Shaw et al. (2017) [76] have presented the advanced RL consolidation agent method for VM allocation that is capable of optimizing VM circulation in the cloud data center while saving large amounts of energy and lowering SLA violations. They established a space for state-action. Action is defined as a combination of any host's utilization rate and the size of the VM to be deployed, and state is well-defined as the entire active host as a percentage of the total host.

Sotiriadis et al. (2018) [77] proposed a VM scheduling strategy that uses extracted data from past VM and host resource utilizations to define host weights based on the resource utilization of hosted VMs on that host. They used SVM to classify VM states based on historical records. They used the resource utilization dataset (percentage of CPU, RAM, and disc usage) in the X-Y planes and expressed the data as vectors. The results of the experiments reveal that, through learning the system's behavior, their method improved physical machine selection.

Mason et al. (2018) [78] using evolutionary NN, created a way to forecast the host's CPU utilization. For network training, optimization approaches such as particle swarm optimization (PSO), differential evolution (DE), and covariance matrix adaptation evolutionary strategy (CMA-ES) are used. The outcomes of the experiments showed that CMA-ES performs better than other optimization strategies and trains networks to predict CPU consumption accurately.

Patel et al. (2019) [79] presented a load-balancing method based on energy-aware VM Migration. They perform it by assigning a lower and higher threshold to an individual host, which specifies whether the host is underloaded or overloaded. Before initiating the migrations, they used a prediction approach that predicts the demand on the host. Their process uses an artificial neural network (ANN) with the dynamic double threshold (DDThr) technique to predict VM movement and energy consumption while considering CPU utilization. Not only does it reduce the number of VM movements, but it also saves energy. Graphs comparing VM movement and energy utilization show that when ANN is combined with existing techniques, both VM movements, and energy utilization decrease slightly, saving a significant amount of electricity. To create a cloud environment, the CloudSim simulator was employed, and Matlab2015a was used to implement ANN. Based on the experiments, the proposed strategy uses less energy and has fewer migrations than the competitive approach.

Kumar et al. (2020) [80] provide a workload forecasting framework based on a NN (WFNN) model with supervised learning. To increase the predictive model's learning efficiency, an upgraded and adaptable differential evolution method has been designed and developed. The algorithm determines the most appropriate crossover and mutation operators. Because of its adaptive nature in pattern learning from sampled data, the learning's prediction accuracy and convergence rate have been seen to improve. The prediction model's performance is assessed using real-world data traces from Google's cluster and NASA's Kennedy Space Center. A Python3 Jupyter notebook is used to implement the suggested model. The results are compared with other recent methods, and improvements of up to 97%, 91%, and 97.2% are observed over backpropagation, self-adaptive differential evolution, and average-based workload prediction techniques, respectively.

Saxena et al. (2021) [81] introduce an energy-efficient resource provisioning and management system to satisfy future applications' dynamic demands. The proposed system addresses power consumption, performance, resource wastage, and QoS depletion by accurately matching the application's expected resource demand with VM resource capacity. Consequently, condensing the whole load onto the smallest number of energy-efficient physical machines (PMs). The proposed work makes contributions in the form of online multi-resource feed-forward NN (OM-FNN) to predict resources, autoscaling of VMs, and allocation of scaled VMs on energy-efficient hosts. The suggested integrated solution has been rigorously evaluated using real resource usage traces from the Google cluster dataset, and it outperforms the other VMPs in terms of resource utilization and power savings by up to 21.12% and 88.5%, respectively. Also, the OM-FNN predictor is more accurate, takes less time, and uses less space than the single-input single-output feed-forward NN predictor.

Malik et al. (2022) [64] focuses on employing a hybrid Genetic Algorithm (GA) and Particle Swarm Optimization with a Functional Link Neural Network (FLNN) to anticipate the multi-resource utilization (CPU, memory, and network bandwidth). For resource usage prediction, the programme employs models from convolutional neural networks (CNN) and long short-term memory (LSTM). Experimental findings using Google cluster traces demonstrate that the suggested model outperforms conventional methods in terms of accuracy. This study's major objective was to examine how well neural networks forecast the use of several resources. FLNN is used for prediction, while hybrid GA-PSO is used to train the network weights. Therefore, to manage a high number of users, the resources need to be scaled dynamically for optimal use, decreased energy consumption, and cost, with better quality of service (QoS).

Table 5 Evaluation of Machine Learning Algorithms for Cloud Data Center Resource Management

Author/ Year	Algorithm/Method	Data set/ Workload	Tools/ Experiment Environment	Objective	Performance Metrics/ Pros	Limitations
J. Rao et al. [65], 2009	RLVCONF	TPC	A testbed of cloud with Xen VMs	VM Auto-Configuration, optimize VMs performance	SLA, Throughput, Response Time, Adaptability, and Scalability	Limited samples quality in model training, N/W, and disk B/W not included.
T. Vinh et al. [66], 2010	Neural Network	NASA, ClarkNet	CloudSim, GridSim	PM Selection, Load prediction	Energy Consumption, drop rate, predictor	Less diversity of workloads and application services
O. Niehorster et al. [67], 2011	SVM	RUBiS	Libvirt 0.8.3 interface, Eucalyptus	Utilization Prediction, Provisioning of VM, SaaS, private cloud	Service Level Objectives (SLO), QoS, self-optimizing, Autonomic Resource Management	Required parallel learning in larger cloud environments, cost estimate misses the SLO, more dataset required.
G. Kousiouris et al. [55], 2011	ANN LR	6 Matlab Benchmark tests	MatlabR2007b	VM performance, VM Analysis	Scheduling decisions, placement of VMs	Detection of workload, Real-world application, premature convergence
S. Islam et al. [68], 2012	NN LR	TPC-W	Java	CPU Prediction, resource management, forecasting resource utilization	SLA Fulfilment, Mean Absolute Percentage Error Root Mean Squared Error, R^2prediction accuracy,	Need more variety of workload generators, required utility functions for prediction, cost, performance.
C. Xu et al. [69], 2012	RL URL	TPC	Real environment	VM Configuration	Service Quality, Throughput, SLA assurance, system utilization	Time complexity, less impact on the quality of final configuration, and traffic perturbations deserve further investigations.
F. Farahnakian et al. [70], 2013	DC-KNN	PlanetLab	CloudSim	Utilization Prediction	SLA Violation, Energy Consumption	RAM and N/W resources not included, required K-NN regression for predicting overutilized and underutilized hosts
F. Farahnakian et al. [71], 2014	RL-DC	PlanetLab	CloudSim	Dynamic Consolidation	Energy Consumption, SLA Violation	Real environment required, optimum solution through trial and error in a dynamic context
M. Patel et al. [72], 2016	SVR	Real Data set of Xen	R language	Predict Dirty page	Migration Time, Total transferred pages	Model can be overtrained or undertrained, no live migration
M. Duggan et al. [73], 2016	AI tech RL RLLM	PlanetLab	CloudSim	Network-aware live VM Migration strategy	Energy Consumption, VM Migration, SLA Violation, PDM, ESV, performance	No real-world cloud applications,
M. Duggan et al. [74], 2017	RNN	PlanetLab	CloudSim	Predict CPU utilization	CPU Utilization, Energy Efficiency, Economy of Scale	RAM and disk utilization required, Back propagation through time prediction accuracy

