Multi-objective load balancing in cloud infrastructure through fuzzy based decision making and genetic algorithm based optimization

Neema George[1], Anoop Balakrishnan Kadan[2], Vinodh P. Vijayan[3]
[1]Department of Computer Science and Engineering, Srinivas University Srinivas Nagar, Mangalore, Karnataka, India
[2]Department of AIML, Srinivas Institute of Technology, Mangalore, India
[3]Department of Computer Science and Engineering, Mangalam College of Engineering, Kottayam, India

Article Info

Article history:

Received Jul 15, 2021
Revised Sep 12, 2022
Accepted Oct 12, 2022

Keywords:

Data mining and analysis
Distributed systems
Fuzzy logic
Genetic algorithm
Intelligent and knowledge-based system
Scheduling

ABSTRACT

Cloud computing became a popular technology which influence not only product development but also made technology business easy. The services like infrastructure, platform and software can reduce the complexity of technology requirement for any ecosystem. As the users of cloud-based services increases the complexity of back-end technologies also increased. The heterogeneous requirement of users in terms for various configurations creates different unbalancing issues related to load. Hence effective load balancing in a cloud system with reference to time and space become crucial as it adversely affect system performance. Since the user requirement and expected performance is multi-objective use of decision-making tools like fuzzy logic will yield good results as it uses human procedure knowledge in decision making. The overall system performance can be further improved by dynamic resource scheduling using optimization technique like genetic algorithm.

Corresponding Author:

Neema George
Department of Computer Science and Engineering, Srinivas University
Srinivas Nagar, Mangalore, Karnataka, India
Email: neemacsemlm@gmail.com

1. INTRODUCTION

Cloud computing can be explained as on-demand availability of services like cloud servers, resources, storage and computing power, which is managed remotely in internet. The term is usually used to define data centers accessible to consumers over the Internet. The service models of cloud computing [1]–[10] are platform as a service (PaaS), infrastructure as a service (IaaS) and software as a service (SaaS) [8], [9]. Hence any user they like to use the above-mentioned services can avail the services after paying the service coast but user always enjoys the uninterrupted services without facing the difficulty of maintaining the same.

Amazon web service (AWS), Microsoft Azure, Server Space, Google Cloud Platform, Adobe Creative Cloud, IBM Cloud Services, and VMware are the major cloud service providers. When the multiple users have multi objective requirement the cloud infrastructure operation is difficult as it will not be able to provide good QoS to all the clients. Service migration amid data servers may reduce the network overhead in a cloud infrastructure and improve QoS to the clients but it will create serious load balancing problems which ultimately degrade the performance of the system. Figure 1 shows basic cloud architecture with various services like infrastructure, applications and platform which can be accessed by multiple users in multiple configurations through internet. Infrastructure services contain services like server, computing power and data storage.

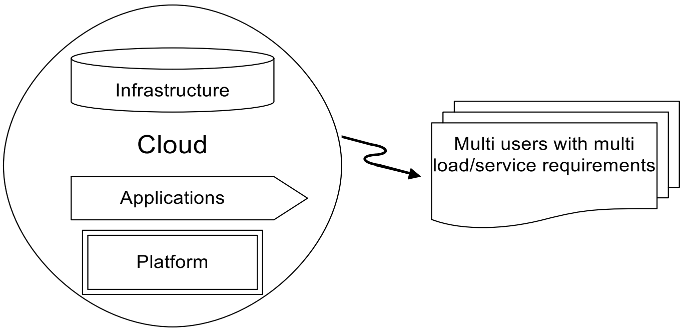

Figure 1. Cloud architecture

Application services include application software, middleware and compiler. Platform services may contain various operating system as a service, all these services can be accessed through browser and internet. While we plan for load balancing in cloud it is important to perform the same without affecting the principles of cloud computing like virtualization, resource pooling, elasticity, metered billing and automatic resource deployment.

2. BACKGROUND WORK

An improved particle swarm optimization (IPSO) [10], [11] algorithm was introduced to increase the virtual machine resource scheduling performance in cloud computing environment. The designed algorithm changed the constant coefficients of cognition and social items in the velocity variation to number of iterations. IPSO algorithm was more balanced as stronger processing ability virtual machines were allocated with more tasks.

