Resource Allocation in Cloud Computing

G. Senthilkumar [1,*], K. Tamilarasi [2], N. Velmurugan [3], and J. K. Periasamy [4]

[1] Department of Computer Science and Engineering, Panimalar Engineering College, Chennai, India
[2] School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, India; Email: tamilarasi.k@vit.ac.in (K.T.)
[3] Department of IT, Saveetha Engineering College, Chennai, India; Email: velmurugan@saveetha.ac.in (N.V.)
[4] Department of Computer Science and Engineering, Sri Sairam Engineering College, Chennai, India; Email: jkperiasamy@gmail.com (J.K.P.)
*Correspondence: gsenthilkumarphd@gmail.com (G.S.)

***Abstract*—Cloud computing seems to be currently the hottest new trend in data storage, processing, visualization, and analysis. There has also been a significant rise in cloud computing as government organizations and commercial businesses have migrated toward the cloud system. It has to do with dynamic resource allocation on demand to provide guaranteed services to clients. Another of the fastest-growing segments of computer business involves cloud computing. It was a brand-new approach to delivering IT services through the Internet. This paradigm allows consumers to access computing resources as in puddles over the Internet. It is necessary and challenging to deal with the allocation of resources and planning in cloud computing. The Random Forest (RF) and the Genetic Algorithm (GA) are used in a hybrid strategy for virtual machine allocation in this work. This is a supervised machine-learning technique. Power consumption will be minimized while resources are better distributed and utilized, and the project's goal is to maximize resource usage. There is an approach that can be used to produce training data that can be used to train a random forest. Planet Lab's real-time workload traces are utilized to test the method. The suggested GA-RF model outperformed in terms of data center and host resource utilization, energy consumption, and execution time. Resource utilization, Power consumption, and execution time were employed as performance measures in this work. Random Forest provides better results compared with the Genetic Algorithm.**

***Keywords*—cloud computing, infrastructure-as-a-service, random forest, genetic algorithm, service level agreements, optimization, resource allocation**

I. Introduction

Cloud Computing is defined as a new-fangled method of dynamically analyzing and maintaining adaptable and virtualized assets on the Internet. Kaefer [1] suggested that Cloud Computing is a framework that enables on-demand access to a dispersed collection of sources like hosts and storage devices, resulting in responsibilities that could be kept and trashed with much less control work. CC proposes a cutting-edge trend in information generation that moves computing facts from the desktop to extensive statistics facilities. It is defined as a web-based application that is delivered as a service. The computational effort in all cloud settings is provided by a collection of knowledge centers spread throughout the globe that are linked by a high-speed network. In cloud computing, a cloud could be a dispersed system of collection. It distributes digital clients with essential static assets at some point on the Internet. Mary *et al.* [2] suggested that the resource control approach aids in the synchronization of data assets to govern work completed by both cloud consumers and providers. It's known as Resource Allocation (RA) from the Resource's providers to the Resource's customers. Because the aid administration allows for dynamic asset reallocation, the user can better use the available capacity. Resource Allocation in CC is the process of allocating public resources over the web to the appropriate cloud apps. Infrastructure as a Service (IaaS) gives assets to contested requests due to a predetermined aid allocation mechanism.

Cloud computing uses the Internet and centralized servers to keep data and programs. Using this way, end-users and organizations can utilize programs without entering and verifying any confidential information on any computer connected to the Internet. Cloud computing allows for more efficient computing by standardizing data, memory, bandwidth, and other resources. Email services like Hotmail, Gmail, and Yahoo are all cloud computing models. The cloud service provider manages all the host and email management software in that cloud. The end consumer can enjoy the benefits of the software on their own. Cloud computing serves more as a convenience than a product by sharing resources, programs, and data among computers and other devices. (i) Software-as-a-Service (SaaS), (ii) Platform-as-a-Service (PaaS), and (iii) Infrastructure-as-a-Service (IaaS) are three types of cloud computing services. Allocating Cloud resources must fulfill the Quality of Service (QoS) limitations given by clients through Service Level Agreements (SLA) and should also reduce energy usage.

Nair *et al.* [3] suggested that in order to execute cloud services as efficiently and effectively as possible, resources must be allocated. Alternatively, it might be considered as any process that ensures that the infrastructure of the service provider can meet the specified requirements of the SLA. In order to satisfy the

Manuscript received January 4, 2023; revised March 9, 2023; accepted March 23, 2023; published October 13, 2023.

demands of cloud applications in an elastic and transparent way, resource allocation is defined as the process of integrating cloud provider activities in order to utilize and allocate finite resources, which may appear limitless to users [3]. The two key parties in cloud computing (users and service providers) benefit from resource allocation schemes. In light of Cloud computing's emphasis on quality and dependability; users may want to predict their workload in order to finish a project ahead of schedule. On the other hand, this could lead to an over-provisioning problem. Providers, on the other hand, want to optimize their profit by using fewer resources per user in order to accommodate more customers and generate more revenue. A lack of resources will result. However, the absence of information sharing between them makes it impossible to distribute resources in a mutually beneficial way.

