Resource Allocation Techniques for Improving QoS in Cloud Computing

GUGULOTH LACHIRAM[1], B SHANKAR NAIK[2]

[1]Research scholar, MUIT,Lucknow, Uttar Pradesh,India.
[2]Associate Professor,CSE Department,CMR Technial campus,Nedchal,Hyderabad, Telangana, India.

***Abstract*: Cloud computing has become a crucial platform for processing data and executing computationally intensive applications on a pay-per-use basis. Resource allocation is the mechanism by which cloud providers provide resources to users based on their adaptable needs. Service Level Agreement (SLA) between service providers and customers has grown more critical as data continues to grow exponentially. This process of resource distribution gets increasingly difficult as a result of limited resources and rising customer demands. In light of the uniqueness of the models and approaches, the primary objective of resource allocation is to minimize the related overhead costs. The thematic taxonomy of resource allocation dimensions is examined, along with the articles that fall within each category. Focusing on resource and request validation, we propose the Multi-Agent-based Dynamic Resource Allocation (MADRA) strategy, a multistage framework utilizing the QoS-based Resource Allocation (QRA) algorithm, and the Artificial Immune System Directed Acyclic Graph (AIS-DAG) model for optimal resource allocation use to improve QoS and scalability.**

***Keywords*: Cloud computing, AIS-DAG, Resource allocation, Resource scheduling, QoS.**

1. INTRODUCTION

The modern technology of cloud computing is based on the service-oriented architecture to provide infrastructure, platform, and software as a service. In cloud computing, resource allocation involves assigning processing tasks to a pool of resources in the cloud infrastructure, which consists of multiple computers. The purpose of this cutting-edge technology is to provide clients with pay-per-use payment services. As a new technology, cloud computing faces difficult challenges that necessitate a clear depiction of activities and relationships, with the end goal of promoting the strategic development and implementation of cloud computing in mind. Technically, it is a combination of server virtualization technology, other resources, and various technologies [1].

In cloud computing, resource allocation involves the scheduling and provisioning of resources with consideration of the available infrastructure, service level agreements, cost, and energy factors. For example, a cloud service provider manages resources based on the on-demand pricing model, all the while ensuring excellent Quality of Service (QoS) and user satisfaction [2]. Similarly, the resources must be assigned so that each application receives the necessary resources without exceeding the cloud environment's limit. Similarly, resource allocation is responsible for addressing the issue of underutilized applications by enabling service providers to allocate resources to each module [3].

Cloud computing provides low-cost, high-quality services to consumers [4]. While data centers provide an abundance of storage resources and distributed computing models ready to assist with request resource allocation, this leads to suboptimal resource allocation. Another issue faced by large data centers is energy consumption. It has been observed that energy consumption consumes over 20% of large data centers. Reducing energy consumption can save resource providers a substantial amount of energy and money [5]. Elastically utilizing hardware resources and shutting down unused servers is the easiest and most effective method for achieving this goal. This, however, requires careful planning so that data centers do not run out of resources as queries occur.

2. RESOURCE ALLOCATION IN CLOUD COMPUTING

Allocation of resources is the process by which the appropriate resources are assigned to the activities needed by the consumer for these tasks to be completed effectively. In cloud computing, this entails assigning a virtual machine to a job that may have its time constraints. The feasible manner in which these tasks may be assigned to and managed by virtual machines is an additional form of resource allocation strategy in the cloud [6]. Simply, it involves describing when a computational activity should begin or conclude according to a certain condition.

1. Resources assigned.
2. Time taken.
3. Predecessor actions.
4. Predecessor relationships.

In addition, cloud computing resource allocation involves resource disclosure, choice, provisioning, application design, and management. It entails determining when, what, where, and how much resources should be provided to the customer, as seen in Figure 1.