Table 5 (continued)

Author/ Year	Algorithm/Method	Data set/ Workload	Tools/ Experiment Environment	Objective	Performance Metrics/ Pros	Limitations
Q Z Ullah et al. [75], 2017	ARNN	FastStorage	rJava	CPU Prediction usage	CPU Resource Utilization	Need more duration for prediction, size of training data, type of prediction patterns, temporal and spatial complexity
R. Shaw et al. [76], 2017	ARLCA	PlanetLab	CloudSim	Resource management	Energy Consumption, SLA Violation	RAM and N/W bandwidth are not considered, multi-objective optimization techniques are required.
S. Sotiriadis et al. [77], 2018	SVM	YCSB	OpenStack	VM Scheduling, VM placement	CPU Utilization, performance, CPU steal time, prediction of VM resource	Need model for classification and regression, time frame window, behavior of VMs and PMs.
K. Mason et al. [78], 2018	EvolutionaryNN CMA-ES	PlanetLab	CloudSim	Predict CPU consumption, performance	CPU Utilization, Mean Absolute and Squared Error, Multi-Step Prediction Accuracy	RAM and disk utilization not included, prediction accuracy of a system trained only on the PlanetLab.
D. Patel et al. [79], 2019	DT with ANN	Real System	CloudSim, Matlab2015a	VM Migration based Load balancing, performance, and accuracy	Energy Consumption, CPU Utilization, VM Migration	CPU, bandwidth, RAM parameter are not considered, ANN should be integrated with a cloud server for continuous load assessment.
J. Kumar et al. [80], 2020	WFNN	Google Cluster Trace, NASA, and Saskatchewan servers' weblogs	Python3 Jupyter notebook	QoS, Resource Utilization,	Performance, predict upcoming workload with precision, Accuracy, forecast accuracy, faster convergence	Forecast multivariate workload traces, computational complexity, high computation costs
D. Saxena et al. [81], 2021	NN	Google cluster dataset	Python version 3	Resource provisioning, VM Placement, VM Allocation	Performance, Power consumption, QoS, Resource Usage	Network traffic not included, manual selection of nodes in the I/O layer of OM-FNN predictor, more EC due to communication-intensive VMs.
S. Malik et al. [64], 2022	FLNN, CNN	Google cluster traces		Multi Resource utilization	Prediction of Resources, Accuracy, Resource Utilization	Predicting disk utilization, cost-effectiveness, and network, multi-variate resource utilization datasets.

Table 6 Comparison of benchmark concerning Energy Consumption for Table 5

S.No.	Reference	Year	Algorithm/ Method	Benchmark Algorithm	Energy reduced in %
1	[66]	2010	NN PP Mode	NM	46.7
2	[70]	2013	DC-KNN	LR	1.6
3	[71]	2014	RL-DC	LR	12.5
4	[73]	2017	AI tech RLRL-LM	Lr-Mmt	3
5	[76]	2017	ARLCA	Lr-Mmt	44.7
6	[79]	2019	ANN	DDT	2.4
7	[81]	2021	OM-FNN	BF-VMP	88.5

A brief description of the above detailed literature review and algorithms developed using machine learning methods with different workload data is given in Table 5. Table 6, summarises the work, methods, and comparison with their benchmark algorithm to evaluate energy consumption by different researchers. Figure 12 depicts the percentage difference in energy reduction or energy savings in graphical form. The implementation of these algorithms has been tested for different settings. The authors have already mentioned the host specification, characteristics of virtual machine, workload datasets, simulators environment, and other criteria for comparing the proposed method to their benchmark algorithm.

Virtual machine management using statistical techniques

Statistical methods are used in research planning, analyzing, data collecting, meaningful interpretations, and reporting the findings of various virtual machine management. In cloud computing researchers work on mean, standard deviation, regression, ARIMA, PPRGear, etc. to detect overload and underload hosts, resource prediction, VM allocation, VM migration, and VM placement to save energy consumption.

Cao et al. (2012) [82] proposed strategies for dynamically combining VMs in a virtualized data center to reduce SLAV and energy utilization. The authors suggested detecting host overload, VM selection, and allocation strategy. The author's uses mean and standard deviation CPU utilization metrics to determine overloaded hosts. The extension of the maximum correlation (MCE) strategy was utilized to select VMs for migration with mean and variance-related computations for VM allocation. Experiments using PlanetLab traces on CloudSim revealed that the new framework, which consists of the policies listed above, outperforms the previous policies in the requisites of energy utilization and overall QoS. However, it performed slightly worse in the requisites of energy utilization. As a result, managing the energy-performance trade-off is difficult.

Farahnakian et al. (2013) [83] Using PlanetLab historical data, a linear regression method was proposed to forecast the upcoming CPU use of the host (LIRCUP). Authors discovered a relationship between expected and current CPU use, where expected utilization is a dependent variable and current utilization is an autonomous variable. The LIRCUP algorithm detects overloaded hosts and maintains SLA and energy utilization by transferring some VMs from the overburdened hosts by comparing the expected CPU utilization value with the present utilization.