Fuzzy-based multi-dimensional resource scheduling and queuing network (F-MRSQN) [12] method was introduced for integrated scheduling and load balancing algorithm F-MRSQN method used minimum resource and time for scheduling and load balancing in cloud environment infrastructure. Fair load balancing was achieved by multi-dimensional load optimization algorithm through increasing number of virtual machines. Cloud workflow scheduling strategy [13], [14] was introduced to achieve efficient scheduling process in cloud computing environment. Cloud workflow algorithm was introduced for performing scheduling optimization. Heuristic-based dynamic load-balancing algorithm was introduced utilized that in turn monitored the virtual machines in a continuous manner resulting in significant resource utilization [15]–[22].

Priority aware longest job first (PA-KJF) method that efficiently predicts overloading hosts, therefore, minimizing the number of migrations [15]. Heuristic-based dynamic load-balancing algorithm was introduced utilized that in turn monitored the virtual machines in a continuous manner resulting in significant resource utilization PA-KJF method that efficiently predicts overloading hosts, therefore, minimizing the number of migrations. For optimizing the load and effficient scheduling of resources [16] for each cloud user request with the efficient evolution of the data center, multi-objective resource scheduling optimization technique was applied by multi constraints through resource scheduling in infrastructure cloud services.

On Apache Spark, a parallel application towards accelerating N-FINDR [23] unmixing method and support vector machine (SVM) classifier in a fusion-based hyperspectral image classification in a wireless sensor application creates a trade-off between computational overhead and energy consumption. The cloud resource management is proved as a combinatorial optimization problem where the complexity belongs to NP-hard. When compared with classic techniques like, reinforcement learning (RL) as a special model of machine learning devised techniques like DeepRM, DeepRM_Plus [24] could offer 37.5 percentage faster with respect to the convergence rate. And the above two techniques are much beeter in case of parameters like average-weighted turnaround time and the average cycling time. The application of the modern metaheuristic whale optimization algorithm (WOA) [25] for the cloud task scheduling with multi-objective optimization model could impove the performance of a cloud system for a certain computing resource which could also contribute to improve accuracy and convergence speed in searching for the optimum task scheduling plans.

3. PROPOSED SYSTEM

The load balancing in cloud environment [17], [18] can be of two types, viz., static and dynamic load balancing. Through load balancing, it is expected to improve parameters like overall performance, system stability, quality of service (QoS) [19], fault tolerance [20] as it is essential to improve the service. Static load balancing never gives better performance as load demand and configuration requirements changes frequently and dynamically. The conventional dynamic algorithm also never gives good performance as the requirement is multi-objective. The bio-inspired algorithm like genetic algorithm as an optimization tool is expected to give improves solution without overhead due to its simplicity and operational principle.

3.1. Fuzzy based decision making and genetic algorithm-based optimization

Figure 2 shows the proposed system where the multi objective requirement of users are considered and decision making is done using fuzzy logic [10], [21], [22], [26]–[28] without compromising the QoS. As the actual load on the system is dynamic and heterogeneous in nature with multiple objectives to be considered a dynamic scheduling of resources is required where genetic optimization is the sufficient tool. Figure 3 shows the overall algorithm used for a stepwise approach to solve the cloud resource scheduling using the fuzzy decision making and genetic alogorithm based optimization. The initial decision making based on fuzzy rules which uses human expertizes mainly. The heterogeneous nature of environment and user requirement cerate lot of loads inbalance and which ultimately degrades system performance. The application of genetic algorithm with suitable selection, cross over and mutation technique can inprove overall system performance.

