IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

PSO-Based Advanced Genetic Algorithm and Its Role in Cloud Computing Resource Scheduling

Sujeet Kumar, M.Tech Scholar, Department of Computer Science & Engineering, Bhabha Institute of Technology, Kanpur, India.

Akash Niras, Associate Professor, Department of Computer Science & Engineering, Bhabha Institute of Technology, Kanpur, India.

Sunny Kumar, Assistant Professor, Department of Computer Science & Engineering, Shri Ramswaroop Memorial University, Lucknow, India

Abstract—The discussion in this article expresses the computability of various resource scheduling methods and allocation policies in several dimensions. Because there are so many distinct computer systems that are available in a cloud computing environment, each of which has a varied capability, resource scheduling is a challenging and sophisticated process. Resource allocation is often scheduled for the Process, which informs the user of the resources and preferences that are accessible. Cloud computing solutions, which link numerous users to the same physical infrastructure, let users rent computer resources on demand, and charge on a pay-as-you-go basis, have witnessed a stratospheric surge in popularity in recent years. Utility computing is another name for cloud computing. These cloud computing environments provide the impression to cloud users that they have access to an infinite amount of computer resources. Because of this, cloud customers can modify how quickly they use resources to keep up with changing demand. Because cloud computing enables more users to seek more cloud services at once, there should be a clause that guarantees all resources are made effectively available to users who have made requests in order to meet their needs. However, this shouldn't be done at the expense of the cloud's resources performing poorly. Undoubtedly, the process of maximizing the resources that are divided over several virtual computers is the most challenging part of cloud computing.

Keywords—Cloud Computing, Resource Scheduling, Resource Scheduling Algorithms, Resource Management, etc.

I. INTRODUCTION

A ground-breaking breakthrough made possible by the technology known as cloud computing has significantly changed how computer services are delivered. As a direct result of the growth of the Internet and the World Wide Web, cloud computing has changed how users of information and communication technology access numerous resources. By offering resources in a dynamic and virtualized manner through the internet, it has made it possible to move the emphasis away from local or personal computation and towards computation that is centred on a datacenter. The use of cloud computing has the potential to transform it into a fifth utility that is paid for per use, much like how current utilities like water, electricity, gas, and telephone service are offered [1]. Cloud computing is a new technical development that enables the provision of computer resources like processing, storage, services, networks, and applications in an abstracted, virtualized, controlled, and dynamically demand-driven way via the internet. Or, to put it another way, cloud computing is the delivery of computing services via the internet. The resources that enable us to obtain the services aren't kept on our local network; rather, they're spread across the Internet. End consumers have access to all of this capabilities over the internet at any time and from any location in the world, just like they would with any other service. Due to the availability of a pool of virtualized resources that can be customised to match each user's unique needs, which introduces the concept of elasticity, users are spared of the burden of managing their hardware, software, storage, and

networks [2]. Due to these similarities, cloud computing and the internet are frequently contrasted in modern life.

Cloud computing is distinguished by a number of distinctive features, including virtualization, heterogeneity, metered service and pricing, elasticity, and resource pooling, even if it has its roots in previously known computing paradigms like Cluster and Grid Computing. To offer the unique characteristics that cloud computing does, it must overcome a variety of challenges, such as security and privacy, resource scheduling, scalability and fault tolerance, energy efficiency, interoperability, and many others. It is the duty of resource providers to address any issues that may develop when offering services to end users in order to ensure that customers continue to have faith in cloud computing. Customers desire services that are as cheaply, quickly, and dependably as they can get them, while service providers aim to maximise their earnings and return on investment on the other side. In order to satisfy the needs of both ends, an efficient structure for the management of resources must be in place. As one of the most challenging issues in cloud computing from both the standpoint of cloud providers and cloud consumers, resource scheduling is heavily stressed in this study.

The phrase "cloud computing" refers to a relatively new computer paradigm that makes it possible to easily do high-performance computations. Users can utilise their own services and apps whenever they need to and have the choice to just pay for the resources they really use thanks to the well-managed way that the cloud makes its services available. In order to keep expenses down while sustaining datacenters, cloud computing is concerned with remote services [3]. Infrastructure, platform, and software services are just a few of the service types that cloud computing provides. Additionally, it provides deployment models for community, hybrid, private, and public clouds. The cloud also has features like self-service, network accessibility, resource sharing across several servers, measurement services, and elastic services. [4]

Basic Cloud Services

Software as a Service (SaaS): It enables the user to independently access or use software stored on the cloud infrastructure. This programme can be accessible from a range of devices, including laptops, desktops, smartphones, and tablets, via the use of a web browser or client application. Because the CSP is in charge of delivering all necessary services, the user is not required to manage the services [5].

Platform as a Service (PaaS): Through the use of programming languages and supporting tools, users and clients can create software applications. Customers and users won't need to be concerned with license agreements, and there won't be any requirement to install any language on the personnel system. The service distributor may be in charge of managing the infrastructure that the consumer uses.

Infrastructure as a Service (IaaS): The user has access to servers, networks, physical storage, processing power, and other fundamental computing resources where they can deploy and run applications. The operating system and any other programmes regarded as system software can be included in this group of programmes. Although the user could have command over the operating system and other layers, it is not necessary for the user to exert control or management over the infrastructure [5].

Storage, communications, databases, and a host of other services are managed by cloud computing, but they all fall under the purview of the first three services mentioned. see at figure 1.

