Soft Computing
https://doi.org/10.1007/s00500-023-08281-y

FOCUS

Improved quality of service with fuzzy-based optimised resource allocation with energy consumption-based IoT network application

VijayaKumar Chandarapu[1] · Madhavi Kasa[1]

Accepted: 19 April 2023

Abstract
Cloud computing provides web-based services for accessing, manipulating, and modifying data. Cloud computing is unquestionably a breakthrough because it relies on physical infrastructure that is far more expensive than cloud computing itself. End users can take advantage of various Cloud computing features, including self-service, limitless resources, quick elasticity, and measurable services. These resources are available to anybody with an Internet connection. These resources and services are available to users on a pay-per-use basis. Load balancing is a critical concern in cloud computing design. It is possible to use a variety of techniques to allocate resources in cloud computing. Server efficiency necessitates the development of algorithms and strategies that reduce energy usage and allocate data efficiently with minimum load on the server groups. Unlike traditional computing, cloud computing offers a number of advantages. As a result, it provides utility-based services to its customers on demand. Aside from providing IT services to its consumers, this computer environment charges for each use. A growing number of jobs necessitate virtual machines in order to be completed in a timely manner. Cloud computing's huge user growth has made load balancing a top priority. This research concentrates on load balancing, which distributes the load among numerous servers with minimum energy consumption to meet the increasing demands of users. In this research, a method of allocating resources with optimised technique to a cloudlet while simultaneously taking into account the estimated power consumption while servicing the cloudlet is considered. The cloud user has the option of participating in energy-efficient resource allocation if they choose to match energy consumption as a consideration when selecting a VM in the suggested technique. The selection criteria include not just the execution or completion time but also an estimate of the data centres power consumption. This paper proposes an efficient normalised load balancing optimised time triggered resource allocation (NLB-OTT-RA) model with minimum energy consumption in the cloud environment to improve the performance. The proposed model is compared with the traditional models, and the results show that the proposed model performance is better.

Keywords Resource allocation · Power consumption · Quality of service · Data centres · Load balancing · Virtual Machine · Time Triggered Resource Scheduling

1 Introduction

Because of its dependability, cost-effectiveness, and scalability, cloud computing is a popular choice among businesses and individuals alike for delivering secure and trustworthy services. Cloud computing is the most popular method of cutting computing costs for end users, especially in the IT industry. IT processing, construction, and management are all influenced by this approach, as are the methods and technologies used. There are four main components to cloud computing: virtual machines (VMs), physical machines (PMs), data processing with resource

✉ Madhavi Kasa
Kasamadhavi1@yahoo.com

VijayaKumar Chandarapu
vijay.chandarapu@gmail.com

[1] Department of Computer Science and Engineering, Jawaharlal Nehru Technological University College of Engineering, Jawaharlal Nehru Technological University, Ananthapuramu, Andhra Pradesh 515002, India

Published online: 09 June 2023

allocators, and users themselves. The cloud delivers the services and resources to the users based on the service level agreement (SLA) to accomplish the user's work, which includes control over resources such as CPU, memory, and physical memory. VMs in the cloud are typically allotted to PMs in order to provide services to customers. In order to increase resource efficiency and decrease costs, VMs require an accurate prediction of the workloads and quality of service (QoS) requirements. The vast amounts of power needed to run the data centres that house the cloud services have an adverse effect on the environment because of the high levels of carbon dioxide emissions. Reducing cloud data centre energy usage is therefore necessary to boost CSP profits, lower user costs, and lessen the environment's impact on CSP.

In order to improve cloud energy conservation, it is critical to look into the power flow in traditional data centres and understand how power is distributed. The power consumption of cooling appliances is large, but it is proportional to the power consumption of IT equipment. An innovative technology for reducing cooling power consumption is the use of free cooling by large corporations. Rather than relying on conventional refrigeration, these solutions use naturally cold air or water to cool data centres. A reduction in power consumption has been achieved as a result. Zero-refrigeration systems, which are feasible in many countries, can reduce power consumption up to 100 per cent. Before beginning any work on power and energy modelling or assessment, it is critical that users have a firm grasp of the relationship between the two. The paucity of power measuring equipment in modern data centres means that models that forecast power requirements, as well as VM migration maintenance costs, are becoming increasingly popular for power monitoring. In order to effectively organise and schedule virtual machines in a way that reduces data centre energy expenditure, simulations that rely on knowledge such as resource consumption or information provided by service management are helpful.

Resource and service distributions must be carried out in a methodical manner to ensure the same loads are applied to all resources at any given moment and to increase resource efficiency. The system's performance will suffer greatly if there is any sort of load imbalance. While maintaining the load, it is important to consider the energy consumption. The term green cloud computing refers to cloud computing that makes optimal use of resources while also consuming less energy. Both cooling and computational resources contribute to the data centre's high energy usage. Around 70% of total energy usage is accounted for by computing resources, with the remaining 30% being accounted for by cooling infrastructure.