Nadjar et al. (2015) [84] present a decentralized scheduling strategy for DCVMs fitted with an auto-regressive integrated moving average (ARIMA) technique to progress resource provisioning by predicting VM resource

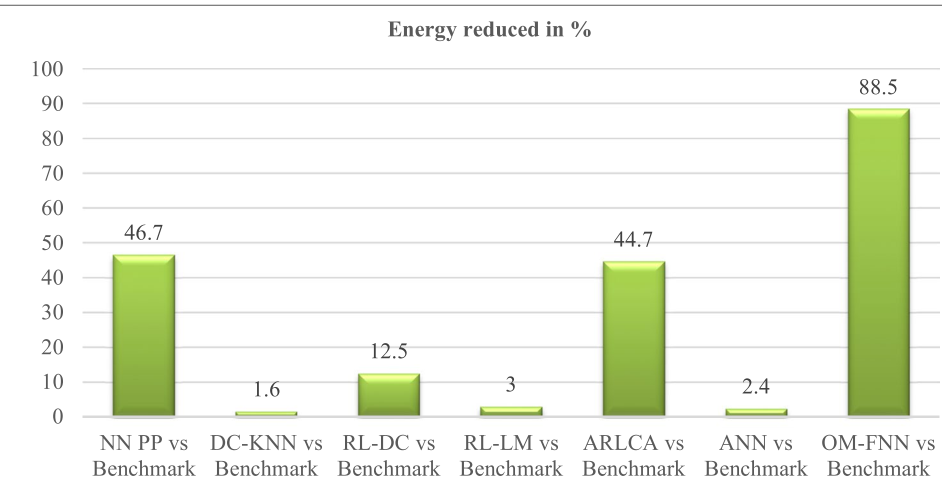

Fig. 12 Energy reduction in Machine Learning techniques vs Benchmark

usage to decrease SLAV and energy utilization in cloud data centers. Global Manager uses first fit decreasing, Cluster Manager uses max load VMP, and Local Manager uses the ARIMA model in their model. As a result, by utilizing ARIMA upper-bound prediction, it is possible to obtain a 90% reduction in migration and SLA violation rates and a 5.4% increase in energy savings. The CloudSim simulator was used to evaluate the method's efficiency with recently proposed approaches that employed the same workload and experimental settings.

Ruan et al. (2015) [85] define performance-to-power ratio (PPR) as conscious virtual machine distribution in energy-efficient clouds. They describe "PPRGear," a novel VM allocation mechanism that takes advantage of performance-to-power ratios for diverse types of hosts. PPRGear can ensure that the host devices use the least amount of power possible. Thus, this drastically lowers the energy usage with minimal performance loss. The proposed algorithm outperforms the competition.

Abdelsamea et al. (2017) [86] introduced multiple regression host overload detection (MRHOD) procedures that practice memory, CPU, and bandwidth to detect host overload and save energy significantly. They used a combination of factors to manage VMs while keeping energy consumption and SLAs low. They also created the hybrid local regression host overload detection (HLRHOD) method based on LR with hybrid variables. This algorithm outperforms single-factor methods.

Khoshkholghi et al. (2017) [87] by developing a method for overloaded host detection using iterative weighted linear regression (IWLR), which takes SLA constraints for data centers into consideration, researchers forecasted a dynamic, cost-effective, and energy-efficient management of virtual machines.

Hemavathy et al. (2019) [88] provide a prediction-based thermal aware server consolidation (PTASC) model, an integration technique that considers numeric and local architecture, as well as service level agreement. PTASC uses a statistical learning approach to consolidate servers (VM Migration). Cloud computing is a method of supplying essential resources by optimizing the usage of data-center resources, which raises energy costs. To reduce energy costs and enhance usage, new energy-efficient methods are proposed that reduce the overall energy consumption of computing and storage.

Lianpeng et al. (2019) [89] Based on the suggested robust simple linear regression (RobustSLR) prediction model, the authors developed a host overloading/underloading detection technique and a novel VM placement strategy for SLA-aware and energy-efficient virtual machine consolidation in cloud data centers. Unlike native linear regression, the proposed approaches update the forecast and slant toward over-prediction by including the error using eight ways of calculating the error. Researchers examined suggested techniques for the test by extending the CloudSim simulator with PlanetLab and random workload. The experimental findings demonstrate that the suggested approach can minimize SLA violation rates up to 99.16% and energy usage up to 25.43%.

Xialin Liu et al. (2020) [50] proposed dynamic consolidation using migration thrashing (MT), which prioritizes VMs with high dimensions, significantly decreasing MT. The degree of migrations required maintaining service level agreements (SLAs) by keeping VMs prone to relocation thrashing on the identical physical servers rather than migrating. Their method improves the relocation thrashing measured around 28%, the number of movements measured around 21%, and the SLAV measured around 19%. When the server is overloaded, their solution detects VMs with sufficient capacity by restricting the transfer of VMs with excessive capacity. The suggested techniques were proven to work by simulating large-scale research setting with a workload data set from many PlanetLab VMs.

Maryam C.-Samani et al. (2020) [90] suggested predictive consolidation of virtual machines (PCVM) using the ARIMA approach, which focuses on the DCVM over the fewest number of real servers. It also reduces the number of unnecessary migrations, detects PM overloading, and enforces SLAs using the ARIMA prediction model. Furthermore, the DVFS approach is utilized to determine the best frequency for heterogeneous physical devices. The experimental findings reveal that, the given framework greatly reduces energy usage while improving QoS characteristics as compared to various baseline techniques. The suggested solution was simulated using MATLAB and CloudSim with real-world PlanetLab workloads.

A brief description of the above detailed literature review and algorithms developed using statistical methods with different workload data is given in Table 7. Table 8, summarises the work, methods, and comparison with their benchmark algorithm to evaluate energy consumption by different researchers. Figure 13 depicts the percentage difference in energy reduction or energy savings in graphical form. The implementation of these algorithms has been tested with different settings. The authors have already discussed the specification of the host, characteristics of the virtual machine, workload datasets, simulator environment, and other criteria for comparing the proposed method to their benchmark algorithm.

Most of the above researchers have used PlanetLab workload traces, as shown in Table 9, or Bitbrains

Table 7 Evaluation of Statistical Methods for Cloud Data Center Resource Management