Mohan *et al.* [4] improved that as an attractive computing model, cloud computing attracts many clients, projects, and internet organizations. The primary objective of both cloud service providers and cloud service consumers is to maximize cloud resources and maximize financial profit. Among the most immediate concerns in cloud computing seems to be the distribution of scarce resources. Consumers' perceptions of resource allocation are shaped by how products and services were spread among such populations. The more productive an economy is, the more resources it would have to work. Customers gain access to the resources they need for a thorough job through using skills as a service. It prevents clients from having to pay for resources that are never used. Businesses might modernize existing operations by allowing customers to access a most up-to-date application and infrastructure changes. Whenever a given goal function requires the best possible job-resource matches in place and time, dispersed systems' allocation of resources and planning play a critical role in determining the best possible match. Given the complexity of the best resource allocation, i.e., efficient distribution with scarce resources and most significant profit, allocating resources to cloud users seems to be a complicated procedure. A sequence replica has been used to determine the cost of services in a cloud. Workflow programs may be implemented more quickly because of dynamic resource allocation, allowing customers to define their policies. A cloud computing infrastructure's resource allocation replication seems that several resources from such a uniform resource group were distributed simultaneously.

With the increase in computer and mobile users in recent years, data storage has become a priority in all fields. Large and small-scale businesses today rely on data, and they spend a lot of money to keep this data up to date. Cloud storage makes IT services available on-demand via Large distributed data centers and high-speed networks. Network virtualization is a recent trend in cloud computing that has emerged as a multifaceted approach to the future internet by facilitating shared resources. Provisioning of the virtual network is regarded as a significant challenge in terms of creating NP-hard problems, minimizing workflow processing time under control resources, and so on.

Dealing with resource allocation and planning in cloud computing is both necessary and difficult. In this work, the Random Forest (RF) and the Genetic Algorithm (GA) are combined in a hybrid strategy for virtual machine allocation. This is an example of supervised machine learning. The project's goal is to maximize resource usage by minimizing power consumption while better distributing and utilizing resources. There is a method for producing training data that can be used to train a Random Forest. The method is tested using Planet Lab's real-time workload traces. The work flow is organized as shown below. Section II describes related works, Section III explains the proposed method, and Section IV describes case studies, findings, and discussions.

II. Related Work

Several studies have been done on analyzing cloud platform resource allocation. Algorithms for allocating resources have been tried using different scheduling methods. Numerous factors like maximum throughput, high productivity, QoS, and SLA awareness are considered while allocating resources in the cloud. An overview of many of the studies directly related to the allocation of resources is given below. Emeakaroha *et al.* [5] suggested that an improved resource scheduling technique based upon thorough research into IaaS (Infrastructure-as-a-Service) cloud systems was suggested to achieve cloud scheduling optimization. The flexible method of spreading virtual machines is explored to utilize physical resources best. The automated development policy has been implemented using an IGA [5].

Prodan and Ostermann [6] suggested that a market-based resource allocation strategy (RAS-M) to reduce the number of resources consumed by large data centers while providing better quality services to cloud clients. As part of cloud computing, resources can be allocated flexibly in, an on-demand manner. Schlegel *et al.* [7] suggested that Service Level Agreements (SLA) regulate how cloud services are delivered. SLA violations must be strictly enforced to minimize costly fines. There seem to be several SLA criteria to examine when allocating services. In this case, a scheduling mechanism with several SLA parameters has been evaluated. There are three types of task scheduling strands in clouds: Three examples of service models have been developed for infrastructure, platform, and software. It is possible to grant secure access to the information to several independent components without incurring any additional cost of maintenance. A major challenge in cloud computing involves ensuring that clients get the most from the resources they've purchased, considering their computing limits. If this challenge isn't met, the cloud system as a whole could be shut down. The strategy aims to distribute resources dynamically and make the best use over a set period [7]. A Rule-Based Resource Allocation Model (RBRAM) is suggested, which allocates resources due to task importance to avoid inefficiencies and utilization of resources. It is possible to rent computing resources via the Internet utilizing cloud computing, known as IaaS. The customer could select from various computer resources based on their specific

requirements. The IaaS concept has been used to allocate resources toward completing all the tasks. There are strict deadlines for completing Real-Time projects.

Using this approach, compute, and network resources are given equal weight in resource allocation decisions. Kumar *et al.* [8] suggested that Federated Computing and Network System (FCN) is a new technology designed to handle this joint allocation. The problem of congestion arises whenever a service request demands the distribution of various resources (like computing power, storage server, and connectivity). To make effective utilization of resources together in a congested scenario, You *et al.* [9] presented a congestion control mechanism. Because various resources are being allocated concurrently to meet each service demand, this method can be seen as sustainable management of available resources. It is a requirement of cloud computing services that allocate bandwidth to enter the distribution skill simultaneously. Yan *et al.* [10] suggested here to give resources together in a cloud computing system with a limit on electric power capability. Both administration ability and platform bandwidth are required simultaneously. Several resources from such a generic resource group get simultaneously assigned to every search for a specific age under the resource distribution method for such a cloud services environment. Provision ability and capacity are evaluated here.

As a result, this part comprises various studies relating to IaaS cloud computing's resource allocation. While numerous solutions to scheduling, QoS allocation of resources, unpredictability, energy usage, and other problems were brought forward, none of them could achieve a real-time allocation of resources. Since then, we've proposed a new model dependent on reinforcement learning-based machine learning algorithms that help us get a higher success rate regarding the allocation of resources.

A. Cloud Architecture

The cloud computing architecture can be divided into three tiers as shown in Fig. 1. There are three types of resources: (a) resource, (b) platform, and (c) application. The first tier, the Resource layer (infrastructure layer), was also composed of two characteristics. (i) Computes on a physical and virtualized basis, (ii) Networking, Storage Resources and Memory Devices suggested by Mochizuki and Kuribayashi [11]. For example, customers can only pay for the section they interact with within storage rather than acquiring a disc or knowing nothing about the area of the record [12].