There are two parts to the energy consumption problem: server-side activities and networking communications. One of the most important ideas is to reduce operational costs while maximising resource allocation. In the platform as a service segment, this can be achieved. Load balancers and schedulers are used to balance the resources, predict the load, and reduce energy consumption. Capabilities and infrastructure are either allocated or de-allocated in cloud computing because there is no need for a large up-front expenditure, integrating into the cloud minimises overall operation and maintenance costs. Figure 1 shows a cloud infrastructure. With cloud computing's scalability, customers have the opportunity to scale up or down their computing needs as they see fit. The objective of resource allocation is to distribute resources in such a way that environmental equilibrium is preserved while doing so. Algorithms for scheduling resources are used to keep the system in balance and to maximise performance. The cloud computing infrastructure components are shown in Fig. 1.

Using the network's processing resources to do complex activities that require large-scale computation is the key technology of cloud computing. Many aspects, such as load balancing, make span, and energy usage, must be taken into account while allocating resources. In cloud computing, selecting the best possible resource nodes for a task must be taken into consideration, and they must be chosen based on the work's specific qualities. It is critical that cloud resources be distributed not only to meet user-specified QoS criteria via service level agreements (SLAs) but also to cut down on energy usage.

An application's resources and run-time support are met by virtual machines. Execution of a programme can be made possible in two ways: by generating an instance of the virtual machine required by the programme and scheduling the request for physical resources, otherwise known as resource provisioning. To represent the operating system notion, the VM is used: a software abstraction of a computer's hardware. The underlying physical machine is adequately resembled by a virtual machine, which runs

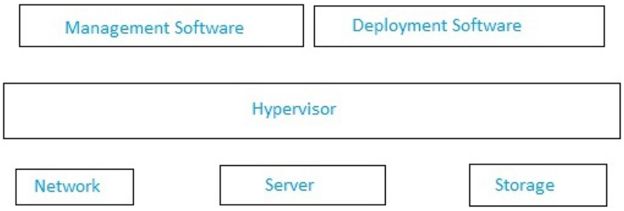

Fig. 1 Cloud computing infrastructure components

existing software without modification. Server consolidation, live migration, and security isolation are just a few of the advantages that VM technology has gained in recent years in data centres and cloud computing settings. In a cloud computing environment, numerous services can be encapsulated within a single virtual machine to satisfy the needs of the user. Server-based VMs enable a multi-OS environment with support for numerous software programmes. A physical machine can host one or more VMs, depending on how many VMs are needed. Using visualisation technologies, cloud computing environments can perform dynamic load balancing between hosts.

Cloud data centres typically use optimisation strategies to reduce energy consumption and SLA breaches. Analysing cloud parameters of mathematical models, optimisation methods provide an efficient energy consumption solution. Resources, energy consumption, and SLA violations are the exclusive emphasis of these optimisation strategies. In order for the models to be effective, they must include a variety of factors, including energy consumption reduction, QoS enhancement, and SLA violations. Energy models and numerous input factors such as VM migration and resource allocation are used in optimisation approaches to determine the best parameter settings for reducing energy usage. Clustering, on the other hand, selects cluster heads for similar tasks and nodes that consume the least amount of energy. The input vector for cloud energy usage is neither changed by the clustering methods, nor is a search process performed. The cloud parameters can be used to train machine learning models, which can then forecast the likelihood of excessive energy usage in future. There are three main types of optimisation methods: search optimisation, clustering models, and machine learning models.

2 Literature survey

The cloud data centre optimisation approaches are the most often used techniques to reduce energy consumption and SLA violations. Analysing cloud parameters of mathematical models, optimisation methods provide an efficient energy consumption solution. Numerous optimisation techniques in resource allocation is analysed by Zhou et al. (44). In the cloud, resource allocation, energy consumption, VM migration, SLA, and QoS are addressed using the firefly algorithm, the whale optimisation algorithm, the heuristic technique, and the cat swarm optimisation. Resources, energy consumption, and SLA violations are the exclusive emphasis of these optimisation strategies. In order for the models to be effective, they must include a variety of factors, including energy consumption reduction, QoS enhancement, and SLA violations.

On the basis of assigning cloud demands to the most suitable cloud server, Zhang et al. (39) suggested an energy-efficient task planning architecture for the cloud. Noted, the process of organising cloud service requests optimally so that the task may be completed in time allotted with low consumption of energy resources is known as green cloud scheduling. In addition, if the task is not well scheduled, it will use more energy. Using specified characteristics, this scheduling system divides arriving cloud service requests into various categories before assigning each request to the cluster that is most suited to the corresponding category. Consolidation of VMs, correct resource allocation, and modifying virtual machine consumption were all investigated by Shojafar et al. (29). Genetic algorithms and machine learning were used to minimise cloud energy usage and to increase the fitness function of the distributed cluster of a distributed server. Using this energy model, the cloud configurations may be evaluated with greater accuracy in the simulator or other models. It does not, however, account for the additional energy expenses of transferring live VMs.

There are two algorithms defined by Liu et al. (18): distributed resource allocation (DRA) and energy saving. Power distribution unit (PDU) connects to the system for monitoring and recording energy consumption. VMs' idle resources can be better utilised by using the DRA algorithm. It also minimises the amount of energy consumed by the cloud cluster by using the energy conservation approach. To be more precise, the reduction amounts to

39.89 per cent of total energy usage. In order to reduce energy usage in a data centre, Al-Dhuraibi et al. (2) designed an architecture for managing VMs. This method solely takes into account the energy consumption of the CPU.