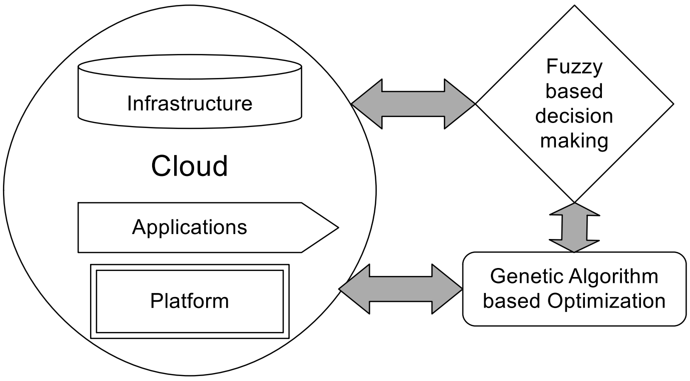
Infrastructure
Cloud
Applications
Platform
Fuzzy based decision making
Genetic Algorithm based Optimization

Figure 2. Fuzzy based decision making and genetic algorithm-based optimization architecture

```
Algorithm for Multi-objective cloud resource scheduling
Step 1: Collection of heterogeneous requests from various client.

Step 2: Fuzzy based resource allocation based on membership function and fuzzy rules.
        1. Fuzzification and membership fixing
        2. Fuzzy engine and fuzzy rules
        3. Defuzzifcation and final decision.

Step 3: Measuring the performance parameters and load.

Step 4: Optimization using Genetic Algorithm.
       1. Representation of parameters
       2. Initial population selection
       3. Cross over and muatation
       4. Calaculation of fitness function and optimum value
       5. If acceptable “stop “else go back to selection.
Step 5: Repeat step 1 periodically or when quick degradation in performance
```

Figure 3. Alogithm for multi-objective cloud resource scheduling

3.2. Fuzzy based decision making

Figure 4 shows a fuzzy logic system where all the input parameters are fuzzified using any of fuzzification method. The major decision is done at inference engine using rule base made based on expert knowledge. Finally, defuzzification is done, which the final decision is given to system. Figure 5 shows the sample membership function for input load, the various linguistic variables are LOW, AVERAGE and HIGH, the number of linguistic variables can be increased based on the user requirement similarly multiple parameters can be considered as input membership function based on the multi-objective requirement of user.

The membership function (MF) considered is a triangular MF due to the nature of input variable. The triangular membership function can be defined as by considering Figure 6 and corresponding point a, b and c. In (1) can be used to calculate membership value for any ‘x’ value which is nothing but measures load value.

$$\mu_A(x) = \begin{Bmatrix} 0, if\,(x \le a) \\ \frac{x-a}{b-a}, if\,(a \le x \le b) \\ \frac{c-x}{c-b}, if\,(b \le x \le c) \\ 0, if\,(x \ge c) \end{Bmatrix} \quad (1)$$

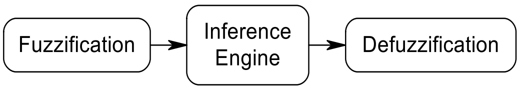

Figure 4. Fuzzy system

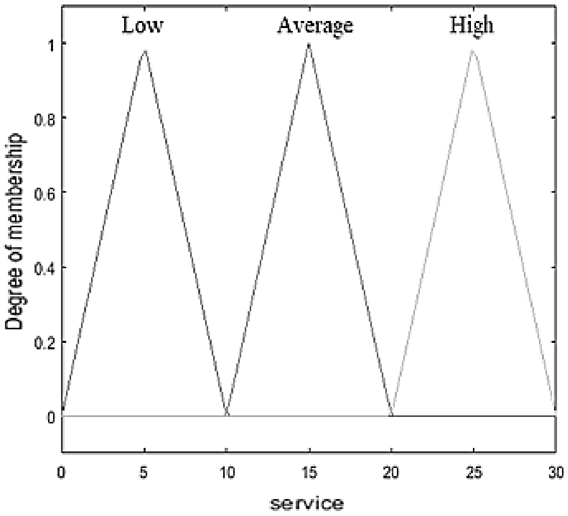

Figure 5. Sample membership function for input load

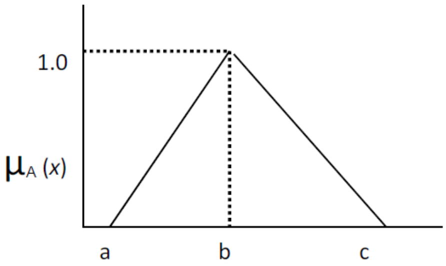

Figure 6. Triangular membership function

Figure 7 shows the output membership function for the system where linguistic variables are POOR, GOOD and EXCELLENT, the fuzzy inference system will map the output to corresponding degree of linguistic variable based on the rule base. The output membership function is Gaussian MF which can be represented as Gaussian (x: c, s) where c represents the mean and s represents standard deviation. Now it is important that the accuracy of the decision always depends on a rule base and the type of membership function selected. In the sample case the input numbers if function is a triangle membership function and output membership function is a normal distribution, the rule base can be e always improved through availing best expert knowledge.