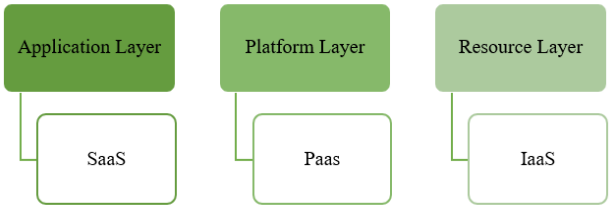

Figure 1. Cloud computing architecture.

Aarthi [13] suggested that the application layer (also known as software as a service) restores the system's functions. If you use SaaS, you won't have to install and run an extensive software package on your computer. Using a pay per use sample instead of purchasing a high-priced software product can help you save money overall. There are four deployment models in the cloud mannequin. (1) Private Cloud: Organizations only use the cloud infrastructure seldom. It could be managed either by the company or by a third party. It might be on-premises or off-premises. (2) Community Cloud: In this cloud, the infrastructure is shared amongst the exclusive corporations, assisting a specific group of people who had similar difficulties. Mission, safety standards, policy, and compliance issues, to name a few. Government authorities or a third party could manage it. It can be found both on and outside the grounds. (3) Public Cloud: obtaining cloud ownership through a corporation that provides cloud services. It is accessible to the general population as well as extensive industry crews. (4) Hybrid Cloud: Each public and private cloud has a hybrid cloud charter. It may remain different companies, but it is bound together by controlled processes that allow for the transfer of records and software.

B. Cloud-Based Resource Allocation

As previously said, Resource Allocation is a method for assigning evaluable assets to the necessary cloud apps in internet-based cloud computing. Although scarce resources are shared, one of the most pressing issues in Cloud Computing is Resource Allocation. Resource Allocation is concerned with the distribution of services among customers from the consumer's point of view. Resource Allocation advantages include the fact that the user does not need to install any hardware or software in order to access, develop, and store the application through the Internet. There are no restrictions on where you can go or how you can get there. The data and application can be accessed by the customer from any location and on any platform in the universe. Cloud providers disseminated their resources throughout the internet during times of scarcity. Next, we'll look at four distinct ways to use the cloud provider's calculating limits.

- Reservation in advance: Reserved assets must be always available in this case.
- Best effort: Restoring assets as quickly as possible under the circumstances is done at the best effort level.

Immediate: A client's request can be accepted or rejected based on available resources. If the client's request is accepted, resources are made available to them right away.

- Deadline sensitive: When the planning calculation of Haize can assure that it will be completed before its time limit, it is just preemptible. It's considered pre-emptive, although they're only able to preempt to a certain degree.

C. Techniques for Scheduling Resources

Techniques for allocating resources in the cloud environment can be employed in a wide variety of ways.

Experts in this field are always developing and inventing new methods to meet the demands of the times. Even if new technologies are expected daily, there are a few that cannot be avoided and whose contributions are critical. We'll go over a few examples of that here, which is shown in Fig. 2.

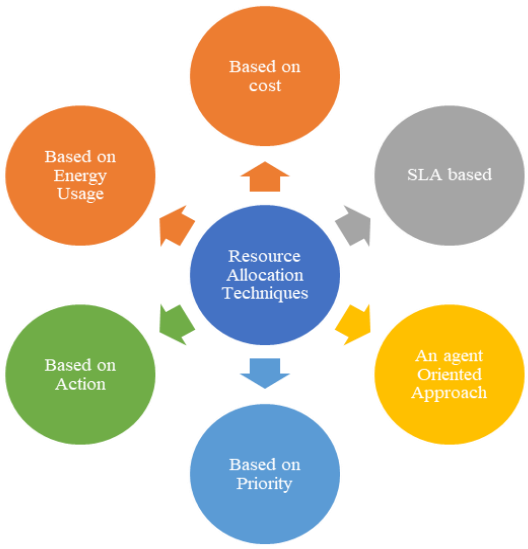

Figure 2. Resource allocation technique.

D. Cost-Based Prioritization of Resources

Devagnanam [14] discussed the reservation plan and the on-demand plans that were discussed in detail. When compared to the on-demand plan price, the reservation plan price should be lower. Because the customer has already agreed to pay the amount. However, because of the uncertainty surrounding the pricing of supplier resources, no mention is made of the user's future demand price. A new method has been proposed that incorporates the long-term plan into the solution. Cooperative cloud markets were established by this author to allocate resources. The goal is to keep costs down while yet providing the consumer with the resources they need. Costs can be reduced by using this algorithm. Jen [15] discussed the fundamentals of the market-oriented cloud were laid forth.

E. SLA-based Resource Allocation

With this method, the user and the service provider agree on a service level agreement. The SLA must meet the needs of the end user. Zhang *et al.* [16] introduced an SLA-Based Resource Allocation model. Multiple cloud service providers were explored. The Nash equilibrium notion of game theory is used to distribute resources to the consumers in this strategy. Multi-cloud service provider resource allocation is a game. There are servers for each CSP and requests are routed accordingly. Yang *et al.* [17] examined how to reorganize resources. Resource reallocation was proposed by the authors. Buildings are broken up into two halves. The SLA contract and resource scheduling are handled by the data center and customer, respectively, in the first section, and the control chart is shown in the second section to detect SLA violations for host performance. Evaluation is based on the use of a utility function.

F. An Agent-Oriented Approach to Allocation of Resources

Recently, several studies have been conducted on the topic of cloud resource optimization. When it comes to cloud resource optimization, Zheng and Wang [18] developed a solution based on map reduction [19]. PSO is concerned with the ability of a flock of birds to fly toward a common goal while also conducting their own independent searches in the process. It is possible for each particle to communicate with one another and update its new location. Every particle has a unique benefit, but there is also a universally beneficial one. The particle is moving at a certain rate. There are three parts to the cloud resource optimization framework: time, money, and satisfaction of the end user.