Rahimi et al. (25) devised a strategy for implementing a computer infrastructure for information and communication technology centres (ICTs). Cloud computing, according to recent studies, will become the norm in technological advancements and will be tremendously useful to businesses. Administrative, instructional, research, and teacher education ICT systems and facilities are provided by all educational institutions' ICT units. For the first time, a novel algorithm based on an agent-based automated system structure has been presented by Sotiriadis et al. (31). This algorithm not only looks for integrated solutions but it also recognises and lowers the cost of virtual machines exclusively used by on-demand services. In the classification of cloud computing systems, Yu et al. (37) raised several concerns about power management. Cloud data centres' power consumption was also examined through virtualisation, migrations, and work system architectures. The design and monitoring of matching processing times between data centres and inbound jobs will benefit from the implementation of a new management paradigm.

It was found that the current infrastructure for assigning resources and its potential significance to cloud computing, which is projected to take centre stage on the future Internet, was evaluated by Huanget al. (14). As a researcher, the author is likewise concerned about network consciousness and the constant optimisation of network resource allocation, as well as highlighting issues that the research group should investigate further. A strategy for allocating network resources in cloud technology based on dynamic observations was also proposed.

Based on the application of analytics processing modelling methodologies and randomised evaluation metrics, simulation studies were made by Gai et al. (8) to demonstrate the strategy's efficiency. In order to accommodate both online and batch requests, consideration is given. A cloud data centre planning issue has been identified within the parameters of the service scalability technique by Diallo et al. (6). As a result of using the suggested programming style, less energy is consumed during the execution of jobs. Using this strategy, energy consumption can be minimised significantly more than other strategies, according to the data collected. To help with the estimation problem, Ghanavati et al. (10) conducted in-depth study and notations on a number of quality evaluation measures, including context switching, processing duration, processing periods, and reaction time.

Zhong et al. (40) devised a sleep-state selection algorithm that can cut power usage at run time. The interval lengths are predicted statistically. Using a hybrid approach, Xu et al. (36) found a way to reduce energy consumption in the cloud while also assuring quality of service (QoS). A hybrid VM selection strategy and a low utilisation host policy are employed to reduce energy consumption in this method. In addition to reducing CO2 emissions and energy consumption, this method has a substantial impact on health-related issues. An optimisation approach developed by Ghahramani et al. (9) aims to reduce the energy consumption of data centres that are distributed area wise. Dynamic requests between DCs and users can be handled with an intelligent heuristic algorithm. Both a niche genetic algorithm and a random depth-first search are used to accomplish this. The hybrid technique developed by Zhou et al. (42) optimises energy in a way that has both good migration results and low energy consumption. ACO and a gravitational search algorithm are two of the algorithms incorporated into the suggested system. Gravitational search is a local search method that exploits the law of gravity, in which all agents try to gravitate towards the agent with the most gravitational mass, which is regarded the best agent. ACO aids in solving NP-hard problems and is used for dynamic VM consolidation.

3 Proposed model

In today's world, cloud computing is a critical tool. It is modelled to deliver services like as computation, software, data access, and storage to users without any prior knowledge of the server providing this service. Servers, networking equipment, and cooling systems consume a large amount of energy in large and virtualised data centres in order to provide their clients with efficient and dependable services. For both service providers and end users, high energy usage increases operating expenses. Additionally, a significant amount of carbon dioxide is released into the atmosphere, endangering wildlife and the ecosystem. In the last few years, a number of large-scale studies have been conducted on the energy use of data centres. Physical resources alone are not enough to produce this difficulty; it is how they are utilised.

Managing many VMs on a single server is made possible with virtualisation, which is a very effective high technology. It is in reality possible to dynamically shift VMs from one server to another in real time using migrations and server consolidation techniques. The server will also be shut down if there are no virtual machines on it. These VMs may be able to distribute the load more evenly. To ensure compliance with SLAs and the overall quality of virtual services (QoVS), it can be done without disrupting the service. However, poorly managed VMs can lead to a decrease in performance when the demand for resources is increased. Cloud providers must establish a middle ground in between energy performance of data centres and SLA in order to deliver QoS described by SLA. To reduce energy consumption and maintain a high level of performance in cloud data centre environments, efficient interventions to manage data centre resources are needed.

Dynamic power management policies can only be implemented if a model of dynamic power consumption can be developed. An accurate model must be able to estimate how much power the system will use at any given time based on the system's operational characteristics. Power monitoring capabilities built into modern computer servers can be used to achieve this goal. A server's power consumption can be monitored in real time using this device, and accurate power consumption statistics can be collected. It is possible to derive a power consumption model for a specific system using this information. In order to implement this strategy, statistical data for each target system must be gathered.