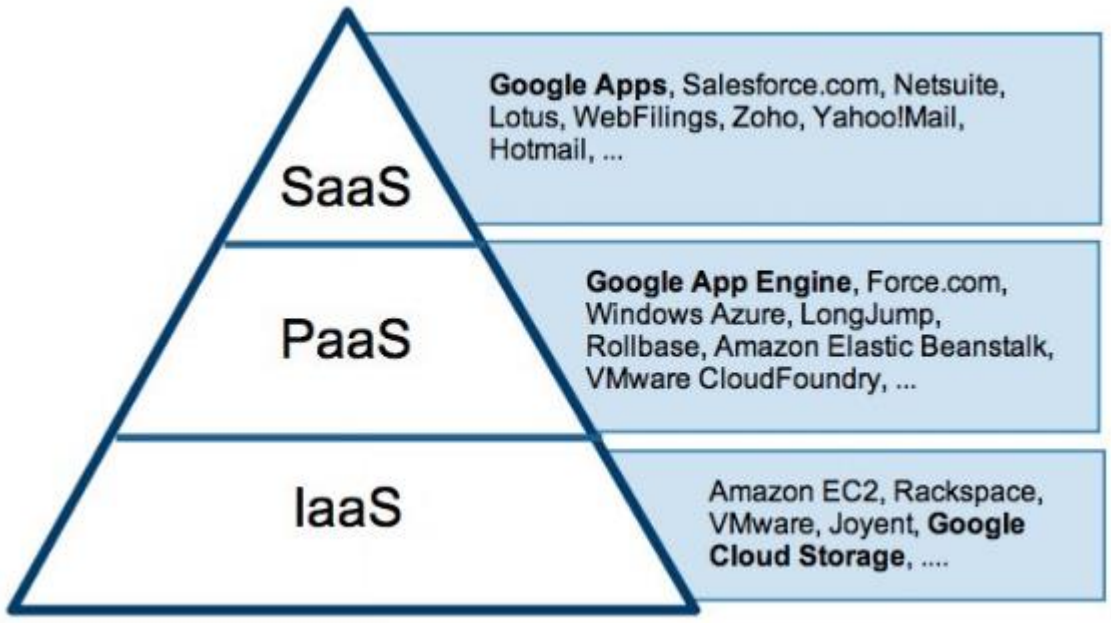

Figure 1: Basic Structure of Cloud Computing

Basic Cloud Models

Private Cloud: Access to the cloud infrastructure is available to businesses with a big customer base. It might be independently managed, run, and controlled by oneself or another service provider, or it could be joined with another, and it may set up shop in the same or different places.

An internal or corporate cloud is another name for private cloud. Organizations use it to construct and operate their own data centres, either internally or through a third party. Open source tools like Eucalyptus and Open stack can be used to deploy it.

Community Cloud: The organizational structure and the amenities may be utilized by a specific group of customers from any company that engages in the practice of sharing thoughts. It might be independently governed, run, and managed by its own service provider or another one. Additionally, it might merge and it might set up shop in the same or different places.

In order to share information between an organization and a particular community, a collection of various organizations can access systems and services through a community cloud. One or more community-based organizations, a third party, or a combination of them own, manage, and run it.

Example: Health Care community cloud

Public Cloud: Over the network, members of the general public have easy access to both the organizational structure and the facilities that it maintains. It is conceivable that it may be self-governed, operated, and managed by the self and/or another service provider, or that it could be combined with another entity. In addition, it is conceivable for it to have been installed in other locations, and it is present in the sites where the service provider has installed the services. Additionally, it is feasible for it to have been installed in similar locations.

Public cloud is open to all to store and access information via the Internet using the pay-per-usage method.

In public cloud, computing resources are managed and operated by the Cloud Service Provider (CSP).

Example: Amazon elastic compute cloud (EC2), IBM SmartCloud Enterprise, Microsoft, Google App Engine, Windows Azure Services Platform.

Hybrid cloud: The inclusion of two or more than two cloud infrastructures each have certain distinctive entities, but each also have some bonding by or proprietary technology that ensures the application and data likelihood [5].

Hybrid Cloud is a combination of the public cloud and the private cloud. We can say:

Hybrid Cloud = Public Cloud + Private Cloud

The services that are running on the public cloud can be accessible by anybody, whilst the services that are running on a private cloud can only be accessed by the users of the organization. This provides hybrid cloud computing with a level of security that is only partially safe.

Example: Google Application Suite (Gmail, Google Apps, and Google Drive), Office 365 (MS Office on the Web and One Drive), Amazon Web Services.

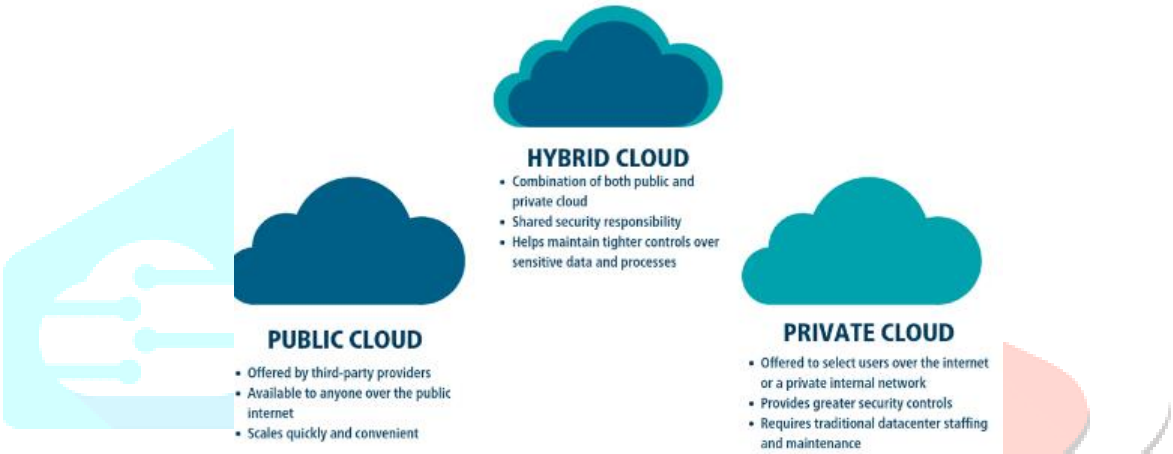

Figure 2: Types of Clouds

Table 1: The difference among all Clouds

Parameter	Public Cloud	Private Cloud	Hybrid Cloud	Community Cloud
Host	Service Provider	Enterprise (Third Party)	Enterprise (Third Party)	Community (Third Party)
Users	General Public	Selected Users	Selected Users	Community Members
Access	Internet	Internet, VPN	Internet, VPN	Internet, VPN
Owners	Service Provider	Enterprise	Enterprise	Community