Author/ Year	Algorithm/Method	Data set/ Workload	Tools/ Experiment Environment	Objective	Performance Metrics/ Pros	Limitations
Z. Cao et al. [82], 2012	EV_MCE	PlanetLab	CloudSim	Host Overload Detection, VM Selection, DVMC, IaaS	Energy Consumption, QoS, SLAV, VM Migrations, SLATAH	Little worse than previous work for EC, simulation, required real infrastructure
F. Farahnakian et al. [83], 2013	LiRCUP	PlanetLab	CloudSim	Detection of Overloaded and Underloaded PM, Utilization Prediction	Energy Consumption, SLA Violation, power cost, CPU usage prediction	Simulation, Prediction utilization is approximated as a function
A. Nadjar et al. [84], 2015	ARIMA MSV_ML	PlanetLab	CloudSim	Dynamic Consolidation of VM, IaaS	SLA, Energy Savings, Migration count, performance, predicting resource usage, number of active hosts	Required real infrastructure, less accuracy in predicting VM and host resource requirements in the near future
X. Ruan et al. [85], 2015	PPRGear	PlanetLab	CloudSim	VM Allocation, VM Migration	Host Utilization, Energy Consumption, SLA, Shutdown times, migration times	Performance degradation, required primitive characteristics of host computers, heavy workload
A. Abdelsamea et al. [86], 2017	MRHOD HLRHOD	PlanetLab	CloudSim	PM Overload and Underload Detection, VMC, Better predictions of host overloading	Migration of VM, Energy Consumption, SLA Violation	Complex, required real cloud, less host utilization
M A Khoshkholghi et al. [87], 2017	PCM (IWLR,V-VMS,BRB,MRUHD)	PlanetLab	CloudSim	Utilization Prediction, VM Consolidation, workload-independent	Energy Utilization, VM Relocation, SLAV, heterogeneous physical servers, IaaS environment	N/W topology required, performance aware strategy, only CPU usage considered.
Hemavathy et al. [88] (2019)	PTASC model, Extended Multiple Linear Regression (EMLR)	Own data	CloudSim	workload prediction, load balancing, server consolidation, VM scheduling, VMP, resource provisioning,	energy efficiency, cost reduction, Response time	Required numeric and local architecture into consideration, statistical learning method, intrusion detection, and prevention systems needed.
Lianpeng et al. [89] (2019)	RobustSLR	PlanetLab and random workload	CloudSim simulator	VM consolidation, host overload, underload detection	Energy consumption, SLAV, SLATAH, PDM, Average SLAV, VM migration, host shutdown	Required RAM and N/W usage to improve energy efficiency and SLAV
X. Liu et al. [50], 2020	DCMMT	PlanetLab	CloudSim	VM Migration Thrashing, Dynamic Consolidation	VM Migration, SLA Violation, Thrashing Index, SLATAH, PDM	Real-world cloud platform needed, required workload statistical properties
Maryam C.-Samani et al. [90] (2020)	PCVM.ARIMA	PlanetLab	CloudSim	DCVM, Resource utilization, VM Placement	Energy Consumption, Migration of VM, SLA Violation, hosts shutdown	Memory and disk utilization are required for future prediction, required reduction in CO_2 emission, failure tolerance, and security.

Table 8 Comparison of benchmark concerning Energy Consumption for Table 7

S.No.	Reference	Year	Algorithm/Method	Benchmark Algo	Energy reduced in %
1	[82]	2012	EV_MCE	DVFS	84
2	[83]	2013	LiRCUP	LR	49
3	[84]	2015	MSV_ML	MAD_MMT_2.5	5.4
4	[85]	2015	PPRGear	THR_RS	69.31
5	[86]	2017	MRHOD	LR & LRR	20
6	[87]	2017	PCM (IWLR, V-VMS, BRB, MRUHD)	LR_RS/LR_MC	28
7	[89]	2019	RobustSLR	Threshold-based heuristics	25.43

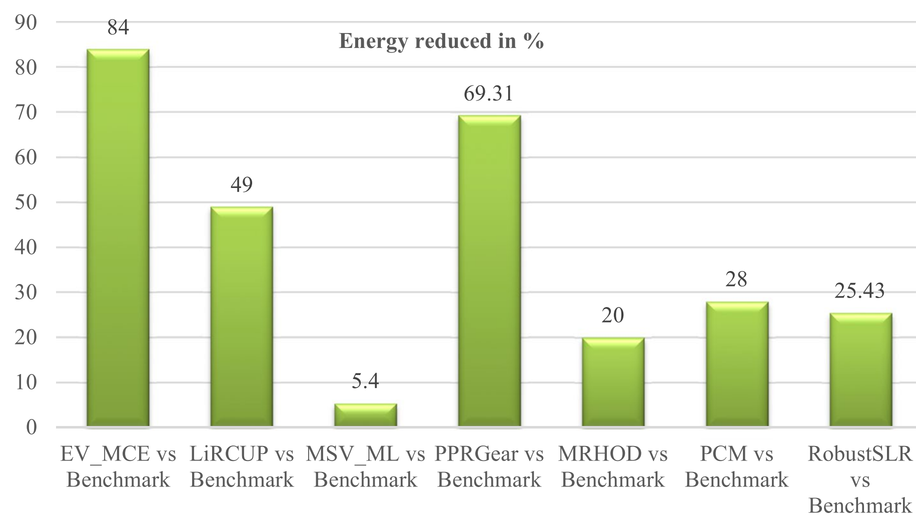

Fig. 13 Energy reduction in Statistical techniques vs Benchmark

Table 9 PlanetLab Workload traces with statistical features [27]

S.No.	Date	VM number	Mean (%)	SD(%)	Q1(%)	Median(%)	Q3 (%)
1.	03/03/2011	1052	12.31	17.09	2	6	15
2.	06/03/2011	898	11.44	16.83	2	5	13
3.	09/03/2011	1061	10.70	15.57	2	4	13
4.	22/03/2011	1516	9.26	12.78	2	5	12
5.	25/03/2011	1078	10.56	14.14	2	6	14
6.	03/04/2011	1463	12.39	16.55	2	6	17
7.	09/04/2011	1358	11.12	15.09	2	6	15
8.	11/04/2011	1233	11.56	15.07	2	6	16
9.	12/04/2011	1054	11.54	15.15	2	6	16
10.	20/04/2011	1033	10.43	15.21	2	4	12

workload traces, as shown in Table 10 for simulation in CloudSim, Matlab, Java, or other environments are given below. Half of the 800 physical nodes in PlanetLab's simulated data center are HP ProLiant ML110G4 systems, while the other half are HP ProLiant ML110G5 systems, as depicted in Table 11. For the smooth conduction of simulation, the power modeling has been configured in CloudSim as shown in Table 11.

