Ph.D. Research Proposal

Doctoral Program in "Department Name"

RT-NRT: Real-Time and Non-Real-Time Tasks Oriented Scheduling by SLA Constraints and VM Load Balancing in Cloud Data Centers

by

<Name of the Candidate>

<Reg. No of the Candidate>

<Supervisor Name>

<Date of Submission (DD MM 20YY>

I. INTRODUCTION / BACKGROUND

Cloud computing is an emerging paradigm that utilizes computing resources over online distribution service. Cloud data centers are designed to effectively utilize the cloud resources such as networks, storage, services, applications and servers. Power consumption is one of the most important issues for the operation and maintenance of cloud data centers. Cloud applications consume huge amount of energy, high operational cost and carbon emission to environment. Many previous works addressed several cloud computing main issues such as VM migration, placement, scheduling and load balancing etc. In this specialty, task scheduling is always hot topic. Task scheduling is performed with support of dedicated and reliable cloud services which increases dynamic scheduling process. In this process, load balancing and task scheduling is performed based on weighted random and feedback mechanism. Resources in cloud are shorted by weight randomly and then it acquired corresponding information to make load filter. At last, it achieved self adaptively to system load through feedback mechanism in load balancing [1]. Cost minimization problem in the hybrid cloud data centers is overwhelmed using temporal task scheduling algorithm (TTSA). TTSA effectively transforms incoming tasks into private CDC and public cloud in which cost minimization problem is modeled as mixed integer linear program and solved by hybrid combination of Simulated Annealing and Particle Swarm Optimization (SA-PSO). TTSA method effectively increases the throughput and reduces the cost of private CDC [2].

A dynamic task quantum variable is calculated using combination of Shortest Job First (SJF) and Round Robin (RR) algorithm to effectively schedule task into cloud. This algorithm performed two main processes, first to balance waiting time between short and long task. Second it separates the task queue into two namely Q1 and Q2. Q1 contains short task whereas Q2 contains long task, while scheduling it takes two tasks from Q1 and one task from Q2. Using this algorithm waiting time of task is reduced [3]. A static load balancing strategy that concerns about minimization of Makespan which can be achieved through Multi-Rumen Anti-Grazing algorithm (MARG) which contains three phases. At first phase, it calculates two

matrices namely task completion time and task execution time with communication delay (TECD). Second phase, it allocates task to respective VM and in third phase reschedule task into new VM by comparing new Makespan value with old Makespan value [4].

Scheduling problem in cloud computing is solved by hybrid load balancing technique BSO. BSO is combination of Bacterial Foraging Optimization (BFO) and Particle Swarm Optimization (PSO) algorithm. Local search in scheduling is achieved using BFO algorithm and global search is achieved using PSO algorithm. . In this process, jobs are collected as batches and for each job it allocates specific resources. After resource allocation process, Makespan, operational cost and utilizations are calculated. Using these values jobs are scheduled to proper virtual machines [5]. Throttled Modified Algorithm (TMA) is used to improve the response time of Virtual Machine in the cloud computing. TMA load balancer performs load balancing by updating, maintaining two index variable namely busy index and available index. In available index, status of VM is '0' whereas in busy index status of VM is '1' [6, 7].

Load balancing in virtual machine performed successfully using hybrid combination of Teaching-Learning-Based Optimization (TLBO) and Grey Wolves Optimization algorithms (GWO) which increases throughput. Scheduler is responsible to allocate jobs on virtual machine in distributed system and provides scheduling for resource allocation. This method increases throughput greater than PSO and BBO algorithms [8, 9]. In this process, Resource Intensity Aware Load (RIAL) algorithm is used which dynamically allocates different weights to different resources based on their usage intensity which significantly reduces the time and cost to avoid load balance. It avoids unnecessary migrations using strict migration algorithm [10]. Virtual machine migration is minimized using stochastic load balancing scheme. In this process, migration of virtual machine considers distance between source and destination physical machines which interns minimizes the VM migration overhead in load balancing [11]. Minimization of Makespan is performed using Ant Colony Optimization algorithm (ACO) which find the optimum resources for batch of tasks in dynamic environment of cloud computing. Implementation results from simulation showed that proposed algorithm produces the better results than existing algorithm [12].

1.1 Research Outline & Scope

In this research work, novel energy efficient algorithm using heuristics VM consolidation and task scheduling are developed to address the problem of efficient delay reduction in cloud data centers for all kinds of tasks such as real-time and non-real-time.

1.2 Research Objectives

The main objective is to fulfill the users delay constraints while submitting their tasks and also reduce resource consumption. To improve efficiency for real-time task scheduling in cloud datacenter is the prime objective of this research work.