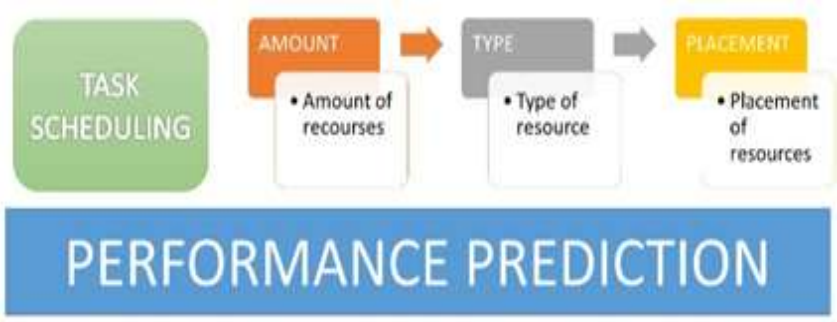

Figure 1. Cloud resource allocation basic elements

Figure 2 shows that the allocation of resources is done in the following steps generally:

1. The consumer will submit the request to the resource allocator.
2. The request will be added to the queue list.
3. The resource allocator informs the allocation unit about the request.
4. The allocation unit asks for the requested resources from the Infrastructure as a Service (IaaS).
5. If the resources are available, then IaaS respond positively.
6. The allocation unit creates a Virtual Machine (VM) from the VM pool according to the request.
7. The resource allocator is informed after creating a VM.
8. Requests are de-queued.
9. resources are allocated.

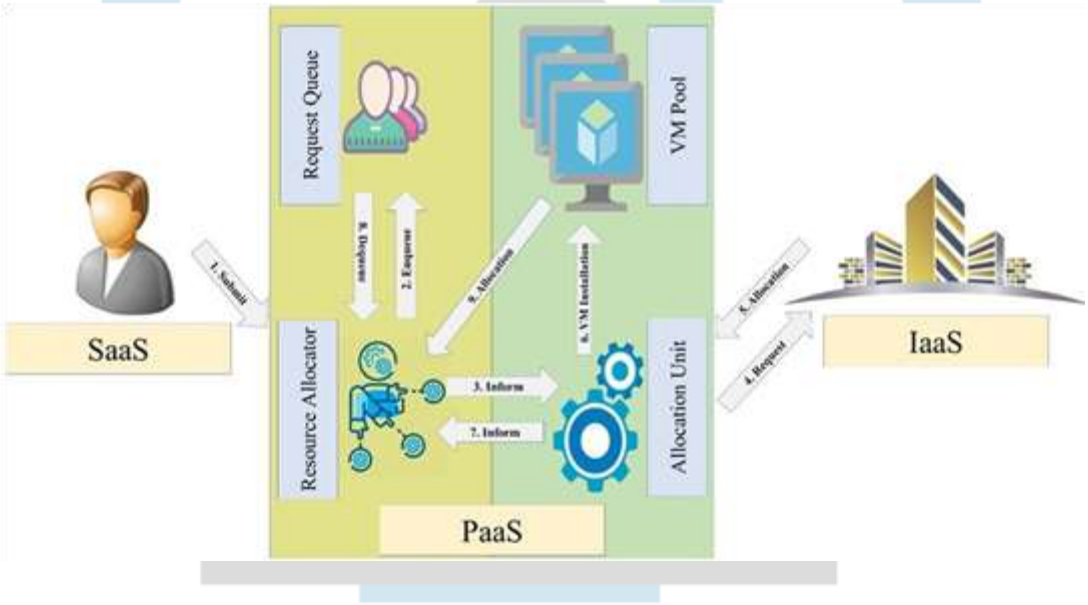

Figure 2. The basic flow of resource allocation in cloud computing.

Both cloud service providers and cloud service users do not exchange information, which further complicates the process of assigning resources. For instance, a cloud service provider will not provide the amount and kind of resources he has, since corporate departments prohibit the sharing of such information. On the other hand, cloud service customers do not reveal the application type and workload to outsiders, including service providers. To optimize the resource allocation for the consumer's request, cloud service users do not make their requests public since they do not exactly know what is available. In addition, cloud service providers are unable to allocate resources optimally to cloud service consumers' applications since there are no or few insights into their workload patterns [7].

Allocation of resources to users based on their application use patterns is one of the most significant issues of cloud computing. The data center will execute these unanticipated requests across the internet. We have identified the following resource allocation difficulties in cloud computing.