Table 10 Bitbrains workload traces have statistical properties [27]

S.No.	Date	VM number	Mean(%)	SD(%)
1.	01/08/2013	1238	11.21	26.33
2.	02/08/2013	1237	7.60	17.52
3.	03/08/2013	1234	5.10	13.16
4.	04/08/2013	1233	8.48	21.11
5.	05/08/2013	1232	9.43	21.67
6.	06/08/2013	1231	8.63	23.19
7.	07/08/2013	1218	7.73	17.49
8.	08/08/2013	1209	10.78	24.07
9.	09/08/2013	1207	7.06	16.93
10.	10/08/2013	1205	8.64	21.62

Result analysis

The result of the review paper work is to find the current research outcomes in energy-efficient resource management as stated in different sections. Table 2, and Fig. 10 represent the saving of energy by up to 90% by different researchers using the heuristic method. The objectives addressed in the evaluation of this method were VM placement, VM allocation, VM migration, and resource utilization. In next section, the authors' metaheuristic approaches were performed to address the objectives of VM consolidation, load balancing, resource management, PM overloading, VM migration, and VM placement. Metaheuristic methods in Table 4 and Fig. 11 showed an improvement in energy savings of up to 95%. Similarly, machine learning algorithms were presented to address the objectives of VM performance, prediction usage of resources, VM scheduling, dynamic consolidation, and resource management. With this approach, the reduction in energy consumption up to 88% has been shown as compared with other methods which has been illustrated in Table 6 and Fig. 12. In the last approach mentioned in this paper, researchers used statistical methods to perform host overload/underload detection, dynamic consolidation of VMs, utilization prediction, and VM allocation. This approach reduces the energy consumption up to 84%, as shown in Table 8 and Fig. 13. The outcomes of the review work are measured in terms of SLA, energy consumption, and the number of migrations against the different numbers of VMs. This review work focuses on energy utilization by different approaches in consolidating virtual machines. The results show that there has been an improvement in energy saving in the outcome of all the researchers by using different techniques. Other research outcomes include the use of integrated and combined approaches for utilization prediction, utilization, virtual machine consolidation, overload detection, VM selection, VM migration, and VM placement.

Major issues, suggestions, and future works

In this paper, the authors have outlined energy-efficient strategies for cloud computing. Several methods have been investigated, and their findings with parameters are listed in the tables. This paper can help people to find out the pros and cons of proposed energy-efficient algorithms that are motivated by researchers.

One of the main issues in cloud computing is using energy effectively, which necessitates the development of an eco-friendly environment. To meet SLAs, service providers must provide continuous power to data centers. This way, the data centers consume a large amount of energy and raise the cost of investment. However, the rising demand for cloud infrastructure has significantly increased the data center energy usage, which has become a crucial concern. As a result, energy-efficient solutions are necessary to reduce this energy utilization. Another significant challenge is the system's reliability degradation because of the high frequency of consolidation and deploying VMs on PMs. Cloud efficiency is the capacity to make greater use of cloud resources at the lowest feasible cost. Other issues that must be addressed include scheduling challenges while PM-VM mapping for each user task, resource utilization prediction accuracy, overload, and underload host detection problems, and adaptive threshold estimation. Moreover, VM selection from the overloaded host, access to a real cloud data center to perform an experiment in a real environment, and improving user satisfaction along with the service providers are also various research challenges.

Most of the researchers have performed simulations in the CloudSim framework in an Infrastructure as a Service (IaaS) environment. In CloudSim, development

Table 11 Watts of electricity usage at several load points[a]

Server	Active idle	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
IBM X3250 M3	42.3	46.7	49.7	55.4	61.8	69.3	76.1	87	96.1	106	113
IBM X3550 M3	58.4	98	109	118	128	140	153	170	189	205	222
HP ProLiant G3	105	112	118	125	131	137	147	153	157	164	169
HP ProLiant G4	86	89.4	92.6	96	99.5	102	106	108	112	114	117
HP ProLiant G5	93.7	97	101	105	110	116	121	125	129	133	135

[a] https:/www.spec.org/power_ssj2008/results/power_ssj2008.html

tools, middleware technology, database management, resource computation, etc. help create and control cloud applications. Logical architecture is based on local and global managers. Cloud architecture is the organization of various components, including applications, databases, on-demand resources, storage, middleware, network devices, and software capabilities to provide services. Increased power use is a longstanding problem in today's computer environment. The rise of applications using complex data has resulted in the construction of large data centers, which has increased the need for energy. According to the above analysis of energy-efficient strategies, the majority of the effort to minimize energy utilization in data centers is done by utilizing dynamic VM consolidation and resource management methods. Some researchers suggest multi-objective [91] algorithms that primarily address SLA, QoS, and resource usage while consuming less energy in cloud data centers. There has been little work done on heterogeneous physical devices, which requires considerable attention from the scientific community. Some major issues in current energy management techniques are prediction utilization of different resources [64]; mapping of VMs to PMs; host overload issues; VM selection from overloaded hosts; access to a real cloud data center; and VM placement. As VM placement is an NP-hard problem, metaheuristic approaches are the best suitable technique, which increases the complexity.

This research contributes significantly to provide important information related to the reduction of data center energy consumption, financial expenses, and the provision of QoS, hence assisting in the development of a strong, competitive cloud computing sector. This is especially crucial in the current green environment, where customers are becoming more environmentally concerned. Furthermore, according to recent research, data centers are a huge and rapidly rising energy-consumption sector of the economy, as well as a substantial source of CO_2 emissions. Also, the research done by [92] the use of blockchain technology and cloud solutions facilitates and improves not only the aggregation of data and secured access to it, but also has a huge impact on the reduction of CO2 emissions and reduces the carbon footprint. Hence, reducing greenhouse gas (GHG) emissions is an important energy policy objective for many nations, as well as achieving the United Nations Sustainable Development Goal (SDG) to transform the world by 2030. As a result, global research efforts should focus on the open problems described in this work to improve energy-efficient resource management approaches in cloud computing systems. Also, the researchers' plan should be centred on reducing energy use and increasing resource use without hurting the performance of the system.

Summary and conclusion

Data centers consume a tremendous amount of electricity for computing user applications as well as cooling their equipment. Improving energy efficiency in data centers may reduce greenhouse gas (GHG) emissions, air pollution, and the amount of water utilized in power generation. So, minimizing energy consumption has been a key challenge in recent years. As a result, it is one of the key study areas in cloud computing. Many researchers are concentrating their efforts on lowering the energy usage of data center infrastructures. This review article looks at virtual and physical machine consolidation strategies using various methodologies to save energy. These strategies look at global energy conservation and resource management. As a result, resource usage increases, and data center energy consumption decreases. This paper aims to identify energy consumption research that has been conducted using various heuristics, metaheuristics, machine learning, and statistical methods. VM selection and migration, host CPU usage prediction, overload detection, and VM placement have been used to manage resources and efficient use energy. The energy savings achieved through various strategies are compared in this paper. Various researchers tested several strategies in cloud data centers to reduce energy consumption and SLAV. In the heuristic approach, researchers have saved from 5.4% to 90% of energy with their proposed method when compared with the existing methods. Similarly, the metaheuristic approach reduces energy consumption from 7.68% to 97%. The machine learning method and the statistical method save energy from 1.6% to 88.5%, and 5.4% to 84% respectively when compared to the benchmark approaches considering a variety of settings and parameters. So, energy saving can be maximized up to 90% using different approaches in respect of consolidation of VMs, prediction of workload traces, utilization of resources, host underload/overload detection, VM selection, VM migration, and VM placement. The results of this study could help researchers come up with new ideas for research that will add to their knowledge and make it easier to use energy efficiently in cloud computing. So, the overall outcome of this review paper is to understand different techniques of energy-saving applied in the cloud data centers. As the field of cloud computing is increasing day by day and its application area is increasing, focus must be on the different methods of energy consumption in cloud data centers.