A multistage process is involved in allocating resources to physical hosts and optimising the use of resources in the data centre. It is allotted to the physical host that promises to run it with the least amount of energy consumption. In the next step, simulated annealing is used to optimise the power consumption of all the hosts in the data centre, with the goal of reducing it as much as possible. Load balancing methods in cloud computing architectures are used to reduce energy usage while distributing the workload among servers. When it comes to load balancing and task scheduling, the system's primary goal is to help reduce energy usage in a cloud environment by transferring loads from overworked servers to underutilised servers. Despite the fact that many of these systems have been built to accomplish this goal, the amount of energy consumed can be greatly reduced by combining algorithms.

The proposed model considers weather forecasting streaming data that are provided to cloud. The data are provided to the cloud as a set of tasks in which resources are allocated for execution, and load balancing is performed for improved efficiency. To reduce energy usage, this work follows a specific technique for balancing effort and resource allocation. An efficient normalised load balancing optimised time triggered resource allocation (NLB-OTT-RA) model is proposed in this research for minimum energy consumption in the cloud environment to improve the system performance. The algorithm explains the process of load balancing and reducing energy consumption.

Algorithm NLB-OTT-RA

{

Step-1: The cloud service provider registers all the resources that are available and the resources will be allocated to the tasks to complete the execution in time avoiding delay. The resource registration is performed as

$$Res[M] = \sum_{r=1}^{M} getaddr\big(Res(r)\big) + VM\big(datacentre(r)\big) + getTime(T)$$

Here getaddr() is used to extract the resource default address of a resource r and datacentre(r) is the datacentre used to execute tasks after resource allocation. Time T is the time instant during resource registration.

Step-2: The task analysing and allocation of resources based on the task burst time is performed. The resource allocation based on task burst time is performed as

$$Task[i] = \sum_{i=1}^{M} len(\,task(i)) + maxres(res(i)) + \max_{0\le i\le M} r * i^2 \varepsilon\ Res[M]$$

Here, len() is used to identify the length of task for burst time calculation and maxres() is used to get the maximum number of resources for task completion. Maximum resources are allocated to the tasks for avoiding delay in the execution.

Step-3: The energy level calculation of every server group handling tasks are performed and the resource energy consumption levels are calculated as

$$Ener(L) = {}_{i=1}^{N} allocEner\big(Res(i)\big) + \max(VMset\big(task(i)\big) + \frac{len(maxtask(VM(i))}{allovEner\big(Task(i)\big)} + Th$$

Here allocEner() is the energy allocated to utilize a resource and VMset is the VMs used for handling tasks. maxtask()n is used for detecting the tasks handling by VM and Th is the threshold value used for maintain the additional energy level as a Threshold limit.

Step-4: The task allocation to a server group is performed by allocating required number of resources to the server group is performed as

$$TaskScheduler(tasks(i)) = \sum_{i=1}^{M} \delta(\max Util(res(i))) + \sum_{i=1}^{M} VMset(\min(task(i)) + \frac{len(Task(i)) + \max Ener(res(i))}{count(VMset)}$$

Here δ is the similar task detection for allocating the task priority to avoid delay and to avoid collisions in VM allocation. The task allocation is performed by distributing the load equally to the VM servers. The max energy levels maintaining resources and tasks are considered for VM allocation to tasks.

Step-5: The load on the server group is calculated and the load normalization is performed by allocating the additional resources and balancing the tasks with the server groups. The normalization process is performed as

$$\begin{aligned} &Load\big(VMset(i)\big) \\ &\qquad = \sum_{i=1}^{M} \max\Big(VMset\,\big(len\big(task(i), task(i+1)\big)\big)\Big) \\ &\qquad + \frac{\mathrm{minEner(VMset(task(i))}}{maxcap(VMset(i))} + \delta - allocEner(VMset(i)) \end{aligned}$$

$$\begin{aligned} NormTask(i) = \bigcup\nolimits_{i=1}^{M} &\max(Load(task(i), task(i+1) + \min(energylevel\big(VMset(i)\big) \\ &- \delta - \min(task(i))\varepsilon VMset(i) \end{aligned}$$

$$\begin{aligned} &resRealloc\big(VMset(i)\big) \\ &\qquad = \bigcup\nolimits_{i=1}^{M} \max(res\big(task(i), task(i+1)\big) + \min\big(VMset(i)\big) - NormTask(i) \\ &\qquad + \max(Load\big(task(i)\big)) \end{aligned}$$

Step-6: The time triggered load calculation is performed at regular time intervals and the model balances the load with the server groups. The time triggered load calculations is performed as

$$\begin{aligned} TTLoad(i) = \sum_{i=1} &\min\Big(Load\big(task(i), task(i+1)\big)\Big) + getTime(T) \\ &+ \frac{maxlen(VMset(task(i))}{len\big(task\big(VMset(i)\big)\big)} \end{aligned}$$

Step-7: The energy consumption reduction is performed using the sleep and resume mode in which the server group will shift from sleep and resume modes during the utilization of resources. The process is performed as

$$AllocEner\big(task(i)\big) = MAX\_LIMIT$$

$$\begin{aligned} &\boldsymbol{VMset\big(Tasks(i)\big)} \\ &\qquad = \sum_{\boldsymbol{l=1}} \boldsymbol{maxlen\big(task(i)\big) + Res(i) + getVMexec\big(task(i)\big)} \\ &\qquad \boldsymbol{- ener\big(task(i)\big)} \end{aligned}$$

```
If(allocEner(i)<Th)
{
If (exec(task(i))
{
```

```
allocEner=allocEner(task(i))-ener(len(task(i)))

res(i)=maxlen(task(i))-task(VMset(i))

sleep(len(task(i)))

}

else

{

allocEner=allocEner(task(i))+ener(len(task(i)))

res(i)=minlen(task(i+1))+task(VMset(i+1))

sleep(minener(task(i))

}

}
```