From the perspective of a cloud service provider, it is very difficult to estimate user and application demand. Likewise, from the cloud service consumer's perspective, the task must be done on time. Therefore, owing to constrained resources, cloud-supporting resource allocation algorithms must be accessible and efficient.

- The capacity of physical computers should be sufficient to meet the resource requirements of all virtual machines.

- Applications operating on the VM, as well as consumers, need networking services with efficient QoS to ensure the efficient delivery of their application data.
- Auction-based resource allocation has the difficulty of creating a method that proves acceptable, i.e., the distribution of resources is efficient and its price is advantageous for both the cloud service provider and the cloud service customer.
- Service Level Agreement (SLA) breaches must be reduced while optimizing resource usage since, in the majority of circumstances, quality of service influences cost-cutting resource allocation procedures.

Conclusions can be drawn from the preceding discussion that one must take into account the qualities and attributes of both components of cloud computing to provide proficient cloud services and cloud-based applications, i.e., assigning adequate resources to a suitable application at an appropriate time so that the application can successfully utilize the resources [8].

3. RESOURCE ALLOCATION TECHNIQUES

The resource allocation approaches use a variety of methodologies for the productive utilization of resources to satisfy consumer needs. In cloud computing, the strategies for resource allocation may be characterized as follows:

1. **Strategic:** satisfying the consumer's ever-changing demands,
2. **Target resources:** focusing mainly on requested resources,
3. **Auction:** bidding for the resources,
4. **Optimization:** optimizing the resources,
5. **Scheduling:** prioritizing the task for better performance
6. **Power:** better resource allocation with less power consumption,

This categorization is further subdivided into the following subheadings. The previously indicated techniques are then evaluated using the criteria shown in Figure 3. When creating resource allocation techniques, both cloud service provider (cost, resource use, energy, workload, SLA, QoS) and cloud service consumer (execution time, response time, user satisfaction, SLA, QoS) perspectives should be considered [9].

- Cost is one of the most essential characteristics for cloud service providers, as it ultimately decides whether the cost of delivering various cloud services is high or cheap. It is important to note that in this article, the cost parameter only applies to the service provider and not the service user.
- Resource Use: All cloud service providers strive to optimize their resource utilization so that no resources are idle. It is essential to note that smart resource allocation is beneficial for environmental protection and decreases the total cost of data centers.
- Power: Power is another crucial aspect of cloud computing resource allocation. As the energy crisis worsens, limiting power and energy consumption to make cloud services ecologically sustainable has become a top priority.
- Workload: Workload often indicates the system's capacity to manage and process the assigned work. A system's workload should be sufficient for it to perform tasks effectively in the cloud environment. This parameter will define the effort associated with the empirical resource allocation approach setup.
- Execution Time: Both the cloud service provider and the cloud service customer need the shortest possible job execution time. However, executing numerous workloads on the same resource will cause interference between these tasks, resulting in poor performance.
- Response Time: The time it takes the system to respond to a request. From the perspective of cloud service consumers, it should be as cheap as feasible. Response time is a crucial indicator of the system's performance. Low reaction time is essential for computing success.
- User Satisfaction: The degree to which a user is a content with the cloud service provider. Every cloud service provider seeks to fully delight his customers. Through efficient resource allocation in cloud computing, income and user pleasure may be maximized.