Acknowledgments
The authors would like to thank the editor and reviewers for their suggestions and improvement to the manuscript.

Authors' contributions
Throughout all of these stages of the study, the writers contributed equally. The final manuscript was reviewed and approved by all authors. Suraj Singh Panwar and MMS Rauthan contributed to the idea and design, data analysis, and interpretation; Suraj Singh Panwar and Varun Barthwal authored the paper, conducted a literature review, and critically edited it for key intellectual content, and confirmed the final version.

Funding
Not Applicable.
The authors are not affiliated with any entity that has a direct or indirect financial interest in the subject matter covered in the paper.

Declarations

Competing interests
This manuscript has not been submitted to or is currently being reviewed by any other journal or publication platform.

Received: 3 March 2022 Accepted: 12 November 2022
Published online: 17 December 2022

References

1. Mell P, Grance T (2011) The NIST definition of cloud computing
2. Wyld DC (2009) Moving to the cloud: an introduction to cloud computing in government, IBM center for the business of government
3. Sethi N (2019) The cloud environment and its basics: a review. Int J Comput Technol 6(1):82–88
4. Sotomayor B, Montero RS, Llorente IM, Foster I (2009) Virtual infrastructure management in private and hybrid clouds. IEEE Internet Comput 13(5):14–22
5. Kaur T, Chana I (2018) GreenSched: an intelligent energy-aware scheduling for deadline-and-budget constrained cloud tasks, simulation modelling practice, and theory. 82:55–83. ISSN 1569-190X. https://doi.org/10.1016/j.simpat.2017.11.008
6. Albers S (2010) Energy-efficient algorithms. Commun ACM 53:86–96. https://doi.org/10.1145/1735223.1735245
7. Çağlar İ, Altılar DT (2022) Look-ahead energy-efficient VM allocation approach for data centers. J Cloud Comput 11(11). https://doi.org/10.1186/s13677-022-00281-x
8. Conti J, Holtberg P, Diefenderfer J, LaRose A, Turnure JT, Westfall L International Energy Outlook 2016 With projections to 2040, United States. https://doi.org/10.2172/1296780
9. Milani AS, Navimipour NJ (2016) Load balancing mechanisms and techniques in the cloud environments: systematic literature review and future trends. J Netw Comput Appl 71:86–98. ISSN 1084–8045. https://doi.org/10.1016/j.jnca.2016.06.003
10. Malik N, Sardaraz M, Tahir M, Shah B, Ali G, Moreira F (2021) Energy-efficient load balancing algorithm for workflow scheduling in cloud data centers using queuing and thresholds. Appl Sci 11(13):5849. https://doi.org/10.3390/app11135849
11. Singh S, Chana I (2016) Cloud resource provisioning: survey, status, and future research directions, knowledge and information system. 49:1005–1069. https://doi.org/10.1007/s10115-016-0922-3
12. Mohit Kumar SC, Sharma AG, Singh SP (2019) A comprehensive survey for scheduling techniques in cloud computing. J Netw Comput Appl 143:1–33. ISSN 1084–8045. https://doi.org/10.1016/j.jnca.2019.06.006
13. Singh S, Chana I (2016) A survey on resource scheduling in cloud computing: issues and challenges. J Grid Computing 14:217–264. https://doi.org/10.1007/s10723-015-9359-2
14. Chaurasia N, Kumar M, Chaudhry R et al (2021) Comprehensive survey on energy-aware server consolidation techniques in cloud computing. J Supercomput 77:11682–11737. https://doi.org/10.1007/s11227-021-03760-1
15. Bui DM, Tu NA, Huh EN (2021) Energy efficiency in cloud computing based on mixture power spectral density prediction. J Supercomput 77:2998–3023. https://doi.org/10.1007/s11227-020-03380-1
16. Bobroff N, Kochut A, Beaty K (2007) Dynamic placement of virtual machines for managing SLA violations, 10th IFIP/IEEE international symposium on integrated network management
17. Varrette S, Guzek M, Plugaru V, Besseron X, Bouvry P (2013) HPC performance and energy-efficiency of Xen, KVM and VMWare hypervisors, 25th international symposium on computer architecture and high-performance computing
18. Gelenbe E (2009) Steps toward self-aware networks. Commun ACM 52(7):66–75
19. Berl A, Gelenbe E, Girolama M, Giuliani G, Meer H, Dang MQ, Pentikousis K (2010) Energy-efficient cloud computing. Comput J 53(7):1045–1051
20. Luo, L., et al., (2012) A resource scheduling algorithm of cloud computing based on energy efficient optimization methods
21. Buyya R, Broberg J, Goscinski AM (2010) Cloud computing: principles and paradigms, Vol. 87. John Wiley & Sons
22. Zhang Q, Cheng L, Boutaba R (2010) Cloud computing: state-of-the-art and research challenges. J Internet Serv Appl 1:7–18. https://doi.org/10.1007/s13174-010-0007-6
23. Buyya R, Beloglazov A, Abawajy J (2010) Distributed, parallel, and cluster computing (cs. DC), energy-efficient management of data center resources for cloud computing: a vision, architectural elements, and open challenges. arXiv. https://doi.org/10.48550/arXiv.1006.0308
24. Buyya R, Ranjan R, Calheiros RN (2009) Modeling and simulation of scalable cloud computing environments and the CloudSim toolkit: challenges and opportunities, in 2009 international conference on high-performance computing & simulation
25. Calheiros RN et al (2011) CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. Software Pract Exper 41(1):23–50
26. Park K, Pai VS (2006) CoMon: a mostly-scalable monitoring system for PlanetLab. ACM SIGOPS Oper Syst Rev 40(1):65–74
27. Moges FF, Abebe SL (2019) Energy-aware VM placement algorithms for the OpenStack neat consolidation framework. J Cloud Comput 8(1):1–14
28. Shen S, Van Beek V, Iosup A (2015) Statistical characterization of business-critical workloads hosted in cloud data centers. In: Proc - 2015 15th IEEE/ACM international symposium on cluster, cloud and grid computing 2015, pp 465–474. https://doi.org/10.1109/CCGrid.2015.60 arXiv:1302.5679v1
29. Anoep S, Dumitrescu C, Epema D, Iosup A, Jan M, Li H, Wolters L. The grid workloads archive: bitbrains. http://gwa.ewi.tudelft.nl/datasets/gwa-t12-bitbrains
30. Amvrosiadis G, Park JW, Ganger GR, Gibson GA, Baseman E, DeBardeleben N (2018) On the diversity of cluster workloads and its impact on research results. In: Proceedings of USENIX ATC
31. Reiss C, Tumanov A, Ganger GR, Katz RH, Kozuch MA (2012) Heterogeneity and dynamicity of clouds at scale: google trace analysis. In: Proceedings of ACM SoCC, pp 1–13
32. Liu Q, Zhibin Y (2018) The elasticity and plasticity in semicontainerized co-locating cloud workload: a view from Alibaba trace. In: Proceedings of ACM SoCC, pp 347–360
33. Shahrad M, Fonseca R, Goiri Í, Chaudhry G, Batum P, Cooke J, Laureano E, Tresness C, Russinovich M, Bianchini R (2020) Serverless in the wild: characterizing and optimizing the serverless workload at a large cloud provider. In: Proceedings of USENIX ATC, pp 205–218
34. Luo S, Xu H, Lu C, Ye K, Xu G, Zhang L, Yu D, He J, Xu C (2021) Characterizing microservice dependency and performance: Alibaba trace analysis, SoCC'21
35. Buyya R, Beloglazov A, Abawajy J (2010) Energy-efficient management of data center resources for cloud computing: a vision, architectural elements, and open challenges
36. Katal A, Dahiya S, Choudhury T (2022) Energy efficiency in cloud computing data centers: a survey on software technologies. Clust Comput. https://doi.org/10.1007/s10586-022-03713-0
37. Zhang Y, Liu J (2022) Prediction of overall energy consumption of data centers in different locations, Sensors. 22:3704. https://doi.org/10.3390/s22103704
38. Teng F et al (2017) Energy efficiency of VM consolidation in IaaS clouds. J Supercomput 73(2):782–809
39. Zhou Z et al (2018) Minimizing SLA violation and power consumption in cloud data centers using adaptive energy-aware algorithms. Future Gene Comput Syst 86:836–850