Characteristics of Cloud Computing

The following is a list of the essential characteristics of cloud computing:

- On-demand self service
- Multi-tenancy and resource pooling
- Broad network access
- Rapid elasticity and scalability
- Resource pooling
- Measured and reporting service
- Automation

- Resilience
- Large Network Access
- Work from any location
- Comfortable payment structure
- Service Excellence
- Easy maintenance
- Flexibility
- Economical and Security
- Availability

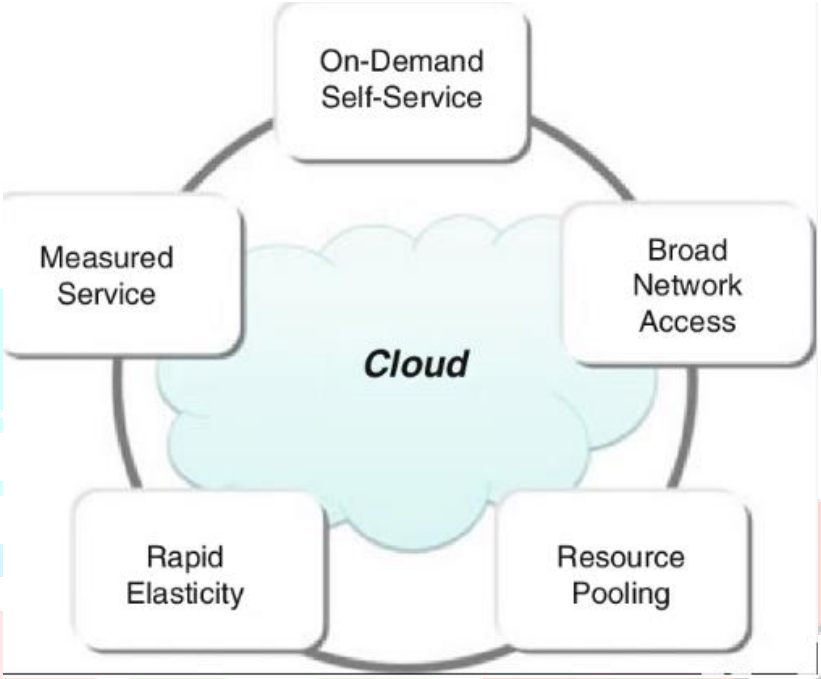

Figure 3: Cloud Computing Characteristics

II. SCHEDULING IN CLOUD COMPUTING

If something is on the schedule, it indicates that it is going to take place at a particular point in time at some point in the future. When it comes to the task of distributing available resources, distributed computing provides its users with a wide selection of scheduling algorithms from which to pick. Authentication of an appropriate kind is required in order for the distributed system to make complete use of all of the algorithms that it possesses. Utilizing the scheduling strategy to its full potential will allow for the greatest feasible rise in the throughput of the system. Working in a cloud environment makes it impossible for standard approaches to achieve the level of efficiency that is necessary for the job. The cloud computing system organizes the algorithms for work scheduling into the following categories: batch mode, sequential mode, online mode, and random mode. Online mode is the most recent addition to this list. When the information is received by the system in batch or sequential mode, all of the resources are already ordered in a chain and have united to form a set. This is true regardless of whether the information is received in batch or sequential mode. In this scenario, the algorithm will become operational at the time intervals that have been established in advance. The algorithms Fcfs, RR, min-min, and max-min are some examples of those that can be used in batch mode. Other algorithms that can be used in batch mode include max-min.

Since cloud computing is an online technology that is also heterogeneous, the speed of the processors can fluctuate within a shorter amount of time. This is because the time it takes to access the cloud can vary. Because of this, the method of scheduling that is done online is more efficient and is suitable for use with the cloud.

The assignment of acceptable resources to cloud workloads is decided by the quality of service requirements of cloud applications. The task of allocating resources in the cloud involves a number of problems, and these challenges must be overcome in order to be successful. There are issues involved in the process of allocating resources in a cloud environment due to the heterogeneity, uncertainty, and dispersion of the available resources. Using the policies that are presently in place for resource distribution will not be sufficient to tackle these difficulties.

The management of resources is an overarching activity that includes several stages of workloads and resources, beginning with the submission of workloads and ending with their execution.

The administration of resources in the cloud consists of two stages:

(i) Resource provisioning, and (ii) Resource scheduling.

The stage known as resource provisioning is the process that determines which resources are sufficient to handle a particular workload based on the quality of service requirements that are outlined by cloud consumers. On the other hand, resource scheduling refers to the process of mapping out and carrying out the execution of cloud consumer workloads in accordance with the specified resources that are obtained through the process of resource provisioning (see Figure 4). In the beginning, the cloud consumer will put in a request for the execution of their project by providing workload specifics in the form of a request. On the basis of these data, the broker (resource provisionor) locates the resource (or resources) that are most appropriate for a particular workload and determines whether or not it is possible to supply resources according to the quality of service standards [6]. When the broker determines that the process of providing resources has been properly concluded, it will forward requests to the resource scheduler so that those requests can be scheduled. The broker is also responsible for the release of extra resources into the resource pool, the preservation of information regarding provided resources, and the monitoring of performance in order to add or remove resources. These responsibilities go hand in hand with the broker's responsibility to act as an intermediary. The first stage, which consists of the provisioning of resources, is followed by the second stage, which is scheduling those resources.

The resource queue is used to hold all of the resources that have been provisioned, while the resource pool is used to hold all of the resources that are still available [7]. In the workload queue, items that have already been turned in for processing are attended to. At this stage, it is the responsibility of the scheduling agent to map the available resources to the workload(s) that have been specified, to carry out the workload(s) that have been specified, and then to return the resources to the resources pool once the workload(s) have been successfully finished. When quality of service criteria are taken into account, it might be difficult to overcome the difficulty of scheduling resources to handle acceptable levels of work. When trying to successfully plan the scheduling of resources, it is absolutely necessary to take into account the quality of service needs [8].