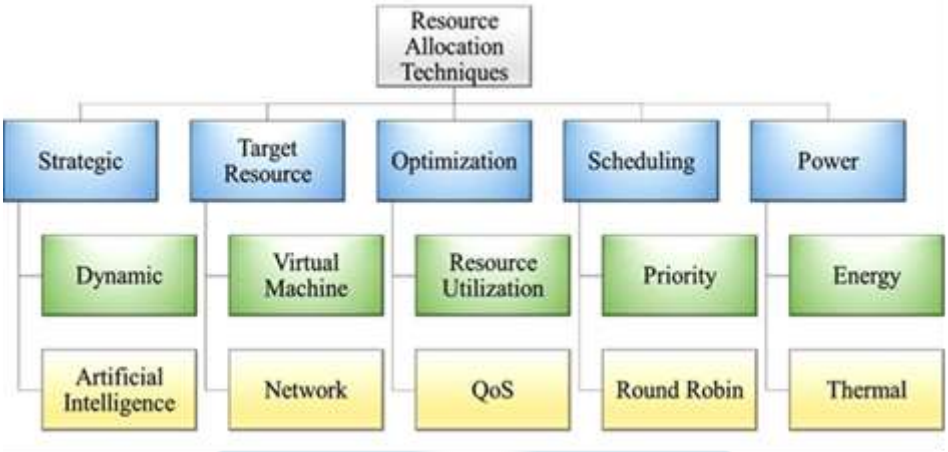

Figure 3. Taxonomy of resource allocation in cloud computing

Other characteristics include SLAs for cloud service providers and cloud service users' quality of service, fraud protection, and income.

3.1. Strategic

Because of the potential qualities for both service customers and service providers, cloud computing is gaining popularity. New needs for strategic-based resource allocation strategies have been used to satisfy customers daily. Strategic resource allocation is further classified as follows:

1) Dynamic resource allocation: technique is used by the cloud service provider to predict the nature of consumers, their demands,

2) Artificial intelligence: imitate nature to schedule tasks among resources.

3.2. Target Resource

The target resource type attribute identifies the primary resource for which the approach was designed. The kind and degree of resources given to the consumer job are determined by the target resource allocation parameters. There are two kinds of goal resource allocation-based strategies.

1. Virtual Machine
2. Network.

The current resource allocation approaches assign resources to tasks at varying granularities. Similarly, virtual machine allocation displays the virtual machine's placement on the actual computer. Furthermore, network failure in a cloud data center might occur as a result of inadequate resource allocation, logical segmentation of physical devices, and scheduling.

3.3. Optimization

The primary goal of optimization is to increase throughput by making better use of physical and virtual resources. This will allow cloud service providers to optimize their earnings by serving as many customers as possible while reducing operating costs by distributing the workload over fewer computers.

The current Resource Allocation Techniques (RAT) aim to achieve many optimization goals, such as

1. Resource Use: effective resource utilisation for environmental safety and reductions in datacenter operating consumption.
2. Quality of Service (QoS): aims to meet customers' numerous satisfaction matrices, such as latency (communication delay), CPU speed, stability, memory, and so on. Noncompliance with execution measures might lead to higher degrees of service performance violation. Quality of Service details is defined by SLAs between cloud service providers and cloud service customers.

SLA violations may have a significant influence on cloud service users' levels of satisfaction. SLA is a contract that governs the quality of service (QoS) between the cloud service provider and the cloud service user. It also incorporates the service cost, with the degree of QoS being balanced by the service cost [11]. The cloud service provider should design its system to meet the QoS requirements of all cloud components. Some QoS-based cloud service consumer-oriented resource allocation approaches attempt to meet their criteria, while others are cloud service provider-oriented and attempt to meet their requirements, which may have a detrimental impact on cloud service customer satisfaction.

3.4. Scheduling

Because of the significant resource cost and execution time, resource scheduling is an important study subject in the cloud environment. Various resource attributes and scheduling criteria are taken into account in various resource scheduling strategies. Because neither the cloud resource supplier nor the cloud service customer wishes to share their data, resource scheduling becomes more challenging. Cloud service providers consider uncertain resources when scheduling and executing workloads.

There are two types of resource scheduling techniques:

1. Priority-based scheduling: allocate resources depending on characteristics such as memory, network bandwidth, and necessary CPU time.
2. Round Robin: RR operates on a time slice, which may be thought of as a little chunk of time.

When assigning cloud computing resources, the cloud service provider prioritises the various user requests. Time, cost, and the number of processor requests are the criteria evaluated in priority-based resource allocation. To tackle the scheduling issue in a service request, the "dynamic priority scheduling algorithm" (DPSA) is used. Consumer jobs are grouped into task units in DPSA based on their individual needs when they are received and processed to plan appropriately and provide productive service.