When determining how much work must be done and where resources should be allocated, the proposed algorithm is employed. The ant colony method has been utilized by Seyedehmehrnaz [20] to assign assets to consumers. An ant's foraging strategy informs this strategy. An ant-inspired algorithm uses NP-solving techniques to mimic the behavior of wild ants in their search for food. Paths can also be kept as short as possible using this approach. The ACO algorithm is run on the work that the work agent collects from the customer. It has been shown that pheromones have a weight that is used to communicate. The Ant colony method selects the nodes for the network. This policy governs the allocation of cloud resources. Fruit Fly Optimization (PFOA) was proposed by Seyedehmehrnaz [20] to allocate resources and arrange tasks in the cloud. To get the general public's attention, an analysis is being produced. Sadashiv and Goudar [21] improved that priority-based k-mean algorithm is used to allocate the tasks. When the improved scent value is not available, the search continues. Resource allocation can be made more efficient by reducing costs. Goutam and Yadav [22] presented a new optimization algorithm in order to reduce implementation costs while still meeting customer demands. The technique is designed to be used in a cloud data centre to create web applications.

The cloud provider and web applications had been used as inputs. Using an algorithm, cloud assets can be consolidated in the most efficient way possible Genetic algorithm was used to predict cloud data centre resources. Based on the current statistics, further resources are expected to be available. The Genetic algorithm is used to complete the forecasting. The forecasting outputs are used to assign Virtual Machine.

G. Based on Priority, the Allocation of Resources

Neethu and Babu [23] proposed an approach to accomplishing high-priority tasks. In addition, the idea of reusing the virtual machine is discussed. The new work does not necessitate the creation of new virtual machines. It is possible to carry out a high-priority task by suspending a low-priority one. If a virtual machine has completed its single task, the hanged job might be reopened in that virtual machine. A mechanism for allocating resources was presented by Iniya and

Ranjith [24]. Peer-to-peer cloud computing is assumed. The need for a larger amount of processing power is prioritized above other tasks. Priority sets, such as high, medium, and low, are created using the K algorithm, and the jobs in the list are sorted accordingly. The cost of a resource is constantly and continuously assessed in the suggested technique based on the current demand for that resource. There is a lot of debate on the internet about the introduction of assets from peer clouds. Mehiar *et al.* [25] presented a technique to assign resources with Fault Tolerance. From the lowest priority task to the highest priority task, pre-emption begins. Here, a more sophisticated reservation mechanism is put out. By using this method, a priority-based work list is generated and then prioritized tasks are assigned. Jobs with high priorities might be reserved. Additionally, this approach can be used for cloud application failure tolerance.

H. Allocation & Energy efficiency of Resources

Iniya and Ranjith [24] suggested that the "auction technique" is a resource allocation strategy used in the market-oriented cloud. In the Action Method, resources are bought and sold based on user preferences or requirements. The primary functions of this approach are optimal price detection and determining the auction winner. Portable Format for Analytics (PFA) is a tool for locating the ideal buyer and seller. This study focused on the auction participants' preferences for selling and purchasing resources at the price they were satisfied with. Combinatorial auctions are employed in this case (both consumers and providers are participating). Dishonest players in the auction can be identified and eliminated by the identifiers. When society has confidence in the system's participants, it can approach the system. The system is responsible for delivering high-performance computing resources. Virtual Machine using Section and Replacement policy proposed by Itachoudhary *et al.* [26] Resources are allocated through identifying overloaded resources. Energy consumption by using idle server is assigned for on demand request by selection and replacement of virtual machine to save the energy consumed by servers.

I. Allocating Resources in Light of Energy Consumption

In the mobile cloud, energy consumption is a big concern. This topic has been the subject of numerous studies [27]. Chang *et al.* [27] suggested that aims to reduce energy consumption in cloud data centres. The proposed framework accomplishes the following goals. Determine the number of resources required and the number of virtual machine requests that will be made. It is estimated how many Physical Machines (PMs) are required to meet user demands. Reduces the amount of energy used by cloud data centers. Data classification, workload prediction, and power management are all included in this system. Fatma *et al.* [28] outlined a strategy for allocating resources while simultaneously lowering energy consumption. Virtual machine migration can be made more efficient by using this strategy. CPU use and virtual machine migration were predicted using a time series forecasting method developed by the authors. Rajathi and Devagnanam [29] conducted that IaaS cloud has been considered using IaaS cloud as a foundation. They have presented two models: one for optimizing server power, and the other for reducing server resource waste in an IaaS cloud using a methodology that measures waste. Virtual Machine requests from a variety of users with varying resource requirements were taken into account for the resource wastage model, while power usage on the server was taken into account for the server power model.

J. Adaptive Management of Resources

In the Cloud Computing system, confidence in managing apps and resources has been established through an adaptive management strategy known as "obscurity". As a result, this article examines how obscurity strategies are evaluated and applied in the Cloud Computing part. To handle the unpredictable workload, Brownout is an autonomous adaptive framework that qualifies or disqualifies arbitrary sectors, such as application components or services. In order to remain flexible, this brownout mechanism was devised. The vital activity of the model can be ensured by temporarily disabling arbitrary areas of the model. It is possible to increase the pace of request acceptance by deactivating certain arbitrary activities and reusing those resources for more important ones. In addition, a brownout allows for increased resource promotion while keeping the apps functional in order to avoid heavy workloads.