It is necessary to investigate the difficulties that can arise in the scheduling of resources in order to carry out tasks without compromising other quality of service standards.

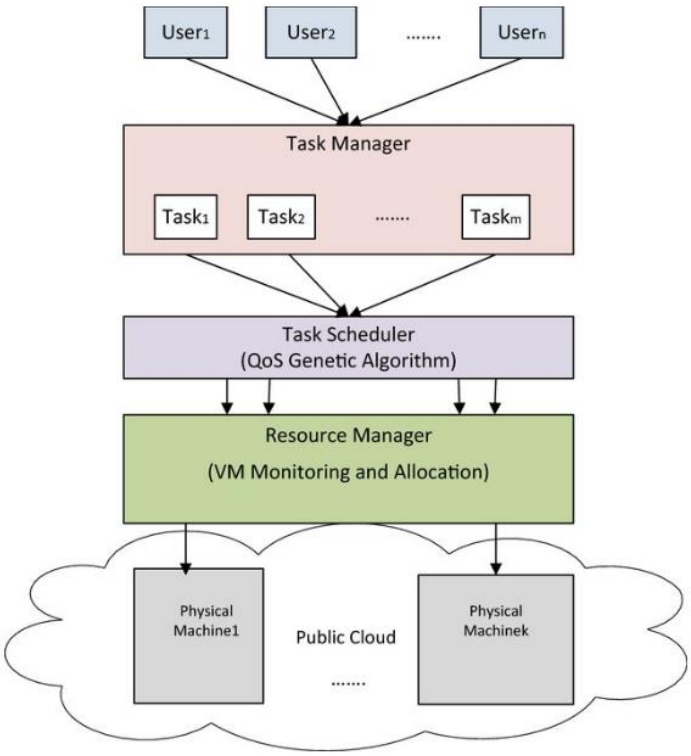

Figure 4: Resource Scheduling in Cloud

In cloud computing, resource scheduling has been an important major topic of research due to the lengthy execution time and high cost of resources. Different kinds of resource scheduling algorithms, or RSAs, each have their own unique sets of criteria and parameters for the scheduling of available resources. The provision of resources is the initial level of resource management, which is something that has already been discussed in our prior review work [10]. This corpus of scientific study will focus its attention on the resource scheduling process, which is the second step in the process of managing available resources. The expenses associated with execution, as well as execution time, energy consumption, and other quality of service concerns like as dependability, security, availability, and scalability, can all be lowered when resources are scheduled correctly. The cloud user and the cloud provider are two separate but equally important participants in a cloud environment. The user of cloud services is

responsible for submitting workloads, while the cloud provider is responsible for supplying the resources necessary for the execution of workloads. Both the provider and the consumer have different requirements: the provider wants to maximize utilization of resources while earning as much money as possible with the fewest investments possible, and the consumer wants to complete their workload(s) with the fewest resources and in the shortest amount of time possible. Both of these goals are possible, but the provider wants to maximize utilization of resources while earning as much money as possible with the fewest investments possible. The execution of several workloads on a single resource, on the other hand, will result in interference between those workloads, which will lead to substandard performance and a loss in customer satisfaction. Those two outcomes will be negative for business. Requests that would result in an uncertain environment are typically denied by service providers because providers aim to maintain a high level of service quality [8]. The suppliers take into account the volatility of some resources while arranging and carrying out the jobs. It is more challenging to schedule the resources due to the fact that neither the people who utilise the resources nor the people who provide them are willing to share information with one another. The dispersion of resources, uncertainty over their availability, and the heterogeneity of those resources are three challenges that are connected with resource scheduling [7]. These challenges are insurmountable for the resource scheduling algorithms (RSAs) that are typically used. As a consequence of this, it is essential to guarantee that workloads in the cloud are carried out in an efficient manner by paying attention to the features of the environment provided by the cloud.

Figure 4 demonstrates how the information required to manage the workload may be received from the cloud consumer through the use of the Cloud Workload Management Portal. The queue is the location where the entire number of cloud workloads that a user has submitted to be handled. The information that is provided by cloud consumers is used to determine how the resources that are necessary for the execution of cloud workloads are distributed across those workloads. If the required resources are not already present in the resource pool, the resource provisioner will not provide the required resources to the workload in order for it to be able to execute it in the cloud environment. This is because the workload requires the required resources in order to complete its tasks. In the event that the required resources are not readily available in accordance with the QoS requirement, the Workload Management System (WMS) will request that the workload be resubmitted with amended QoS criteria in the form of a service level agreement [10]. After the provisioning of resources has been properly finished, the workloads are then forwarded to the resource scheduler to be assigned. After then, the resource scheduler will inquire as to whether the workload has been submitted for the resources that have been allotted. After this is complete, the findings will be sent back to WMS via the resource scheduler, and the cloud workload will be provided with information regarding the resources. By doing an analysis of the data supplied by the cloud consumer, the policy selector is the component that establishes which scheduling strategy is suitable for the workload in question [7].

Cloud environment and a scheduler that, depending on the choice that was made by the policy selector, will carry out a number of different scheduling rules in order to schedule tasks. The scheduling rules assign each workload a certain percentage of the total available resources. This percentage is decided by the cloud. The resource scheduler is responsible for allocating the new cloud workloads to the appropriate slots in the cloud, taking into account the characteristics of each individual task. First things first, get cloud workloads scheduled, and then, depending on the scheduling policies, find cloud workloads that have been mapped as effectively as feasible, as well as appropriate and accessible resources. The responsibilities that need to be carried out are sent to the dispatcher so that they can be carried out. Only in the case that the workloads will be carried out in compliance with the QoS requirements mentioned in the SLA will they be distributed to be finished. The resource monitor is utilized in order to investigate the current state of the scheduling of resources, such as determining whether or not the required number of resources is provided. The information that is contained within the QoS monitor pertains to the QoS parameters, and its function is to analyze whether or not all of the workloads are working within the range that was defined for them. The information that is contained within the QoS monitor can be found by clicking on the icon that is located on the bottom left of the screen. Assuming that the deadline is a quality of service criterion, it is the responsibility of the quality of service monitor to evaluate whether or not the workloads have been finished before to the deadline that was set. A breach of the service level agreement (SLA) occurs when the work is carried out later than the desired deadline.