Round Robin is the most frequent and commonly used scheduling method, making it ideal for allocating resources to tasks effectively. The RR method was created to allocate CPU time among the jobs that have been organized. Similarly, CPU time is split across the jobs in each task lineup in a queue list. Few academics have sought to enhance CPU response time [12]. A novel Round Robin-based approach that calculates and allocates the dynamic time quantum without any task arrangement based on the arrival time of jobs in the ready queue.

3.5. Power

As hosting and data centers require substantial amounts of electricity, power consumption has become a major factor. Poor resource use and hotspot issues may emerge if no effective resource utilization strategy is used. Proper resource allocation strategies may not only minimize power usage; it is further classified into two categories [13].

1. Energy-conscious: anticipate benefit level execution measures to be expanded under power dissemination and power utilization criteria.
2. Thermal-aware allocation: forecasts the thermal consequences of task placement and resource allocation based on the expected thermal impact.

4. Results and Discussion

4.1. Multi-Agent-based Dynamic Resource Allocation (MADRA)

MADRA (Multi-Agent Dynamic Resource Allocation) technique for Distributed Cloud. The agents are employed to improve QoS in terms of resource allocation. The whole process is broken into multiple sub-processes and allocated to agents on the Cloud server in this suggested work. The green navigator agent is used to route the request to the most relevant resources. The service analyzer agent understands and evaluates user service requirements. Based on the condition of load and energy in the virtual machine (VM) management and energy monitor, it determines whether to approve or refuse the request. The user profile agent gathers information on the cloud user's behaviors and attributes. Critical cloud requests are provided special privileges and prioritized above other users. The QoS parameter monitoring agent examines user requests and matches them with available resources. If there is a mismatch, the resources are removed. It computes a quality score (QS) for each matched resource in terms of time, energy, and availability, as shown in Eq (1). The resources are then rated according to their quality score. The service scheduling agent chooses the resource with the best quality score and assigns it to the appropriate incoming request, as shown in Eq (2). The pricing agent determines the cost of services depending on the time, demand, and availability of resources [13].

$$QS(i)=QS_{i}(Time)+QS(Energy)+QS_{i}(Availability) \text{ ---------- (1)}$$

Where i=1,2, 3……n

$$Maximums=max(QS(i)) \text{ where } i=1,2,3\ldots\ldots n \text{ ----------(2)}$$

4.2. Artificial Immune System-Directed Acyclic Graph (AIS-DAG)

Resource Allocation Using AIS-DAG The AIS-DAG model is used for resource allocation to solve an optimization issue in which the affinity function (fitness) is contracted with Energy (Ei) efficiency and Makespan (Ms). Eq defines the affinity function in the AIS algorithm (3). The primary role of the AIS-DAG model is to assign the best resource based on optimal values such as less energy, less time, available resource, and taking the least cost for processing the work in the resource.

$$aff(x)=e^{minEi+minMs} \text{ ----------(3)}$$

The effectiveness of the suggested ways is examined by modeling them in the Green Cloud (GC) Simulator software application. The NS-2 simulation platform serves as the foundation for GC. It can build Cloud components such as data centers, switches, servers, and network connectivity. Three-tier architecture-based application simulations are supported by GC. Network size, number of nodes, communication frequency, and Communication Computational Ratio are the simulation parameters (CCR). As illustrated in Fig. 4, the proposed study demonstrates that AIS-DAG is the best in terms of reaction time. In terms of relevance, the AIS-DAG model improves reaction time when comparing incoming requests vs. response time. When compared to MADRA and QRA techniques [14, 15], the AIS-DAG model [16] reduced reaction time by 20% and 10.56 percent, respectively.