40. Khosravi A (2017) Energy, and carbon-efficient resource management in geographically distributed cloud data centers
41. Kaur T, Chana I (2015) Energy efficiency techniques in cloud computing: a survey and taxonomy. ACM Comput Surv 48(2):1–46
42. Clark C, Fraser K, SH, J., & Hansen, E. (2005) Warfield, live migration of virtual machines. In: Jul C, Limpach I, Pratt A (eds) In proceedings of the 2nd ACM/USENIX symposium on networked systems design and implementation (NSDI), pp 273–286
43. Beloglazov A, Buyya R (2012) Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers, concurrency and computation. Pract Exper 24(13):1397–1420
44. Srikantaiah S, Kansal A, Zhao F (2008) Energy-aware consolidation for cloud computing. Clust Comput 12:1–15
45. Beloglazov A, Buyya R (2010) Energy-efficient resource management in virtualized cloud data centers,10th IEEE/ACM international conference on cluster, cloud and grid computing
46. Beloglazov A, Abawajy J, Buyya R (2012) Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing. Future Gene Comput Syst 28(5):755–768
47. Ghobaei-Arani M et al (2018) A learning-based approach for virtual machine placement in cloud data centers. Int J Commun Syst 31(8):1–18
48. Wang H, Tianfield H (2018) Energy-aware dynamic virtual machine consolidation for cloud datacenters. IEEE Access 6:15259–15273
49. Bhattacherjee S et al (2020) Energy-efficient migration techniques for cloud environment: a step toward green computing. J Supercomput 76(7):5192–5220
50. Liu X et al (2020) Virtual machine consolidation with minimization of migration thrashing for cloud data centers. Math Probl Eng 2020:1–13
51. Jangiti S, Shankar Sriram VS (2020) EMC2: energy-efficient and multi-resource- fairness virtual machine consolidation in cloud data centers. Sustain Comput: Inform Syst 27:100414. ISSN 2210–5379. https://doi.org/10.1016/j.suscom.2020.100414
52. Garg V, Jindal B (2021) Energy-efficient virtual machine migration approach with SLA conservation in cloud computing. J Cent South Univ 28(3):760–770
53. Alharbi F et al (2021) Simultaneous application assignment and virtual machine placement via ant colony optimization for energy-efficient enterprise data centers. Clust Comput 24:1255–1275
54. Kaur T, Kumar A (2022) Power aware energy efficient virtual machine migration (PAEEVMM) in cloud computing. In: Satyanarayana C, Samanta D, Gao XZ, Kapoor RK (eds) High performance computing and networking. Springer, Singapore. https://doi.org/10.1007/978-981-16-9885-9_46 Lecture notes in electrical engineering, vol 853
55. Kousiouris G, Cucinotta T, Varvarigou T (2011) The effects of scheduling, workload type, and consolidation scenarios on virtual machine performance and their prediction through optimized artificial neural networks. J Syst Softw 84(8):1270–1291
56. Aryania A, Aghdasi HS, Khanli LM (2018) Energy-aware virtual machine consolidation algorithm based on ant colony system. J Grid Comput 16(3):477–491
57. Goyal et al (2019) An optimized model for energy efficiency on cloud system using PSO & CUCKOO search algorithm. Int J Innov Technol Explor Eng 8(9S):2278–3075
58. Tarahomi M, Izadi M, Ghobaei-Arani M (2020) An efficient power-aware VM allocation mechanism in cloud data centers: a micro genetic-based approach. Clust Comput 24(2):919–934
59. Dubey et al., (2020), A simulated annealing based energy-efficient vm placement policy in cloud computing, 2020 international conference on emerging trends in information technology and engineering (ic-ETITE)
60. Barthwal V, Rauthan MMS (2021) AntPu: a meta-heuristic approach for energy-efficient and SLA aware management of virtual machines in cloud computing. Memetic Computing 13(1):91–110
61. Mirmohseni SM, Javadpour A, Tang C (2021) LBPSGORA: create load balancing with particle swarm genetic optimization algorithm to improve resource allocation and energy consumption in clouds networks. Math Probl Eng 2021:1–15.
62. Salami et al., (2021). An energy-efficient cuckoo search algorithm for virtual machine placement in cloud computing data centers. The Journal of Supercomputing 77(11):13330–13357.
63. Sayadnavard MH, Haghighat AT, Rahmani AM (2022) A multi-objective approach for energy-efficient and reliable dynamic VM consolidation in cloud data centers, engineering science and technology. Int J 26:100995. ISSN 2215–0986. https://doi.org/10.1016/j.jestch.2021.04.014
64. Malik S, Tahir M, Sardaraz M, Alourani A (2022) A resource utilization prediction model for cloud data centers using evolutionary algorithms and machine learning techniques. Appl Sci 12(4):2160. https://doi.org/10.3390/app12042160
65. Rao J et al (2009) Vconf: a reinforcement learning approach to virtual machines auto-configuration. In: Proceedings of the 6th international conference on Autonomic computing
66. Duy TVT, Sato Y, Inoguchi Y (2010) Performance evaluation of a green scheduling algorithm for energy savings in cloud computing, IEEE international symposium on parallel & distributed processing, workshops and Ph.D. forum (IPDPSW)
67. Niehorster, O., et al., (2011), Autonomic resource management with support vector machines, IEEE/ACM 12th international conference on grid computing
68. Islam S et al (2012) Empirical prediction models for adaptive resource provisioning in the cloud. Future Gene Comput Syst 28(1):155–162
69. Xu C-Z, Rao J, Bu X (2012) URL: a unified reinforcement learning approach for autonomic cloud management. J Parallel Distrib Comput 72(2):95–105
70. Farahnakian F et al (2013) Energy-aware consolidation algorithm based on k-nearest neighbor regression for cloud data centers, Department of IT, University of Turku. IEEE/ACM 6th International Conference on Utility and Cloud Computing, Finland
71. Farahnakian F, Liljeberg P, Plosila J (2014) Energy-efficient virtual machines consolidation in cloud data centers using reinforcement learning. In: 22nd Euromicro International Conference on Parallel, Distributed, and Network-Based Processing, pp 500–507
72. Patel M, Chaudhary S, Garg S (2016) Machine learning-based statistical prediction model for improving the performance of live virtual machine migration. J Eng 2016:1–9
73. Duggan M et al (2017) A reinforcement learning approach for the scheduling of live migration from underutilized hosts. Memetic Computing 9(4):283–293
74. Duggan M et al (2017) Predicting host CPU utilization in cloud computing using recurrent neural networks. In: In 2017 12th international conference for internet technology and secured transactions (ICITST)
75. Zia Ullah Q, Hassan S, Khan GM (2017) Adaptive resource utilization prediction system for infrastructure as a service cloud. Comput Intell Neurosci 2017:1–12
76. Shaw R, Howley E, Barrett E (2017) An advanced reinforcement learning approach for energy-aware virtual machine consolidation in cloud data centers. In: 12th international conference for internet technology and secured transactions (ICITST)
77. Sotiriadis S, Bessis N, Buyya R (2018) Self-managed virtual machine scheduling in cloud systems. Inf Sci 433:381–400
78. Mason K et al (2018) Predicting host CPU utilization in the cloud using evolutionary neural networks. Future Gene Comput Syst 86:162–173
79. Patel D, Gupta RK, Pateriya R (2019) Energy-aware prediction-based load balancing approach with VM migration for the cloud environment. In: Data, engineering and applications. Springer, pp 59–74
80. Kumar J, Saxena D, Singh AK, Mohan A (2020) Biphase adaptive learning-based neural network model for cloud datacenter workload forecasting. Soft Comput 24(19):14593–14610. https://doi.org/10.1007/s00500-020-04808-9
81. Saxena D, Singh AK (2021) A proactive autoscaling and energy-efficient VM allocation framework using online multi-resource neural network for cloud data center. Neurocomputing 426:248–264. ISSN 0925–2312. https://doi.org/10.1016/j.neucom.2020.08.076
82. Cao Z, Dong S (2012) Dynamic VM consolidation for energy-aware and SLA violation reduction in cloud computing. In: IEEE 13th international conference on parallel and distributed computing, applications and Technologies
83. Farahnakian F, Liljeberg P, Plosila J (2013) LiRCUP: linear regression-based CPU usage prediction algorithm for live migration of virtual machines in data centers. In: 39th Euromicro conference on software engineering and advanced applications (SEAA), pp 358–364