Need of Resource Scheduling

Finding the relevant resources is the first step in the scheduling process for resources. This is done so that the appropriate workloads can be scheduled at the appropriate times. Increasing the effectiveness with which the available resources are put to use is the second objective of the scheduling of those resources. In other words, the quantity of resources needed for a job should be maintained to a minimum in order to maintain a particular level of service quality, limit the amount of time needed to finish a workload, or maximize the amount of work that can be done. These goals can be accomplished by keeping the number of resources needed for a workload at a minimum. In order to obtain more efficient scheduling of resources, it is important to carry out the most accurate resource workload mapping that can be performed. The second objective of resource scheduling is to determine which workload is the most acceptable and appropriate one to support the scheduling of many workloads in order to be capable of satisfying a wide variety of QoS standards for cloud workloads [11]. These QoS requirements include CPU utilization, availability, reliability, and security, among other things. The second objective of resource scheduling is to determine which workload is sufficient and appropriate to support the scheduling of various workloads. As a result, the amount of time necessary to finish each individual assignment is factored into the scheduling of resources. However, the type of workload, which can either be heterogeneous or homogeneous depending on the degree to which it varies in terms of its quality of service (QoS) requirements, is the most critical component in determining total performance. Homogeneous workloads have fewer variations in their QoS requirements than heterogeneous workloads.

Resource Management

Cloud computing enables the provisioning and scheduling of available resources, as well as the delivery of cloud services that are both guaranteed and reliable, all on the basis of a pay-per-use pricing model. It is quite challenging to supply the service in an efficient manner because demand from different customers of cloud computing might be highly unpredictable. We have compiled a list of the many approaches to arranging our resources based on the research that has been done previously. As can be seen in Figure 5, the process of managing resources in the cloud is broken down into two stages: resource provisioning and resource scheduling.

A centralized agent known as Cloud Resource Manager is in charge of managing the resource allocation and utilization process (CRM). CRM is responsible for managing all of the cloud workloads and resources, as well as mapping the workloads and resources in an effective manner. As can be seen in Figure 6, CRM is comprised of a variety of different entities as well as interfaces. The scaling listener is what is utilized to map the workloads onto the appropriate resources based on the QoS requirements that the user has stated. The cloud consumer will often submit workloads together with their QoS needs to the cloud provider for execution. This is done in the context of resource management. The resources are provided from a set of resources (r1, r2, r3,..., rn) for user workloads (w1, w2, w3,..., wm), with the goal of achieving maximum resource utilization and customer satisfaction. These requirements are based on quality of service criteria. When working in a cloud environment, it is important to have an efficient management of resources so that income can be maximised and user satisfaction can be improved.

In section 1, the cloud Resource Manager's units are broken down into detail. The fundamental classification of cloud resources, as shown in Figure 7, is an essential component of the comprehensive resource management process.