4 Results

Resource allocation in cloud computing refers to the process of allocating available computing resources to the various cloud applications. If the distribution of resources is not carefully regulated, services will be starved. That challenge is solved by resource provisioning, which allows service providers to control the resource for each module. There are a number of different strategies for distributing resources to fulfil the needs of cloud applications, and one of these is known as resource allocation strategy. There must be an integrated approach to resource consumption and allocation to maximise resource usage. The proposed model considers weather forecasting streaming data that are provided to cloud. The proposed model is developed using python and implemented in google Colab. The proposed efficient normalised load balancing optimised time triggered resource allocation (NLB-OTT-RA) model is compared with the existing profit-maximised collaborative computation offloading and resource allocation (PMCCO-RA model), and the results are represented.

In cloud systems, resource pooling is provided as a pool of resources that may be dynamically assigned and reassigned to meet the needs of various customers. To support consumer systems, it has the ability to rapidly scale out and drastically release so that it can swiftly scale in automatically. Faster and more cost-effective services are provided through the cloud. However, cloud providers face a huge problem with resource allocation. Overconsumption of resources has necessitated greater resource management. In addition, because demand and capacity fluctuate over time, the resources needed may be more than those available in the cloud. As a result, dynamic resource allocation strategies enable more efficient utilisation of the available resources. The resources are registered at the cloud service provider, and the resource registration accuracy levels of the existing and proposed models are shown in Fig. 2.

The resources are allocated to the tasks to complete their operations without any delay. The tasks whose load is high will be allocated with additional resources to balance the load and to complete the allocated operations. The proposed model resource allocation accuracy levels are contrasted with the existing ones, and the results are shown in Fig. 3.

In a cloud computing environment, load prediction is a crucial for cost-optimal allocation of resources and energy-saving strategy. To improve prediction accuracy, load classification is required prior to prediction. In this research, a new method for predicting the future load of cloud-oriented data centres is presented. The load prediction of existing and proposed models is shown in Fig. 4.

To ensure that no node in a cloud computing system is overcrowded or underutilised, load balancing is indeed the procedure of redistribution of the workload. Load balancing aims to improve reaction time, execution time, and system stability by distributing the workload. The proposed model load balancing is effective and fast. The load balancing time levels of the proposed model and traditional model are shown in Fig. 5.

The rise of cloud computing has resulted in excessive energy use in data processing, storage, and communication. The high carbon emissions of the data centres are unfriendly to the environment due to the obvious massive energy usage. The cloud energy consumption levels of existing and proposed model are depicted in Fig. 6.