84. Nadjar A, Abrishami S, Deldari H (2015) Hierarchical VM scheduling to improve energy and performance efficiency in IaaS Cloud data centers. In: 5th international conference on computer and knowledge engineering (iccke)
85. Ruan X, Chen H (2015) Performance-to-power ratio aware virtual machine (VM) allocation in energy-efficient clouds. In: IEEE international conference on cluster computing, pp 264–273
86. Abdelsamea A et al (2017) Virtual machine consolidation enhancement using hybrid regression algorithms. Egypt Inform J 18(3):161–170
87. Khoshkholghi MA et al (2017) Energy-efficient algorithms for dynamic virtual machine consolidation in cloud data centers. IEEE Access 5:10709–10722
88. Hemavathy M, Anitha R (2019) Green aware based VM-placement in cloud computing environment using extended multiple linear regression model. In: Emerging trends in computing and expert technology, COMET 2019. Springer, Cham. https://doi.org/10.1007/978-3-030-32150-5_53 Lecture notes on data engineering and communications technologies, vol 35
89. Li L, Dong J, Zuo D, Wu J (2019) SLA-aware and energy-efficient VM consolidation in cloud data centers using robust linear regression prediction model. IEEE Access 7:9490–9500
90. Chehelgerdi-Samani M, Safi-Esfahani F (2020) PCVM.ARIMA: predictive consolidation of virtual machines applying the ARIMA method. J Supercomput 77:2172–2206. https://doi.org/10.1007/s11227-020-03354-3
91. Chen J, Du T, Xiao G (2021) A multi-objective optimization for resource allocation of emergent demands in cloud computing. J Cloud Computing 10(20):1–17
92. Karaszewski R, Modrzyński P, Modrzyńska J (2021) The use of Blockchain Technology in Public Sector Entities Management: an example of security and energy efficiency in cloud computing data processing. Energies 14(7):1873. https://doi.org/10.3390/en14071873

Publisher's Note

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