III. The Proposed Technique

The virtual machine placement method proposed in this work is based on a genetic algorithm and a random forest algorithm. Random forest is trained using a dataset generated by a genetic algorithm, which is an optimization technique. Virtual machines are assigned to physical computers in the dataset for training purposes. In this paper, a hybrid model based on a genetic algorithm and a random forest technique is proposed for virtual machine placement. A genetic algorithm is an optimization technique that will generate the dataset needed to train the random forest algorithm in this case. The training dataset consists of virtual machine allocation or mapping to physical machines.

A. Genetic Algorithm

One of the most popular methods for finding the best answer seems to be a Genetic Algorithm (GA). A VM-to-Physical Machine mapping is generated by a genetic algorithm in this case. Our method includes the following steps as shown in Fig. 3:

Figure 3. Genetic algorithm workflow.

Step 1: Population Initializing
Step 2: Fitness Function
Fitness (i)= β*Power Efficiency
Step 3: Tournament Selection Method
Step 4: Cross over—for producing novel offspring 2 point crossover is selected
Step 5: Mutation—Operations like, rebalancing, swap, move, and swap and move

There are 50 evolutions are used in our method. Hence, virtual machines will be mapped to physical machines in the end.

Creation of the initial population

Applying three different heuristic ranking strategies on the first three of the chromosomes.

```
For i from 3 to PopSize-1 do
   For j from 0 to ChSize-1 do
      produce at random a gene with the parameters
      that has not been generated in previous genes.
      In order to maintain a Topological Order that is
      reliable, Transfer Chromosome i from its current
      position on the left to its new position on the
      right in the queue.
    End for
End for
Push Sub-Task to Priority Queue
   While not Empty (Priority Queue)
      Pop Ti (First Sub-Task) from Priority Queue
            For each processor Pk do
            Insertion-based HEFT Scheduling Policy()
                Assign Ti to the Processor Pk
          End For
    End While
Return makespan
```

Sub-Task allocation to processors

Input: Two parents from the existing population
Output: Two offspring

1. Select at random an appropriate crossing point
2. Separate the chromosomes of the father and mother into left and right parts
3. Create a new offspring, specifically a son Transfer the father's left chromosomal segment to the son's left chromosome segment.
4. Copy genes from the mother's chromosome that do not occur in the father's left chromosome segment to the right chromosome segment of the son.
5. Produce a new offspring, specifically a daughter Transfer the mother's left chromosomal segment to the daughter's left chromosome segment.
6. Copy genes from the father's chromosome that do not occur on the mother's left chromosome to the daughter's right chromosome.

Return (Two selected offspring).

Crossover mapper

```
Input:    Mappers
Output: New Mappers
For each Mapper do Algorithm 5 (A chromosome Ti
that has been picked at random)
   Execute the Map Reduce task manager
   Record the result as M blocks
   Transfer the outcome to the shuffle step.
   The best person should be written to the global
   file in DFS if all individuals have been processed
   successfully
Return (New Mappers).
```

Mutation operation

```
Input:    A chromosome Ti that has been picked at
          Random.
Output:  A new chromosome is being created.
Line 1: from Succ (i) select the first successor Tj
  In the interval [i+1, j-1] Choose a gene
  Tk randomly
 For all Tl  member Pred (k)
   If l < i then
     Swapped (Ti,Tk) to create a new
     generation.
     Return the new offspring
   Else
     Go to Line 1
    End If
 End for
```

Mutation reduces

(1) Give each Reducer the result of Algorithm 6.
(2) Activate the MapReduce task manager.
(3) Begin the process of calculating.
(4) Incorporate the findings into M-sized blocks.
(5) Verify that the termination condition has been met
(6) If all individuals are properly processed then Write (best individual, DFS).
(7) The result should be sent to the cluster.

B. Random Forest Classifier

Random forest is a machine-learning technique that can be used for both regression and classification purposes. It is among the most versatile and easy-to-understand algorithms out there. More trees mean a more resilient random forest, and this is true for all random forests. At each node, a random forest selects a small number of attributes to divide and then calculates the optimal split based on these features as in the training set, making each tree unique. When it's all said and done, it obtains a forecast out of each tree and then selects the best option by either "majority vote" or "performance voting". Fig. 4 shows the RF workflow.

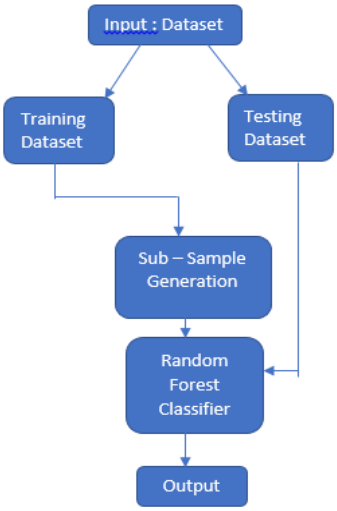

Figure 4. Workflow of random forest classifier.

The execution begins once the virtual machines are installed on the physical machines. After a while, a physical machine may receive more virtual machine's to execute; in this case, the physical machine becomes overloaded. To handle this type of scenario, we used a virtual machine migration technique. When a Physical Machines (PM) becomes overloaded, some of the virtual machines are chosen and migrated to other PMs with fewer loads, ensuring load balance across all physical machines in the data centre. One of the overload detection methods available is in the Interquartile Range (IQR), which is used to detect overloaded physical machines in the data centre. Following that, the virtual machine's to be migrated from the overloaded physical machine are chosen using the maximum correlation policy. It chooses the virtual machine's with the highest correlation of CPU utilization and other virtual machine's.