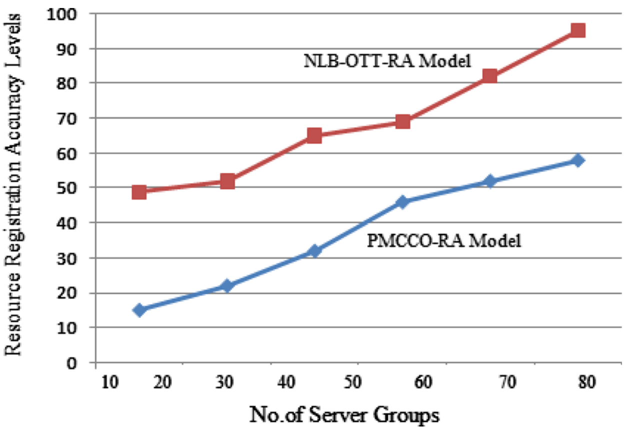

Fig. 2 Resource registration accuracy levels

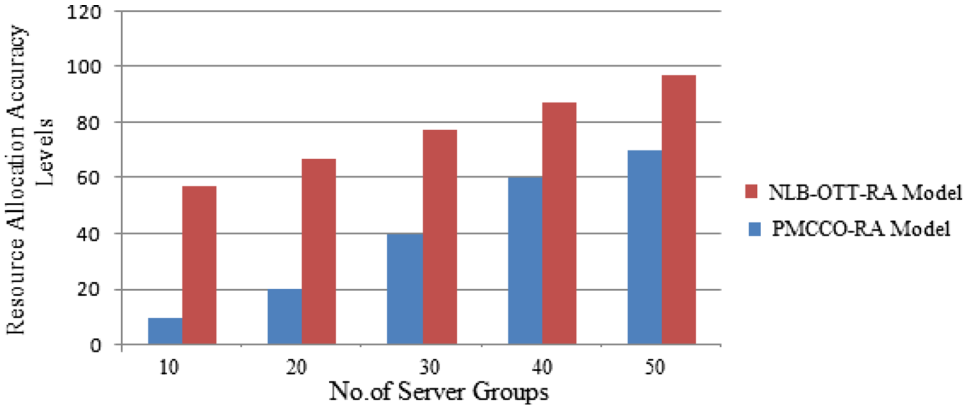

Fig. 3 Resource allocation accuracy levels

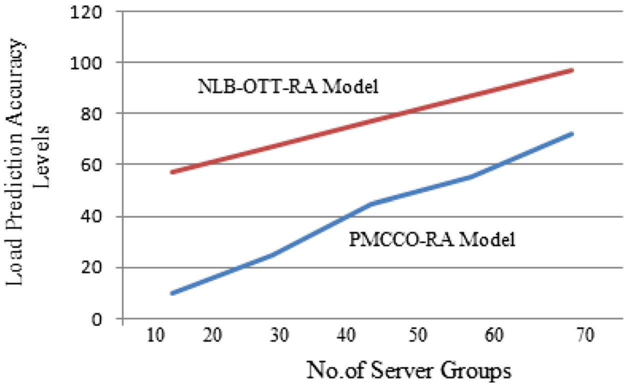

Fig. 4 Load prediction accuracy levels

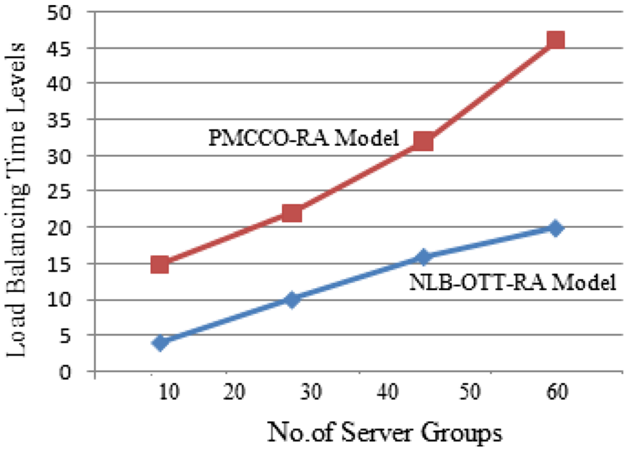

Fig. 5 Load balancing time levels

A cloud server ability to regulate traffic and assure the performance of key nodes with low capacity is referred to as quality of service. Organisations might prioritise select high-performance applications to alter their overall traffic and resource utilisation and task executions. The QoS levels of the proposed and existing cloud environments are shown in Fig. 7.

5 Conclusion

Modern cloud computing places a high value on reducing both consumption and maintenance costs. In reality, data centres require a lot of electricity, which reduces performance and releases a lot of carbon dioxide into the atmosphere. Server virtualisation, relocation, and consolidation are just a few of the technologies that are utilised to optimise network resources and reduce energy usage. Scheduling resources is a critical function in cloud computing environments. Each and every customer request must be recognised and responded effectively without any delay. Reduced energy consumption, cheaper prices, and so on are examples of possible goals. According to ability to discern such as resource use and monitoring, demand data, and other parameters, the resource scheduler provides alternatives for allocating available resources. According to this plan, the number of active hosts would be reduced, allowing idler machines to be put to sleep. To maximise the number of cloud host servers that can be utilised post-service exits, a quadratic mathematical optimisation technique was adopted. This migration strategy is combined with an accurate allocation mechanism to reduce total data centre energy consumption. The presented methods can be used as an energy consumption aware virtual machine scheduler to optimise physical infrastructure management and operators. Performance and cost are directly affected by resource scheduling, as well as indirectly by bad performance, which makes it more expensive or less effective to use extra cloud host servers in the cloud centre. This paper proposed an efficient normalised load balancing optimised time triggered resource allocation (NLB-OTT-RA) model fuzzy with minimum energy consumption based on fuzzy in the cloud environment to improve the system performance. The proposed work is predicated on making use of as much of each resource's capabilities as possible for a specific number of VMs. In future, more number of virtual machine linking and live migrations can be optimised to improve the cloud performance.