$$\mu_A(x,c,s,m) = exp\left[-\frac{1}{2}\left|\frac{x-c}{s}\right|^m\right] \tag{2}$$

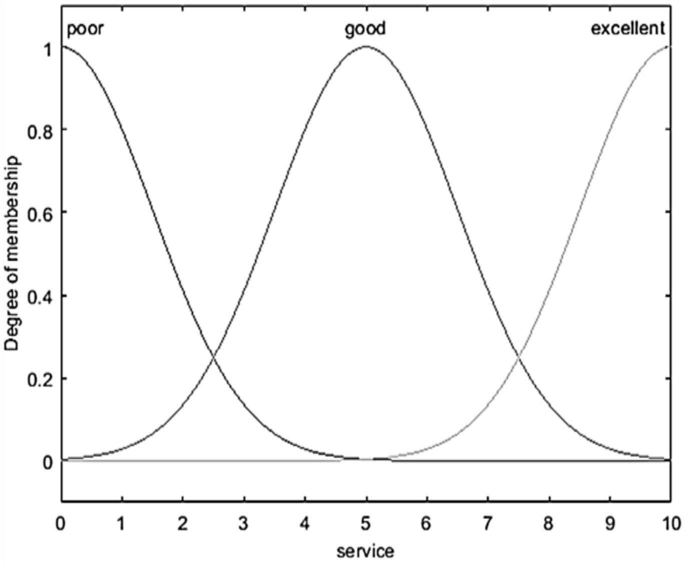

Figure 7. Sample output membership function for service quality

3.3. Genetic algorithm-based optimization and scheduling

Figure 8 shows genetic algorithm-based optimization where the effective scheduling can be done in this multi-objective environment and genetic algorithm is always an excellent tool due to its simplicity and bio-inspired modelling. The selection of crossover technique and mutation technique always have a role in the effectiveness similarly the probability of mutation, percentage of crossover also have significant effect on the output. Here as the requirement is a multi objective the chromosome representation seems to be challenging hence it is important to choose a suitable representation technique.

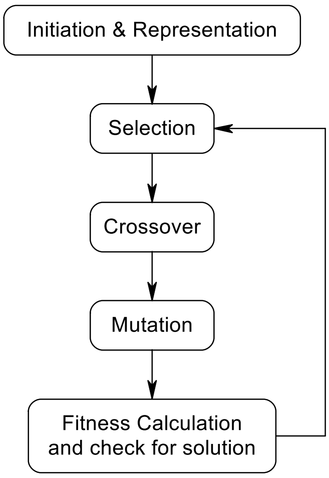
Initiation & Representation
Selection
Crossover
Mutation
Fitness Calculation and check for solution

Figure 8. Genetic algorithm-based optimization

4. EXPERIMENTATION AND RESULT

The simulation of the network has been done in the cloudsim simulator with various possible configuration and varying workload, arrival time. The types of jobs use for testing are batch jobs and long running jobs. In the both case small, medium and large jobs are taken with different task time, CPU request and memory requests. Different workloads like bursty, slow and mixed are considered with variable batch and service type. Table 1 shows the cost measurement for three different workloads.

Table 1. Cost measurement for three different workloads

Cost for various workload in Percentage		
Slow	Bursty	Mixed
30	30	30
35	34	34
40	42	43
45	43	43
50	50	46
55	59	50
60	60	52
65	62	58
70	68	60
75	70	64

Figure 9 shows the cost occurred with reference to various workloads. And from the diagram it is evident that the slow load doesn’t make much improvement in the proposed solution but the bursty load has a slight cost reduction compared to slow load. Finally, the mixed load has a good impact on this scheduling where there is significant reduction in the cost with reference to time.