C. Performance Metrics

1) Energy consumption

Physical Machines (PM) in a data centre consume energy, and this number represents the sum of all their energy consumption. A linear cubic power consumption method is used to calculate PM's energy usage. This power model assumes that when CPU utilization rises, so does the physical host's power consumption.

Consider the following variables for the power model:

a. P_k^{max}: when the host k is fully utilized, the maximum power consumed.
b. P_k^{idle}: k is the host's idle power value.
c. U_k: host k's current CPU utilisation.
d. T: total number of data centre hosts.

$$\sum_{k=1}^{T} P_k = \sum_{k=1}^{T} \left[P_k^{idle} + \left(P_k^{max} - P_k^{idle} \right) \times U_k^3 \right] \quad (1)$$

2) Execution time

From the cloud provider's perspective, it's critical to complete all user requests as quickly as possible. As a result, one of the most important performance metrics for evaluating algorithms is the execution time.

$$Execution\ Time = \sum_{j=1}^{M} CT\_VM_j \quad (2)$$

3) Resource allocation

The Cloud data centre creates various types of VMs to process a user's request based on the user's resource requirements. The VMP technique aims to improve resource utilization by deploying a virtual machine on a suitable physical machine. The CPU resource type has been considered. Based on the number of resources requested by the customer, the Cloud data centre can build a variety of Virtual Machines (VMs). The placement of virtual machines on such an appropriate physical machine is one of the goals of the VMP approach. Consideration has been given to the CPU.

$$maximize \sum_{i=1}^{N} UCPU_i \quad (3)$$

IV. Performance Validation

To test the proposed algorithm, we used the CloudSim 3.0 toolkit simulator. Cloudsim offers a variety of VM provisioning techniques and virtualized resources. We used real workload traces from PlanetLab to conduct the experiment. PlanetLab is a component of the CoMon project, which measures CPU utilization from over 1000 virtual machines running on various hosts in over 500 locations worldwide. We used four different types of virtual machines in our experiment: Micro, Small, Medium, and Extra-Large instances, each with its own MIPS count. There are 800 HP ProLiant G4 and HP ProLiant G5 heterogeneous hosts deployed.

TABLE I. Initial Population Generation in a Random Manner

Virtual Machine	V M_2	V M_1	V M_3	V M_5	V M_7	V M_4	V M_6
Physical Machine	P M_3	P M_7	P M_1	P M_2	P M_4	P M_6	P M_5

TABLE II. Crossover Operation in the Proposed Model

Parent1: 2 Point crossover

Virtual Machine	V M_2	V M_1	V M_3	V M_5	V M_7	V M_4	V M_6
Physical Machine	P M_3	P M_7	P M_1	P M_2	P M_4	P M_6	P M_5

Parent2

Virtual Machine	V M_2	V M_1	V M_3	V M_5	V M_7	V M_4	V M_6
Physical Machine	P M_4	P M_5	P M_2	P M_6	P M_1	P M_3	P M_7

Table II shows the characteristics of these servers. As shown in Table II, the number of 500 hosts and VMs used for simulation ranges from 500 to 650 depending on the data center configuration. PlanetLab dataset is a log file from a real-world data center that contains incoming traffic, task size, VM size requested by task, and VM configuration such as RAM, number of processors, and MIPS count.

The metrics listed below are used to assess the proposed algorithm and other algorithms. It represents the total amount of energy consumed by all Physical Machines (PMs) in the data Centre. The linear cubic power consumption model is used to calculate the energy consumption of PMs. In this power model, as CPU utilization increases, the physical host's power consumption increases linearly. Fig. 5 shows the energy consumption of active PM's. Random Forest is higher in terms of genetic algorithm by 13%, 38%, 49%, and so on.

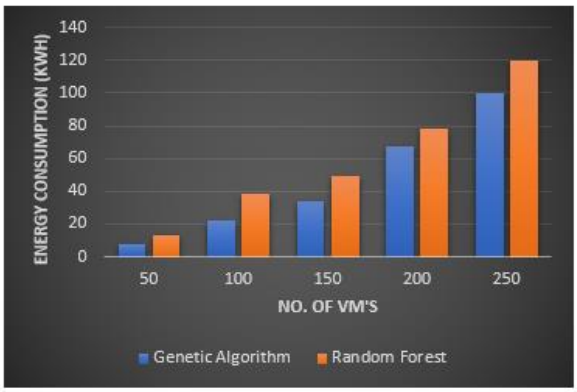

Figure 5. Performance of GA and RF in energy consumption.

From the standpoint of the cloud provider, completing all user requests in less time is critical. As a result, one of the key performance factors used to evaluate algorithms is execution time. Fig. 6 shows the executive time of active PM's. Random forest is higher in terms of genetic algorithm by 20%, 40%, 50%, and so on.

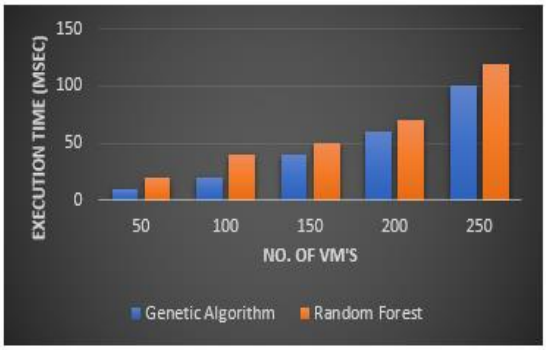

Figure 6. Performance of GA and RF in execution time.