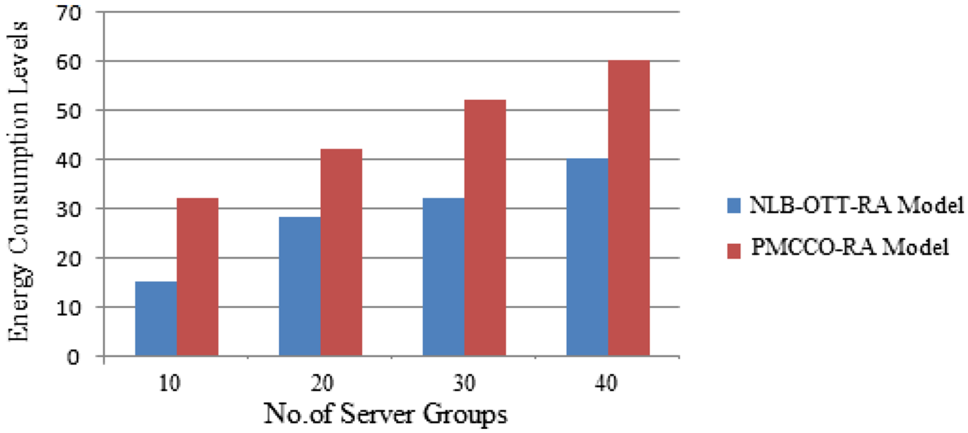

Fig. 6 Energy consumption levels

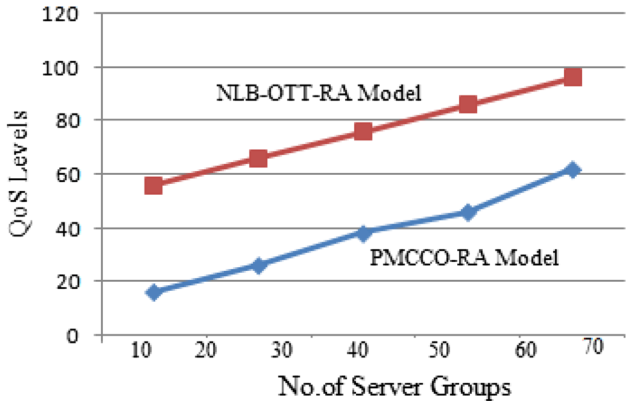

Fig. 7 QoS levels

Funding No funding is applicable.

Availability of data and material Not data and materials are available for this paper.

Code availability The data and code can be given based on the request.

Declarations

Conflicts of interest The authors declare that they have no conflict of interest. The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Ethical approval The article has no research involving human participants and/or animals.

Competing interest The author has no financial or proprietary interests in any material discussed in this article.

References

Addya SK, Turuk AK, Sahoo B, Satpathy A, Sarkar M (2018) A game theoretic approach to estimate fair cost of VM placement in cloud data center. IEEE Syst J 12(4):3509–3518

Al-Dhuraibi Y, Paraiso F, Djarallah N, Merle P (2018) Elasticity in cloud computing: state of the art and research challenges. IEEE Trans Serv Comput 11(2):430–447

Ali Z, Khaf S, Abbas ZH, Abbas G, Muhammad F, Kim S (2020) A deep learning approach for mobility-aware and energy-efficient resource allocation in MEC. IEEE Access 8:179530–179546. https://doi.org/10.1109/ACCESS.2020.3028240

Azizi S, Shojafar M, Abawajy J, Buyya R (2020) GRVMP: a greedy randomized algorithm for virtual machine placement in cloud data centers. IEEE Syst J 15(2):2571–2582

Bashir AK, Arul R, Basheer S, Raja G, Jayaraman R, Qureshi NMF (2019) An optimal multitier resource allocation of cloud RAN in 5G using machine learning. Trans Emerg Telecommun Technol 30(8):e3627

Diallo M, Quintero A, Pierre S (2019) An efficient approach based on ant colony optimization and tabu search for a resource embedding across multiple cloud providers. IEEE Trans Cloud Comput

Lin W, Zhang Y, Wu W, Fong S, He L, Chang J (2020a) An adaptive workload-aware power consumption measuring method for servers in cloud data centers. Computing

Gai K, Qiu L, Qiu M, Zhao H (2020) Cost-aware multimedia data allocation for heterogeneous memory using genetic algorithm in cloud computing. IEEE Trans Cloud Comput 8(4):1212–1222

Ghahramani M, Javidan R, Shojafar M, Taheri R, Alazab M, Tafazolli R (2021) RSS: an energy-efficient approach for securing IoT service protocols against the DoS attack. IEEE Internet Things J 8(5):3619–3635

Ghanavati S, Abawajy JH, Izadi D (2020) An energy aware task scheduling model using ant-mating optimization in fog computing environment. IEEE Trans Serv Comput 15(4):2007–2017

Han G, Que W, Jia G, Shu L (2016) An efficient virtual machine consolidation scheme for multimedia cloud computing. Sensors 16(2):246

Hong C-H, Varghese B (2019) Resource management in fog/edge computing: a survey on architectures infrastructure and algorithms. ACM Comput Surv (CSUR) 52(5):1–37

Horri A, Mozafari MS, Dastghaibyfard G (2014) Novel resource allocation algorithms to performance and energy efficiency in cloud computing. J Supercomput 69(3):1445–1461

Huang Y, Yang R, Cui L, Wo T, Hu C, Li B (2014) VMCSnap: taking snapshots of virtual machine cluster with memory deduplication.

In: Proceedings of the IEEE 8th international symposium service oriented system engineering, 314–319

Hussain M, Wei LF, Lakhan A, Wali S, Ali S, Hussain A (2021) Energy and performance-efficient task scheduling in heterogeneous virtualized cloud computing. Sustainable Comput Inf Syst 30:100517

Jayaraman R, Manickam B, Annamalai S, Kumar M, Mishra A, Shrestha R (2023) Effective resource allocation technique to improve QoS in 5G wireless network. Electronics 12:451. https://doi.org/10.3390/electronics12020451

Liang Y, Hu Z, Li K (2020) Power consumption model based on feature selection and deep learning in cloud computing scenarios. IET Commun 14(10):1610–1618

Liu X, Cheng B, Wang S (2020) Availability-aware and energy-efficient virtual cluster allocation based on multi-objective optimization in cloud datacenters. IEEE Trans Netw Service Manag 17(2):972–985