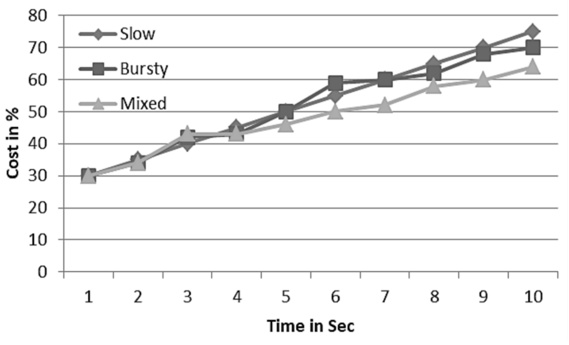

Figure 9. Cost for various workload

Figure 10 shows the overhead occurred due to scheduling and decision making. The decision making is done through fuzzy inference and optimized scheduling is calculated through genetic algorithm. Both of this technique could improve overall system performance but the computational overhead is slightly increased. It is visible from the diagram that the overhead is minimum for low load and it is maximum for the mixed load due to the heterogeneous nature of the load.

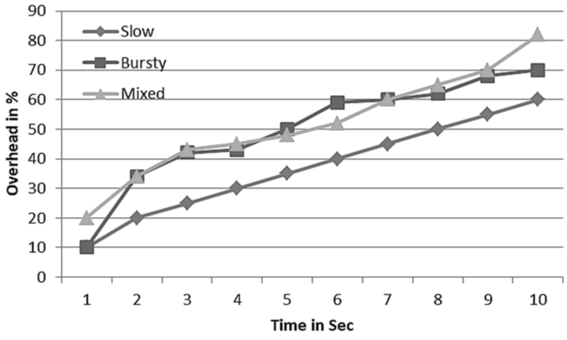

Figure 10. Overhead in scheduling

Figure 11 shows the computation overhead incurred due to application of optimization tool genetic algorithm. It is clear from the graph that as the number of users increase the overhead increases, but in few cases the overhead is random. And due to nature of genetic algorithm a lucky mutation can yield excellent performance with low overhead.

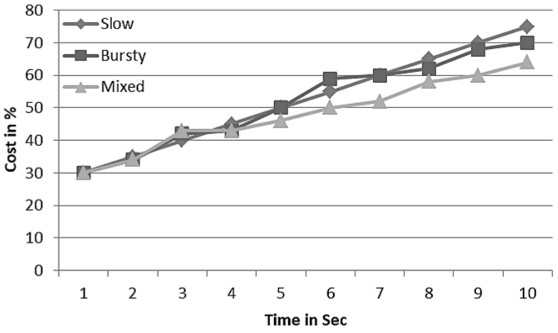

Figure 11. Computational overhead of genetic algorithm

5. CONCLUSION

The high demand for cloud resources like infrastructure, platform and applications has created lot of scheduling issues and also it creates serious load balancing issues and ultimately which degrades system performances. Due the heterogeneous nature of resource requirement with respect to time, it is important to use any dynamic scheduling algorithm to dynamically allocate the resources based on the user requirement. The Genetic based optimization resources gave improves system performance by reducing the overhead and cost. The fuzzy based decision making improves the allotment as it involved human procedure knowledge. The result of system performance for various load conditions like low, average and mixed, shows that fuzzy based decision making and genetic based optimization is effective for mix load condition essentially.

REFERENCES

[1] D. Zhao, M. Mohamed, and H. Ludwig, “Locality-aware scheduling for containers in cloud computing,” *IEEE Transactions on Cloud Computing*, vol. 8, no. 2, pp. 635–646, 2020, doi: 10.1109/TCC.2018.2794344.

[2] N. Tziritas *et al.*, “Online inter-datacenter service migrations,” *IEEE Transactions on Cloud Computing*, vol. 8, no. 4, pp. 1054–1068, 2020, doi: 10.1109/TCC.2017.2680439.