Fig. 7 shows the CPU utilization of active PM's. random forest is higher in terms of genetic algorithm by 13%, 44%, 65%, and so on.

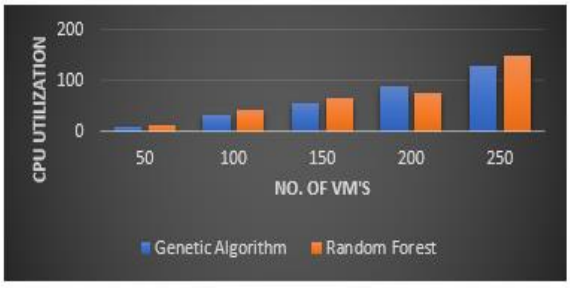

Figure 7. Performance of GA and RF in CPU utilization.

Time complexity is an important factor to consider when studying algorithm performance in cloud computing. There are several studies that demonstrate a comparative analysis of the complexity of genetic algorithm [23, 24]. According to the study, Genetic algorithm finds the best global optimal solution at the expense of a long search time, but it does so better than ant colony optimization and particle swarm optimization algorithms. To reduce the time required for genetic algorithm searching, the optimal solution is trained to a random forest model for training and further prediction, which provides an optimal solution in constant time. The cost overhead is only incurred once, during optimal solution generation using genetic algorithm and model training.

A. Significant Concerns in Allocating Resources

Fatma *et al.* [28], Rajathi and Devagnanam [29] suggested that the following eventualities must be avoided while allocating resources:

As many programs try to access the same resource simultaneously, there may be a problem.

Scarce resources: If there are not enough resources, it will happen.

Separation of resources: This will happen when the resources are divided. Even when sufficient resources are available, the required application will be unable to utilize them.

Overprovisioning: When the application is given more resources than it needs, this is known as overprovisioning.

Less provisioning: A condition called "less-provisioning" occurs when an application has fewer resources than requested.

B. Challenges of the Future

Modern cloud platforms have made a more effective allocation of resources possible. Many scheduling methods have been devised for dynamic and efficient resource allocation although. Since energy consumption directly impacts resource accessibility and management to ensure QoS in application implementations, properly monitoring energy usage is necessary. The management of resources and reducing energy use are also challenging to design objectives because of efficient allocation requirements [4].

RAS-M: Quality of Service (QoS) limitations can be met with RAS-M. Energy-saving can also be improved by assigning virtual machines in the most efficient way possible.

SLA violation: An SLA violation can be avoided by changing various SLA parameters, and in the future, SLA parameters raised in quantity to boost efficiency, hence reducing penalties.

Power saving: Past forecast models conserve electricity by turning off nodes that are not used. However, this doesn't achieve the entire power savings. To successfully save electricity, additional statistical data could be forecasted, so present power management techniques could be defeated.

Congestion control: The current congestion control methods are better, but instead, a superior method could be expected by further developing current strategies.

Despite preserving the required service quality for the program and attaining resource efficiency, cloud computing's modest open research faces a difficult challenge: convincing client software using the cloud. Considering future energy efficiency goals, planning, and software utilization as in the cloud could be studied in the future.

V. Conclusion

For businesses and venture capitalists alike, cloud computing is a game-changer. To reach the pinnacle of cloud computing, you must trade your data innovation viewpoint vehemently. In the end, real-time processing replaces advantage processing. In the context of cloud computing, a strategy for identifying assets is essential for satisfying customer requirements and enhancing the competitive advantage of cloud service providers. In the context of cloud computing, a few techniques to asset distribution have been introduced and sent thus far. In addition, a few resource allocation concerns and their solutions are explored in great length in this paper. It is, however, a technology that is constantly expanding its boundaries in response to the changing needs and demands of its users. As the cloud's most widely shown component, resource allocation makes perfect sense. For the time being, it is solely reliant on the specialized viewpoints. IT organizations must apply components connected to user resource allocation terms for CC to be a fast-expanding enterprise in the IT industry. This approach has the potential to deliver the best results in terms of overall performance and customer satisfaction in the same way. As a result, resource allocation is a difficult and time-consuming task that must be completed as soon as possible whenever a user requests it. A leader in the cloud computing industry will be able to allocate resources to customers while still making a profit and meeting their needs under a variety of conditions. Hence, our results show better results in active PM's. And Random Forest (RF) provides better results compared with Genetic Algorithm (GA). This can only be accomplished by thoroughly analyzing the request and swiftly deciding on an appropriate resource allocation method. Testing of the model will be possible in the future by various approaches to deep learning for better problem solving and performance analysis.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

G. Senthilkumar and K. Tamilarasi developed the theory and proposed model; N. Velmurugan and J. K. Periyasamy worked on data preparation, ML model for the dataset, and verified the analytical methods; All authors discussed the results and contributed to the final manuscript; All authors confirm sole responsibility for the following: study conception and design, data collection, analysis and interpretation of results, and manuscript preparation; All coauthors have seen and agreed with the contents of the manuscript.

References

[1] G. Kaefer, "Cloud computing architecture," *Computer Networks*, vol. 12, pp. 23–33, May 2010.

[2] M. J. S. Mary and D. Mahalakshmi, "An extensive survey on resource allocation mechanisms in cloud computing," *Palarch's Journal of Archaeology of Egypt/Egyptology*, vol. 17, pp. 45–56, July 2020.

[3] T. R. Nair and M. Vaidehi, "Efficient resource arbitration and allocation strategies in cloud computing through virtualization," *IEEE Transaction on Cloud Computing*, vol. 3, pp. 397–401, June 2011.