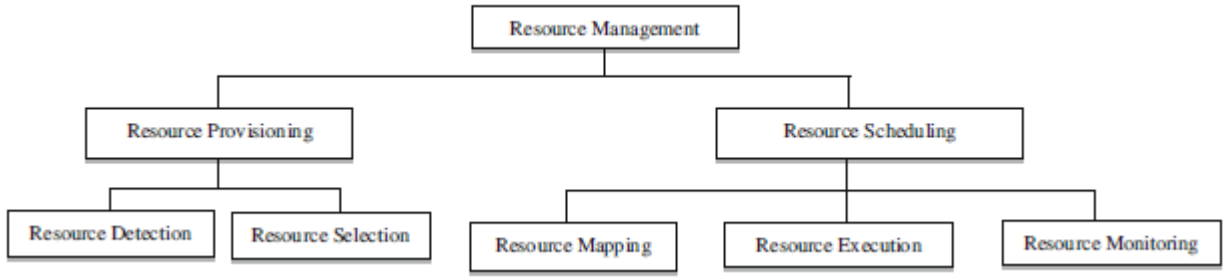

Figure 5: Resource Management Taxonomy

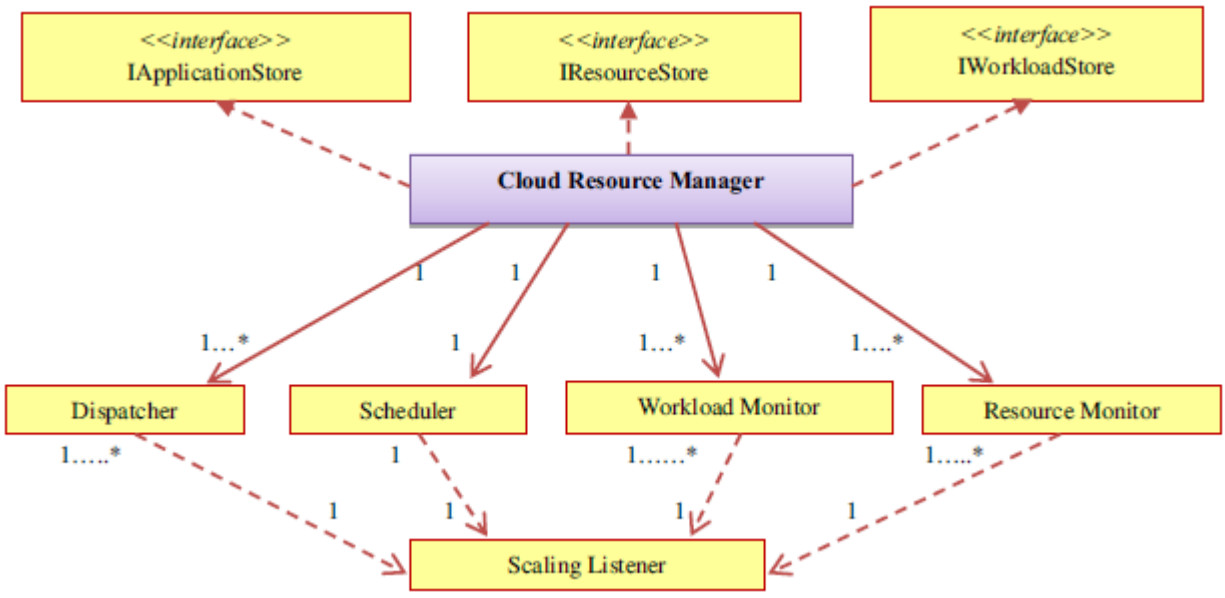

Figure 6: Interfaces and Entities Connected to Cloud Resource Managers

Figure 7: Cloud Resource Classification

Resource Provisioning

The provisioning of resources is a particularly challenging endeavour because the requisite resources are not always easily accessible [6]. [10] The provisioning of resources to workloads is carried out in a manner that complies with the QoS criteria of the cloud applications. According to the findings of the research that was released, one of the challenges associated with cloud resource provisioning is the minimization of execution time. Because assigning the appropriate resources to the appropriate workloads is a difficult operation based on QoS criteria, finding the optimal workload-resource pairs is a major research issue that needs to be addressed in the cloud.

The fundamental resource scheduling and provisioning (Phase-2) model in the cloud is depicted here by Figure 8. Figure 8 depicts the first phase, which involves the cloud consumer sending workload information for the purpose of workload analysis. This information may include QoS parameters, SLA, and workload description. When a job is assigned to a resource provisioning agent (RPA), that agent will query the Resource Information Centre (RIC), which is a database that stores information on all of the resources that are part of the resource pool, and will obtain the provisioning result based on the workload requirements that the user has specified. RPA is responsible for communicating the result of the resource provisioning process to the cloud consumer. If the required resources can only be accessed through the resource pool, then it makes the requested resources available to the workload so that it can be carried out within the context of cloud computing.

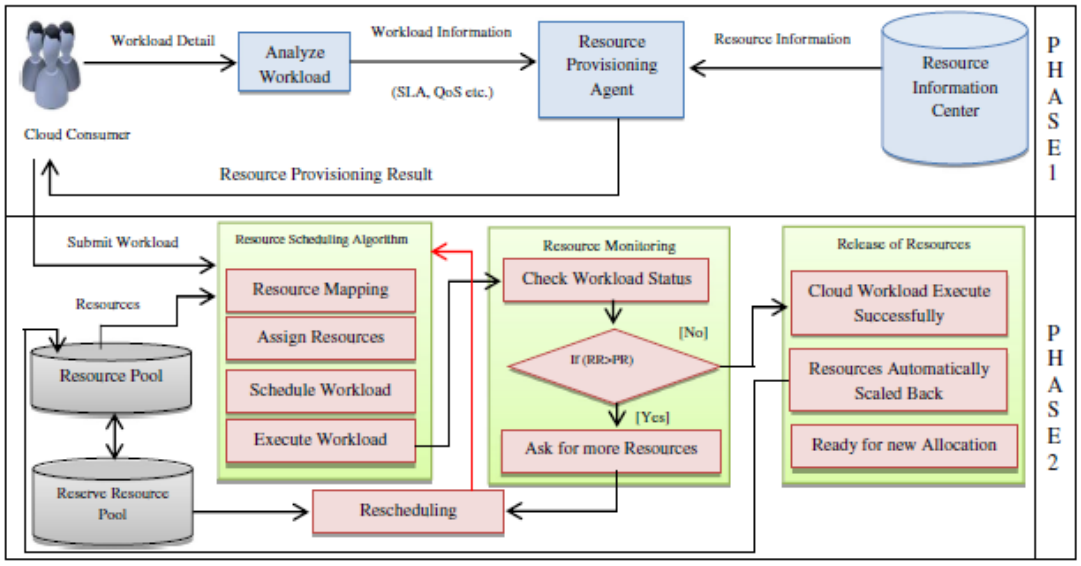

Figure 8: Resource provisioning and resource scheduling in Cloud

Resource Scheduling

Dispersion, unpredictability, and resource heterogeneity are scheduling challenges that cannot be overcome in a cloud environment [2]. Therefore, it is essential to boost the effectiveness of cloud services and cloud-based applications by taking care of certain aspects of the cloud environment. The three elements of resource scheduling are resource mapping, resource execution, and resource monitoring. Resource scheduling is done in Phase 2 after resource supply, as shown in Figure 8.

The cloud consumer first submits the workload for execution. Then, based on the QoS standards that the cloud consumer has established in terms of SLA, workloads are mapped to the right resources in order to maximise QoS parameters. Each cloud customer makes the most of the cloud services that are offered to them by allocating resources according to QoS parameters like throughput, CPU, and memory utilisation, among others. Resource execution's goal is to quickly match the appropriate resources with the appropriate workloads so that applications may utilise the resources effectively. As shown in Figure 8, while a certain cloud workload is operating, the monitoring agent will look at the active workload. If the value of the provided resources (PRs) is more than the value of the requested resources (RRs), it will request more resources. The reserve resource pool provides the appropriate resources to meet the necessary amount of resources for the workload's successful execution by rescheduling. As soon as cloud workloads are successfully finished, the scheduler is ready to start processing new cloud workloads and the free resources are returned to the resource pool. The best strategy to improve performance is to monitor how computer resources are being used efficiently. In order to evaluate resource execution performance, a detailed, intelligent monitoring agent is needed.