II. RESEARCH GAPS

2.1 Common Problem Statement

Scheduling of tasks in Cloud computing is a critical process, since it deals with large amount of data. Further, response time increases for real-time tasks. So that task scheduling must be performed for real-time tasks first and then non-real time tasks are scheduled. Maintenance of this large amount of data is crucial as for improving the utilization of resources in cloud and minimizing response time and completion time. For all these purpose scheduling of tasks is must. Frequent tasks arrival in cloud data centers increases load imbalance issue. Hence, VM load is wholly imbalanced. Some of the crucial existing issues are follows,

- Tradeoff between the makespan and load balancing
- Supports only on the task scheduling and doesn't support load balancing
- Only one objective is considered

2.2 Problem Definition

In [1], the SJF scheduling policy was presented, which handled with numerous iterations. In this process, the minimum available resources were assessed by using the total amount of available computing energy in a cloud data center. In the SJF scheduling, the shortest jobs have a

high level of priority so that jobs are assigned to resources in advance. Therefore, the longest jobs would be waiting a long time for task completion.

In [2], the CWSA was proposed for multi-tenant task. The tasks were submitted to service queue and sorted according to the deadline priority. The scheduler was involved to schedule the workflows based on three policies.

- Scheduling was done by the FCFS algorithm with respect to the arrival of tasks.
- Scheduling was performed by EASY backfilling by considering the deadline of tasks.
- Scheduling was done by the MCT method, which was based on minimum completion time of tasks.

With the FCFS, incoming tasks were sorted by the scheduler in the arriving order of tasks. Based on this order, if the required resources are available to execute the first task, that is immediately processed. Otherwise, the tasks wait until the resource become available for tasks that leads to increased waiting time for allocating the tasks based on resources.

In [3], the usefulness forecast aware VM consolidation model was applied every so often to adapt and optimize the VM Placement as said by the workload. This work aims for migration the full VMs to other under loaded PMs. The main drawback of this work is that if at least one resource i.e., CPU or memory exceeds total capacity, PMs will be considered as overloaded. Therefore, the number of VM migration is high in this process.

III. RESEARCH CONTRIBUTIONS

This proposed work aims to balance the load on the VM and also schedule the arriving tasks. Cloud computing consists of several requests arriving at every minute and has the responsibility to respond to all the requests. The cloud focus on task scheduling because, the user satisfaction lies in the service offered. We have three different components in our system. They are the task manager, the scheduler and the resource manager. Each of these components will play an important role in the process of achieving the objective of the proposed system.

The task scheduler will receive tasks and classified into two types such as real time and no real time and directs the tasks to the scheduler. The resource manager will monitor and provide the VM information to the scheduler. The scheduler will perform the proposed algorithm and achieve the objective.

In this framework, the problems in existing task scheduling and load balancing is addressed in cloud computing. In this work, four algorithms are used namely **Deep Reinforcement Learning** for task scheduling, **Analytical Hierarchy Process (AHP)** is used for ranking the real-time tasks as per the SLA constraints, **Neutrosophic C- Means Clustering** is used for clustering the VM and **Spiral Optimization Algorithm** is used for load balancing.

We propose, **Deep Reinforcement Learning** algorithm for task scheduling. In this process, incoming tasks are represented as a graph in which each graph consists of two or more groups based on priority of task. In grouping process, threshold is set to find the lowest priority (no-real-time) and highest (real-time) priority of each task. To get prioritized task, we are using metrics namely SLA constraints (service rate, makespan and deadline). If any new tasks are arriving after scheduling all previous tasks, then it inserted into free list. DRL updates the priorities of successors, and inserts any newly free successors into the free list. This process is repeated until all tasks are scheduled. After getting scheduled task, we proposed **AHP** algorithm to rank each scheduled task and highest priority task is scheduled first for next process. After prioritization, load balancing is performed in which Virtual machines are clustered using **NCM algorithm** which clusters the VM based on kernel. After clustering of VM, load is balanced using **Spiral Optimization algorithm** which considers weight/capacity of each server and current client connectivity of server which interns balances the load. Highly weighted servers/least connected servers are selected to allocate task which interns improves the response time.