Liu C, Li K, Li K (2021) A game approach to multi-servers load balancing with load-dependent server availability consideration. IEEE Trans Cloud Comput 9(1):1–13

Mavridis I, Karatza H (2019) Combining containers and virtual machines to enhance isolation and extend functionality on cloud computing. Future Gener Comput Syst 94:674–696

Miao Z, Yong P, Mei Y, Quanjun Y, Xu X (2021) A discrete PSO-based static load balancing algorithm for distributed simulations in a cloud environment. Fut Gener Comput Syst 115:497–516

Omer S, Azizi S, Shojafar M, Tafazolli R (2021) A priority power and traffic-aware virtual machine placement of IoT applications in cloud data centers. J Syst Archit 115:101996

Omoniwa B, Hussain R, Javed MA, Bouk SH, Malik SA (2019) Fog/edge computing-based IoT (FECIoT): architecture applications and research issues. IEEE Internet Things J 6(3):4118–4149

Parvizi E, Rezvani MH (2020) Utilization-aware energy-efficient virtual machine placement in cloud networks using NSGA-III meta-heuristic approach. Clust Comput 23:2945–2967

Rahimi MRN, Venkatasubramanian N, Mehrotra S, Vasilakos AV (2018) On optimal and fair service allocation in mobile cloud computing. IEEE Trans Cloud Comput 6(3):815–828

Raj PH, Kumar PR, Jelciana P (2018) Load balancing in mobile cloud computing using bin packing's first fit decreasing method. Proc Int Conf Comput Intell Inf Syst 3:97–106

Ruan F, Gu R, Huang T, Xue S (2019) A big data placement method using NSGA-III in meteorological cloud platform. EURASIP J Wireless Commun Netw 2019(1):1–13

Shanmuganathan V, Suresh A (2023) LSTM-Markov based efficient anomaly detection algorithm for IoT environment. Appl Soft Comput 136:110054

Shojafar M, Canali C, Lancellotti R, Abawajy J (2020) Adaptive computing-plus-communication optimization framework for multimedia processing in cloud systems. IEEE Trans Cloud Comput 8(4):1162–1175. https://doi.org/10.1109/TCC.2016.2617367

Sohani M, Jain SC (2021) A predictive priority-based dynamic resource provisioning scheme with load balancing in heterogeneous cloud computing. IEEE Access 9:62653–62664. https://doi.org/10.1109/ACCESS.2021.3074833

Sotiriadis S, Bessis N, Buyya R (2018) Self managed virtual machine scheduling in cloud systems. Inf Sci 433:381–400

Tuli S, Ilager S, Ramamohanarao K, Buyya R (2020) Dynamic scheduling for stochastic edge-cloud computing environments using A3C learning and residual recurrent neural networks. IEEE Trans Mobile Comput 21(3):940–954

Verbelen T, Stevens T, De Turck F, Dhoedt B (2013) Graph partitioning algorithms for optimizing software deployment in mobile cloud computing. Future Gener Comput Syst 29(2):451–459

Wu H, Sun Y, Wolter K (2020) Energy-efficient decision making for mobile cloud offloading. IEEE Trans Cloud Comput 8(2):570–584

Xiong FU, Zhou C (2015) Virtual machine selection and placement for dynamic consolidation in cloud computing environment. Front Comput Sci 9(2):322–330

Xu X et al (2019) A computation offloading method over big data for IoT-enabled cloud-edge computing. Fut Gener Comput Syst 95:522–533

Yu H, Yang J, Wang H, Zhang H (2019) Towards predictable performance via two-layer bandwidth allocation in cloud datacenter. J Parallel Distrib Comput 126:34–47

Zhang L, Han T, Ansari N (2018) Energy-aware virtual machine management in inter-datacenter networks over elastic optical infrastructure. IEEE Trans Green Commun Netw 2(1):305–315

Zhang C, Wang Y, Wu H, Guo H (2021) An energy-aware host resource management framework for two-tier virtualized cloud data centers. IEEE Access 9:3526–3544. https://doi.org/10.1109/ACCESS.2020.3047803

Zhong W, Zhuang Y, Sun J, Gu J (2018) A load prediction model for cloud computing using PSO-based weighted wavelet support vector machine. Appl Intell 48(11):4072–4083

Zhou Z, Hu ZG, Song T, Yu JY (2015) A novel virtual machine deployment algorithm with energy efficiency in cloud computing. J Cent South Univ 22(3):974–983

Zhou Z, Abawajy JH, Li F, Hu Z, Chowdhury MU, Alelaiwi A et al (2018) Fine-grained energy consumption model of servers based on task characteristics in cloud data center. IEEE Access 6:27080–27090

Zhou Z et al (2018) Minimizing SLA violation and power consumption in cloud data centers using adaptive energy-aware algorithms. Fut Gener Comput Syst 86:836–850

Zhou Z, Shojafar M, Alazab M, Abawajy J, Li F (2021) AFED-EF: an energy-efficient VM allocation algorithm for IoT applications in a cloud data center. IEEE Trans Green Commun Netw 5(2):658–669. https://doi.org/10.1109/TGCN.2021.3067309

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