[3] V. P. Vijayan and E. Gopinathan, “Improving network coverage and life-time in a cooperative wireless mobile sensor network,” *Proceedings-2014 4th International Conference on Advances in Computing and Communications, ICACC 2014*, pp. 42–45, 2014, doi: 10.1109/ICACC.2014.16.

[4] S. E. Mahmoodi, R. N. Uma, and K. P. Subbalakshmi, “Optimal joint scheduling and cloud offloading for mobile applications,” *IEEE Transactions on Cloud Computing*, vol. 7, no. 2, pp. 301–313, 2019, doi: 10.1109/TCC.2016.2560808.

[5] A. S. Abdalkafor, A. A. Jihad, and E. T. Allawi, “A cloud computing scheduling and its evolutionary approaches,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 21, no. 1, pp. 489–496, 2021, doi: 10.11591/ijeecs.v21.i1.pp489-496.

[6] S. Zaineldeen and A. Ate, “Improved cloud data transfer security using hybrid encryption algorithm,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 20, no. 1, pp. 521–527, 2020, doi: 10.11591/ijeecs.v20.i1.pp521-527.

[7] S. Ouhame and Y. Hadi, “Enhancement in resource allocation system for cloud environment using modified grey wolf technique,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 20, no. 3, pp. 1530–1537, 2020, doi: 10.11591/ijeecs.v20.i3.pp1530-1537.

[8] M. Parra-Royon and J. M. Benítez, “Fuzzy systems-as-a-service in cloud computing,” *International Journal of Computational Intelligence Systems*, vol. 12, no. 2, pp. 1162–1172, 2019, doi: 10.2991/ijcis.d.190912.001.

[9] J. K. R. Sastry and M. T. Basu, “Securing SAAS service under cloud computing based multi-tenancy systems,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 13, no. 1, pp. 65–71, 2019, doi: 10.11591/ijeecs.v13.i1.pp65-71.

[10] H. Yu, “Evaluation of cloud computing resource scheduling based on improved optimization algorithm,” *Complex and Intelligent Systems*, vol. 7, no. 4, pp. 1817–1822, 2021, doi: 10.1007/s40747-020-00163-2.

[11] Y. Zhang and R. Yang, “Cloud computing task scheduling based on improved particle swarm optimization algorithm,” *Proceedings IECON 2017-43rd Annual Conference of the IEEE Industrial Electronics Society*, vol. 2017-Janua, pp. 8768–8772, 2017, doi: 10.1109/IECON.2017.8217541.

[12] V. Priya, C. Sathiya Kumar, and R. Kannan, “Resource scheduling algorithm with load balancing for cloud service provisioning,” *Applied Soft Computing Journal*, vol. 76, pp. 416–424, 2019, doi: 10.1016/j.asoc.2018.12.021.

[13] Y. Hu, H. Wang, and W. Ma, “Intelligent cloud workflow management and scheduling method for big data applications,” *Journal of Cloud Computing*, vol. 9, no. 1, 2020, doi: 10.1186/s13677-020-00177-8.

[14] J. K. Konjaang and L. Xu, “Multi-objective workflow optimization strategy (MOWOS) for cloud computing,” *Journal of Cloud Computing*, vol. 10, no. 1, 2021, doi: 10.1186/s13677-020-00219-1.

[15] M. Kumar and S. C. Sharma, “PSO-based novel resource scheduling technique to improve QoS parameters in cloud computing,”

Neural Computing and Applications, vol. 32, no. 16, pp. 12103–12126, 2020, doi: 10.1007/s00521-019-04266-x.

[16] S. Ramamoorthy, G. Ravikumar, B. Saravana Balaji, S. Balakrishnan, and K. Venkatachalam, "MCAMO: multi constraint aware multi-objective resource scheduling optimization technique for cloud infrastructure services," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 6, pp. 5909–5916, 2021, doi: 10.1007/s12652-020-02138-0.

[17] S. Afzal and G. Kavitha, "Load balancing in cloud computing-A hierarchical taxonomical classification," *Journal of Cloud Computing*, vol. 8, no. 1, 2019, doi: 10.1186/s13677-019-0146-7.

[18] C. Jittawiriyanukoon, "Cloud computing based load balancing algorithm for erlang concurrent traffic," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 17, no. 2, pp. 1109–1116, 2019, doi: 10.11591/ijeecs.v17.i2.pp1109-1116.