[4] N. Mohan and E. Raj, "Resource allocation techniques in cloud computing-research challenges for applications," *Computational Intelligence and Communication Networks*, vol. 6, pp. 101–123, September 2021.

[5] V. Emeakaroha, I. Brandic, M. Maurer, and I. Breskovic, "SLA-aware application deployment and resource allocation in clouds," *Computer Software and Applications*, vol. 3, pp. 74–82, May 2011.

[6] R. Prodan and R. Ostermann, "A survey and taxonomy of infrastructure as a service and web hosting cloud providers," *Grid Computing*, vol. 10, pp. 45–57, June 2019.

[7] T. Schlegel, R. Kowalczyk, and Q. Vo, "Decentralized co-allocation of interrelated resources in dynamic environments," *Web Intelligence and Intelligent Agent Technology*, January 30, pp. 1123–1131, April 2008.

[8] K. Kumar, J. Feng, Y. Nimmagadda, and Y. Lu, "Resource allocation for real-time tasks using cloud computing, *Computer Communications and Networks*, vol. 3, pp. 21–30, July 2011.

[9] X. You, X. Xu, J. Wan, and J. Yu, "RAS-M: Resource allocation strategy based on market mechanism in cloud computing," *Grid Computing*, vol. 12, pp. 89–100, September 2009.

[10] J. Yan and W. Li, "Calibrating resource allocation for parallel processing of analytic tasks," *Cloud Computing*, vol. 10, pp. 135–144, October 2009.

[11] K. Mochizuki and S. Kuribayashi, "Evaluation of optimal resource allocation method for cloud computing environments with limited electric power capacity," *Network Information Systems*, vol. 18, pp. 1–12, May 2011.

[12] A. Sivadon and I. Chaisir, "Optimization of resource provisioning cost in cloud computing," *IEEE Transactions on Services Computing*, vol. 5, pp. 12–20, April 2012.

[13] S. Aarti, "A novel agent-based autonomous and service position framework," *Computer and Information Sciences*, vol. 12, pp. 109–120, May 2015.

[14] J. Devagnanam, "Issues and challenges in cloud computing and comparison of algorithms for allocating resource in cloud environment," *International Journal of Research and Analytical Reviews*, vol. 6, no. 2, pp. 104–110, June 2019.

[15] C. H. Jen, "Resource reallocation based on SLA requirement in cloud environment," *IEEE Transactions on Services Computing*, vol. 25, pp. 89–102, July 2020.

[16] Z. Zhang, H. Wang, L. Xiao, and L. Ruan, "A statistical-based resource allocation scheme in cloud," in *Proc. 2011 International Conference on Cloud and Service Computing*, August 2021, pp. 266–273.

[17] Z. Yang, M. Liu, J. Xiu, and C. Liu, "Study on cloud resource allocation strategy based on particle swarm ant colony optimization algorithm," *Cloud Computing and Intelligence Systems*, vol. 2, pp. 67–80, May 2012.

[18] X. Zheng and L. Wang, "A Pareto-based fruit fly optimization algorithm for task scheduling and resource allocation in cloud computing environment," *IEEE Transactions on Services Computing*, vol. 12, pp. 112–121, September 2016.

[19] F. Tseng, X. Wang, L. Chou, H. Chao, and V. Leung, "Dynamic resource prediction and allocation for cloud data center using the multi-objective genetic algorithm," *IEEE Systems Journal*, vol. 12, pp. 1688–1699, November 2018.

[20] M. Seyedehmehrnaz, "Simultaneous cost and QoS optimization for cloud resource allocation, *IEEE Transactions on Network and Service Management*, vol. 14, no. 3, pp. 12-21, September 2019.

[21] D. Sadashiv and R. Goudar, "Priority-based resource allocation and demand-based pricing model in peer-to-peer clouds," *Advances in Computing, Communications and Informatics*, vol. 4, pp. 45–53, June 2014.
[22] S. Goutam and A. Yadav, "Preemptable priority-based dynamic resource allocation in cloud computing with fault tolerance," *International Journal of Communication Networks*, vol. 12, no. 3, pp. 67–76, June 2015.
[23] B. Neethu and K. Babu, "Dynamic resource allocation in the market-oriented cloud using auction method," *International Journal of Micro-Electronics and Telecommunication Engineering*, 45, pp. 90–102, April 2016.
[24] N. E. Iniya and B. Ranjith, "Auction based dynamic resource allocation in cloud," *International Journal of Computer Network*, vol. 6, pp. 50–62, September 2016.
[25] D. Mehiar, H. Bechiri, and R. Ammar, "Energy-efficient resource allocation and provisioning framework for cloud data centers," *IEEE Transactions on Network and Service Management*, vol. 12, June 2015.
[26] A. Itachoudhary, M. C. G. Govil, Girdhar, and I. Singh, "Energy-efficient resource allocation approaches with optimum virtual machine migrations in cloud environment," *International Journal of Parallel Distributed and Grid Computing*, vol. 23, pp. 64–73, May 2016.
[27] Y. Chang, C. Gu, and F. Luo, "A novel energy-aware and resource-efficient virtual resource allocation strategy in IAAS cloud," *IEEE International Journal on Computer and Communication*, vol. 7, pp. 23–34, June 2016.
[28] A. Fatma, M. Sherif, and S. Radhya, "Optimum resource allocation of database in cloud computing," *Journal* of *Egyptian Informatics*, vol. 15, pp. 1–12, July 2016.
[29] A. Rajathi and D. Devagnanam, "Exploring and understanding the cloud environment with resource allocation techniques," *Journal of Science and Technology*, vol. 6, pp. 7–12, August 2021.