III. RESOURCE SCHEDULING ALGORITHMS

The two basic components of resource allocation in cloud computing are static and dynamic resource scheduling, as well as related processes including different kinds of resource scheduling, resource scheduling algorithms, and their evolution. It demonstrates a crucial quality in resource use effectiveness. The most crucial QoS criteria for any resource scheduling system are cost, time, and energy. The Resource Scheduling Algorithm (RSA) is crucial for planning and allocating the best resources to workloads. The algorithms handle the scheduling of workloads to resources in order to guarantee QoS to the cloud workload in accordance with user expectations. There are times when RSAs exhibit dynamic behaviour, scheduling resources following resource supply. These methods, sometimes known as dynamic RSAs, are thought to be more effective than static resource scheduling. Another assumption is that RSAs should be created to prevent resource under- and over-utilization. Figure 9 displays various resource scheduling strategies.

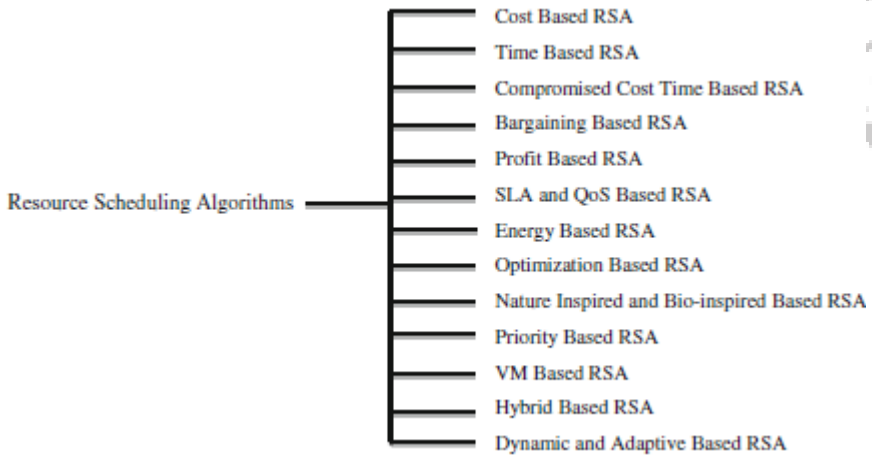

Figure 9: Resource Scheduling Algorithm in Cloud

IV. RESOURCE SCHEDULING TOOLS

CloudSim: It is an extensible simulation toolkit which provides simulation, and experimentation of infrastructures and application environments. Due to workloads, models, resources, applications etc. the existing simulators are not able to execute algorithms effectively but this toolkit overcome this limitation. Through this toolkit, scheduling algorithm can be implemented by extending java classes.

CloudAnalyst: It extends the functionalities of CloudSim to evaluate the behavior of large scaled Internet application and also allows variations in parameters to by performing simulations in repeated manner.

EMUSIM: To evaluate the behavior of service, Automated Emulation Framework (AEF) based EMUSIM is used for emulation in Cloud.

SPECI: SPECI (Simulation Program for Elastic Cloud Infrastructures) is group of two packages: components for experiment execution and data center layout and topology used to examine the behavior of large datacenters under design and size policy.

GroundSim: IaaS based this toolkit is used to detect the events by providing one simulation thread for scientific applications. Real environment can be realized by the integration of GroudSim into the ASKALON.

GreenCloud: It is an extension of CloudSim toolkit to test the results of energy efficient resource scheduling algorithms by calculating energy consumption of communication links, computing servers and network switches.

NetworkCloudSim: It is an extension of CloudSim toolkit to test the behavior of HPC applications and workflows in real cloud environment.

V. CLOUD RESOURCE SCHEDULING: OPEN ISSUES AND CHALLENGES

Even though pay-per-use cloud computing has made it simple to set up scalable computer infrastructures, further advancement needs to be done. However, there are still numerous issues and challenges in this field that need to be resolved. Based on recent research, we have identified a number of unresolved issues in the area of cloud resource scheduling. On the basis of these open themes, resource allocation strategies have been applied to further categorise research difficulties [10]. The literature has identified the following outstanding challenges and issues with resource scheduling in cloud computing [12, 13]:

Resource Scheduling: Dispersion, uncertainty, and heterogeneity of resources are issues with resource scheduling that the cloud environment's standard resource management systems cannot address. So, by taking care of these characteristics of the cloud environment, it is necessary to make cloud services and cloud-oriented applications efficient. Resource scheduling's goal is to assign the correct resources to the right workloads at the right time so those applications may use the resources efficiently.

Quality of Service (QoS): The necessary number of resources are provisioned by the service provider in order to meet the QoS requirements of the cloud service. Based on these QoS standards, SLAs are created, and regular SLA violations are found, which further determines the sanction or compensation in the event of a SLA violation. To minimize or prevent SLA violations, the service provider must dynamically provision a suitable number of resources.

Service Level Agreements (SLAs): It is necessary to have autonomic cloud infrastructures in order to limit the amount of contact that cloud consumers have with the computing environment and to fulfil the QoS requirements that are stated by cloud users in terms of service level agreements (SLA). As a result, an efficient method that can detect SLA violations in advance is a research problem that can prevent performance degradation.

Self-management Service: In this case, a cloud provider's goal is to allocate and release cloud resources in order to meet its SLOs (Service Level Objectives), which will lead to a decrease in its deployment fee. The following steps are often included in these techniques: In order to meet QoS requirements, it is necessary to (i) build an application performance model that predicts the number of application instances needed to handle requests at each individual level; (ii) periodically forecast upcoming demand and define resource requirements using the performance model; and (iii) automatically assign resources using the forecast resource requirements [14-15]. On occasion, the proactive approach allocates resources before they are actually needed by leveraging anticipated levels of demand. Even before monthly demand estimates are available, the reactive approach considers minute fluctuations in demand.

Energy Management: Improving the energy efficiency of cloud computing is one of the most pressing challenges facing the industry. It has been determined that the cost of providing data centres with power and refrigeration amounts for 53 percent of the total operational expenditures made by these facilities. In 2006, data centres in the United States utilized more than 1.5% of the total energy produced that year, and it is anticipated that this proportion will expand by 18% year going forward. As a result,

companies that provide infrastructure are coming under intense pressure to reduce their overall energy consumption. The goal is not only to reduce the cost of electricity in data centres, but also to comply with regulations set forth by the government and with environmental standards. The scheduling of tasks with an eye towards energy efficiency and the consolidation of server resources are two more ways to reduce power usage by turning off unused equipment. Finding a balance that is satisfactory between the reduction of energy consumption and the enhancement of application performance is one of the primary challenges posed by existing methods.