[19] S. Potluri and K. S. Rao, "Optimization model for QoS based task scheduling in cloud computing environment," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 18, no. 2, pp. 1081–1088, 2020, doi: 10.11591/ijeecs.v18.i2.pp1081-1088.

[20] V. P. Vijayan and N. Kumar, "Extending connectivity and coverage using robot Initiated k-nearest dynamic search for WSN communication," *International Journal of Control Theory and Applications (IJCTA)*, vol. 9, no. 41, pp. 1171–1177, 2016.

[21] V. P. Vijayan and N. Kumar, "Coverage and lifetime optimization of WSN using evolutionary algorithms and collision free nearest neighbour assertion," *Pertanika Journal of Science and Technology*, vol. 24, no. 2, pp. 371–379, 2016.

[22] M. A. Rodriguez and R. Buyya, "Deadline based resource provisioningand scheduling algorithm for scientific workflows on clouds," *IEEE Transactions on Cloud Computing*, vol. 2, no. 2, pp. 222–235, 2014, doi: 10.1109/tcc.2014.2314655.

[23] J. Sun *et al.*, "Multiobjective task scheduling for energy-efficient cloud implementation of hyperspectral image classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 587–600, 2021, doi: 10.1109/JSTARS.2020.3036896.

[24] W. Guo, W. Tian, Y. Ye, L. Xu, and K. Wu, "Cloud resource scheduling with deep reinforcement learning and imitation learning," *IEEE Internet of Things Journal*, vol. 8, no. 5, pp. 3576–3586, 2021, doi: 10.1109/JIOT.2020.3025015.

[25] X. Chen *et al.*, "A WOA-based optimization approach for task scheduling in cloud computing systems," *IEEE Systems Journal*, vol. 14, no. 3, pp. 3117–3128, 2020, doi: 10.1109/JSYST.2019.2960088.

[26] B. Hayat, K. H. Kim, and K. Il Kim, "A study on fuzzy logic based cloud computing," *Cluster Computing*, vol. 21, no. 1, pp. 589–603, 2017, doi: 10.1007/s10586-017-0953-x.

[27] M. J. Dos Santos and E. A. D. M. Fagotto, "Cloud computing management using fuzzy logic," *IEEE Latin America Transactions*, vol. 13, no. 10, pp. 3392–3397, 2015, doi: 10.1109/TLA.2015.7387246.

[28] M. Jaiganesh and A. V. Antony Kumar, "B3: Fuzzy-based data center load optimization in cloud computing," *Mathematical Problems in Engineering*, vol. 2013, 2013, doi: 10.1155/2013/612182.

BIOGRAPHIES OF AUTHORS

Ms. Neema George is a Research Scholar in Computer Science and Engg, Srinivas University, Mangalore. Working as an Assistant Professor in Mangalam College of Engg, Kottayam, Kerala. Having 10 years of teaching experience in MLMCE. Master of Engineering in Computer science and Engineering (M.E CSE) from Anna University Chennai and Bachelor of Technology in Computer Science and Engineering (B.Tech-CSE) from MG University, kerala. Her Area of interest Cloud Computing, Machinelearning, Artificial Intelligence and her Research area is Cloud computing. She can be contacted at email: neemacsemlm@gmail.com.

Dr. Anoop Balakrishnan Kadan is Professor in AIML, Srinivas Institute of Technology Mangalore.He has received B.E degree from anna University Chennai in the year 2008, M.tech from VTU Karnataka in the year 2010 and Ph.D from APJ Abdul Kalam Technological University Kerala in the year 2020. His area of research is Machine Learning. He can be contacted at email: dranoopbk@sitmng.ac.in.

Dr. Vinodh P. Vijayan Principal, Mangalam college of Engineering, Ettumanoor, India has completed UG in ECE, PG in CSE and Ph.D. in Computer Science and Engineering in soft computing and Wireless Sensor Networks. His research area or research includes AI, Softcomputing, Datascience and cloud computing. He can be contacted at email: vinodhpvijayan81@gmail.com.