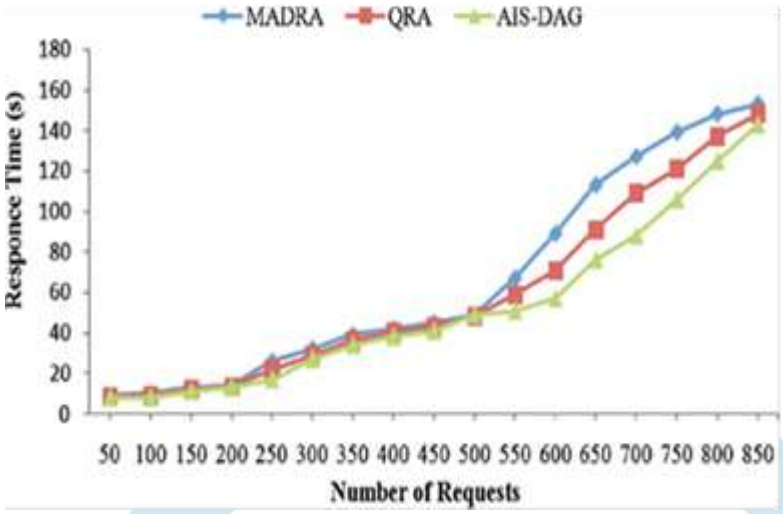

Figure 4. Comparison based on response time

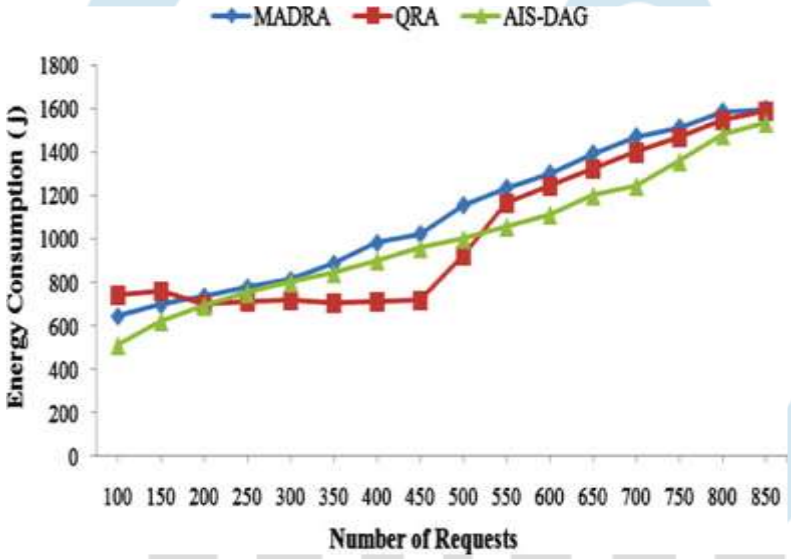

Figure 5. Comparison based on energy consumption

The quantity of energy needed is determined by the number of operations, the size, and the kind of resource. When compared to MADRA and QRA approaches, the AIS-DAG model lowers energy usage by 10.02 percent and 2.33 percent, respectively. The simulation shows that the AIS-DAG model consumes less energy than the other offered approaches. Figure 5 shows the amount of energy spent by all of the recommended strategies.

5. Conclusion

The difficulties covered here are the many resource allocation approaches based on their systems. Cost, energy, reaction time, execution time, workload, resource usage, user happiness, and SLA should all be met through an effective resource allocation approach. quality of service, as well as profit to cloud service providers for Cloud resource allocation, and an efficient communication paradigm with cheap cost, quicker reaction time, and optimal energy use is provided. For the Cloud, an efficient resource allocation mechanism has been developed. Low-cost resource allocation, quicker reaction time, and optimal energy utilisation. All of the techniques offered were examined and contrasted. The experimental findings suggest that the AIS-DAG model increases reaction time, indicating that it has a strong potential for increasing the energy efficiency of cloud data centers. It may also successfully meet the Service Level Agreement (SLA) sought by users. The AIS-DAG energy-efficient resource selection and allocation approach suggested in this study may be applied to the mobile Cloud, which is a trending technology for energy savings and enhanced battery life.

References

[1]. R. Buyya, Ch. Sh. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility", Future Generation Computing System, vol. 26, no. 6, pp. 599–616, Jun. 2009. DOI: 10.1016/j.future.2008.12.001.