Data Security: Cloud computing raises a number of questions, one of which is data security. Cloud service providers, on the other hand, do not typically have access to the physical data security system of data centres; therefore, cloud service providers are need to rely on infrastructure service providers in order to achieve complete data security. Even for a virtual private cloud, the cloud provider can only identify the security setting in a general sense, and they have no idea whether or not it has been fully applied. Creating trust mechanisms at each architectural layer of the cloud is an extremely risky endeavour. At first, the hardware layer needs to have a reliable platform section employing a hardware stable platform. In addition to this, the virtualization platform needs to maintain its privacy by making use of secure VM observers. The migration of virtual machines (VMs) should only be permitted if both the sending and receiving servers are secure.

VI. BENEFITS OF CLOUD RESOURCE SCHEDULING

In the available research, we uncovered a variety of benefits associated with cloud resource scheduling; some of the most important of them are as follows [15-17]:

- Efficiently arranging cloud resources can simultaneously boost the robustness of a workflow and reduce its makespan.
- Decrease the amount of time that cloud workloads spend executing and computing in order to effectively schedule cloud resources.
- Improved resource usage in accordance with various priority needs, while avoiding both under loading and overloading of available resources.
- There will be no delays in the scheduling, and the risk of a resource failure will be reduced as a result of the efficient allocation of resources.
- A decrease in power usage that does not violate the service level agreement (SLA) in effective cloud resource scheduling.
- Reduce the percentage of missed user deadlines that are caused by resource scheduling that occurs after resource provisioning.
- With efficient scheduling of cloud resources, the amount of time spent waiting for pending tasks is reduced.

VII. CONCLUSION

In this study, the findings have been examined using a range of different methods, including resource classification and resource schedule development. A detailed classification of resource scheduling algorithms and their subtypes, a comparison of resource scheduling algorithms, resource scheduling aspects, resource distribution policies, and resource scheduling tools have also been provided in accordance with the research questions. Recent research has shown that resource provisioning strategies can be used to make resource scheduling algorithms more efficient. It might also be challenging to find the best workload-to-resource mapping if you do not have an effective resource provisioning approach. Gaining a thorough understanding of the QoS requirements imposed by the task would be more advantageous than concentrating on analysing workload and resource requirements. This would make it possible to allocate resources more effectively.

It is vital to first uncover the advancements achieved in cloud research in order to locate the advanced resource scheduling research. We summarised the previously published content and presented it as a systematic sequence of resource scheduling. With the aim of identifying research needs for the purpose of future research, this study presents an extensive and complete literature evaluation of resource allocation in the subject of cloud computing in general, and cloud resource scheduling in particular.

References

[1] Z.Liu, "Research on Computer Network Technology Based on Cloud Computing," in Proceedings of the 9th International Symposium on Linear Drives for Industry Applications, Springer, 2014.

[2] Peter Mell and T. G. "The NIST Definition of Cloud Computing". National Institute of Standards Technology Special **Publication. pp 80-145. 2011.**

[3] J.Lee and S. Kim, "Software Approaches to Assuring High Scalability in Cloud Computing," in IEEE International Conference on E-Business Engineering, 2010.

[4] Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R. H., Konwinski, A., & Zaharia, M. (2009). Above the clouds: A berkeley view of cloud computing (Vol. 4, pp. 506-522). Technical Report UCB/EECS-2009-28, EECS Department, University of California, Berkeley.

[5] B.Furht and A.Escalante, in Hand Book of Cloud Computing, Springer, 2010.

[6] T.Chieu, A.Mohindra and A. Karve, "Scalability and Performance of Web Applications in a Compute Cloud," in e-Business Engineering (ICEBE), 2011

[7] Singh, S., Chana, I.: QRSF: QoS-aware resource scheduling framework in cloud computing. J. Supercomput. 71(1), 241–292 (2015)

[8] Chana, I., Singh, S.: Quality of service and service level agreements for cloud environments: issues and challenges. In: Cloud Computing-Challenges, Limitations and R&D Solutions, pp. 51–72. Springer International Publishing (2014)

[9] L.Vaquero, L.Rodero-Merino and R. Buyya, "Dynamically Scaling Applications in the Cloud," ACM SIGCOMM Computer Communication Review, pp. 45-52, January 2011

[10] Singh, S., Chana, I.: Cloud resource provisioning: survey, status and future research directions. Knowl. Inf. Syst. 44, 1–50 (2015).

[11] Singh, S., Chana, I.: QoS-aware autonomic resource management in cloud computing: a systematic review. ACM Comput. Surv. 48(3), 39 (2015).

[12] C.Fehling, F.Leymann, R. Retter, W. Schupeck and P. Arbitter, Cloud Computing Patterns, Springer, 2014

[13] Singh, S., Chana, I., Buyya, R.: Building and Offering Aneka-based Agriculture as a Cloud and Big Data Service. Big Data: Principles and Paradigms, pp. 1–25. Elsevier (2016)

[14] Pascual, J.A., Lorido-Botr´an, T., Miguel-Alonso, J., Lozano, J.A.: Towards a greener cloud infrastructure management using optimized placement policies. J. Grid Comput. 13(3), 375–389 (2015)

[15] J.McCabe, Network Analysis,Architecture, and Design, Elsevier, 2007.

[16] F.Galán, A. Sampaio, L. Rodero-Merino, I. Loy, V. Gil and L. Vaquero, "Service specification in cloud environments based on extensions to open standards," in Proceedings of the Fourth International ICST Conference on COMmunication System softWAre and middlewaRE, 2009.

[17] A.Younge, G.Laszewski, L. Wang, S. Alarcon and W. Carithers, "Efficient Resource Management for Cloud," 2010.