[2]. S. Gong, B. Yin, Z. Zheng, and K.-Y. Cai, "Adaptive multivariable control for multiple resource allocation of service-based systems in cloud computing", IEEE Access, vol. 7, pp. 13817–13831, 2019. DOI: 10.1109/ACCESS.2019.2894188.

[3]. S. H. H. Madni, M. Sh. A. Latiff, Y. Coulibaly, and Sh. M. Abdulhamid, "Recent advancements in resource allocation techniques for cloud computing environment: A systematic review", Cluster Computing, vol. 20, no. 3, pp. 2489–2533, 2017. DOI: 10.1007/s10586-016-0684-4.

[4]. Q. Qi and F. Tao, "A smart manufacturing service system based on edge computing, fog computing, and cloud computing", IEEE Access, vol. 7, pp. 86769–86777, 2019. DOI: 10.1109/ACCESS.2019.2923610.

[5]. A Berl, E. Gelenbe, M. Di Girolamo, G. Giuliani, H. De Meer, M. Dang, and K. Pentikousis, "Energy-efficient cloud computing", The Computer Journal, vol. 53, no. 7, pp. 1045–1051, 2010. DOI: 10.1093/comjnl/bxp080.

[6]. [Ch. Guo, H. Wu, K. Tan, L. Shi, Y. Zhang, and S. Lu, "Dcell: A scalable and fault-tolerant network structure for data centers", ACM SIGCOMM Computer Communication Review, vol. 38, no. 4, pp. 75– 86, 2008. DOI: 10.1145/1402958.1402968.

[7]. B. Sotomayor, R. S. Montero, I. M. Llorente, and I. Foster, "Resource leasing and the art of suspending virtual machines", in Proc. of the 11th IEEE International Conference on High Performance Computing and Communications, 2009, pp. 59–68. DOI: 10.1109/HPCC.2009.17.

[8]. M. Shojafar, S. Javanmardi, S. Abolfazli, and N. Cordeschi, "FUGE: A joint meta-heuristic approach to cloud job scheduling algorithm using fuzzy theory and a genetic method", Cluster Computing, vol. 18, no. 2, pp. 829–844, 2015. DOI: 10.1007/s10586-014-0420-x.

[9]. X. Lu, J. Zhou, and D. Liu, "A method of cloud resource load balancing scheduling based on improved adaptive genetic algorithm", Journal of Information & Computational Science, vol. 9, no. 16, pp. 4801–4809, 2012.

[10]. S. Son, G. Jung, and S. Ch. Jun, "An SLA-based cloud computing that facilitates resource allocation in the distributed data centers of a cloud provider", The Journal of Supercomputing, vol. 64, no. 2, pp. 606– 637, 2013. DOI: 10.1007/s11227-012-0861-z.

[11]. S. Singh and I. Chana, "QoS-aware autonomic resource management in cloud computing: A systematic review", ACM Computing Surveys (CSUR), vol. 48, no. 3, article no. 42, 2016. DOI: 10.1145/2843889.

[12]. J. Tan, T.-H. Chang, and T. Q. Quelc, "Minimum energy resource allocation in fog radio access network with fronthaul and latency constraints," in Proceedings of the 2018 IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), pp. 1–5, IEEE, Kalamata, Greece, June 2018.

[13]. X. Xu, Y. Li, T. Huang et al., "An energy-aware computation offloading method for smart edge computing in wireless metropolitan area networks," Journal of Network and Computer Applications, vol. 133, pp. 75–85, 2019.

[14]. Kandan M, Manimegalai R (2015) Multi agent based dynamic resource allocation in cloud environment for improving quality of service. Aust J Basic Appl Sci 9(27):340–347.

[15]. Kandan M, Manimegalai R (2017) QRA: a multi-stage framework for improving QoS in resource allocation. J Adv Res Dyn Contr Syst 9(5):131–141.

[16]. Kandan M, Manimegalai R (2016) AIS-DAG: artificial immune system for directed acyclic graphs model based fair resource allocation for heterogeneous cloud computing. Asian J Inf Technol 15(19):3673–3686.