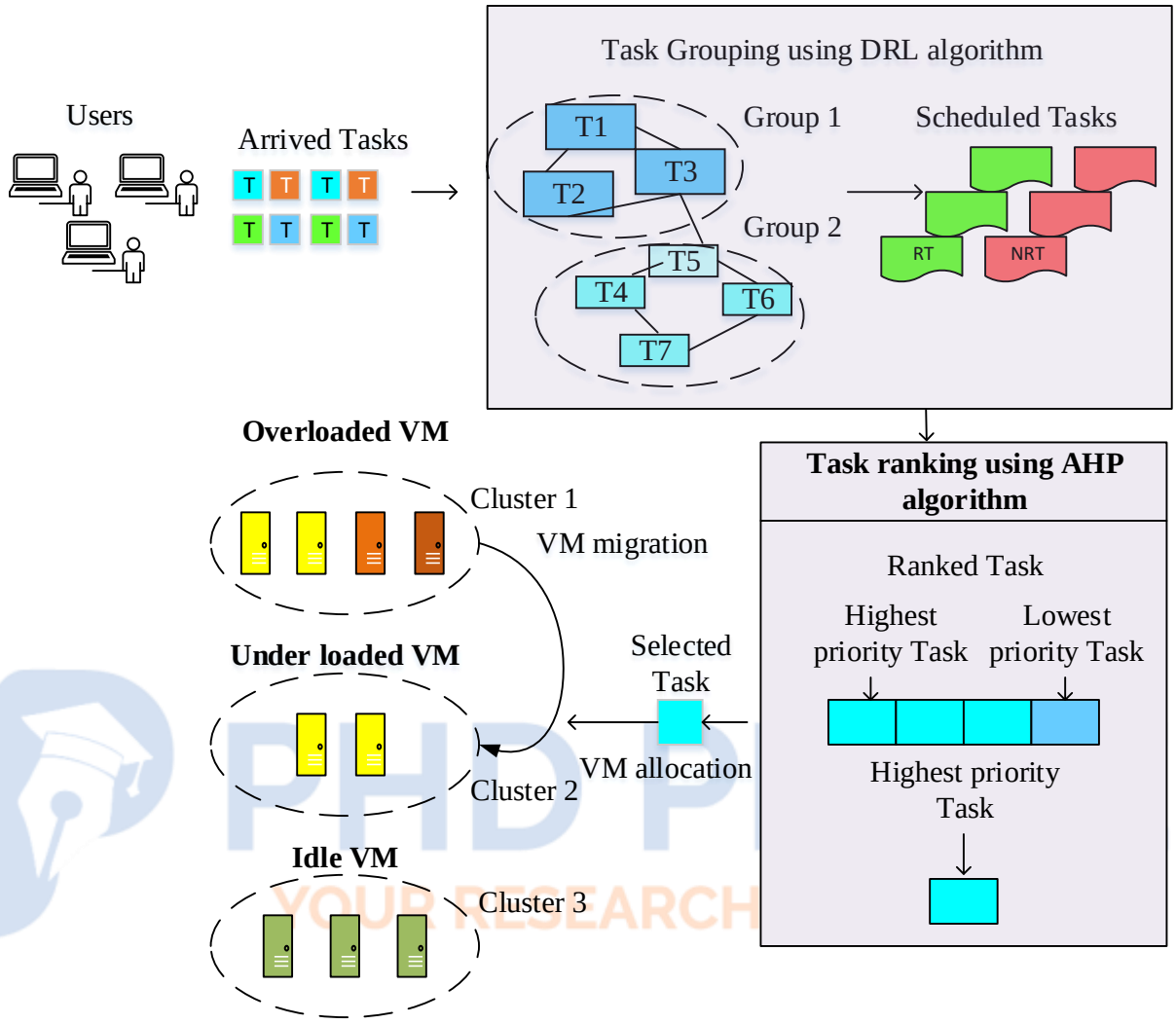
Performance Evaluation

In performance evaluation step, following metrics are calculated,

- Response time vs. Number of Tasks
- Makespan vs. Number of Tasks

- Resource utilization vs. Number of Tasks
- Service reliability vs. Number of Tasks
- Energy Consumption vs. Number of Tasks

SYSTEM ARCHITECTURE



IV. RESEARCH NOVELTIES

- Energy efficient resource allocation is satisfied with significant constraints such as unnecessary wastage elimination, dynamic resource scalability, efficient resource utilization and location independence. In order to conserve energy, deep reinforcement learning with heuristic algorithms are presented
- For VM allocation and to remove VM eliminating process, Spiral Optimization Algorithm is proposed. It is an optimization analysis method and utilizes it to predict the future utilization tendency. Hence, frequent VM migration can be avoided.

V. PREVIOUS WORKS & LIMITATIONS

Paper 1

Title – Task scheduling scheme based on sharing mechanism and swarm intelligence optimization algorithm in cloud computing

Concept

In this paper, the author proposes a hybrid intelligent optimization algorithm of fusion sharing mechanism was proposed to realize dynamic scheduling of cloud tasks. First, the virtual machine scheduling is encoded as bees, ants and genetic individuals. Then, using artificial bee colony (ABC), ant colony optimization (ACO) and genetic algorithm (GA), the optimal solution is found in each neighborhood. Finally, by a mechanism of sharing, three algorithms regularly exchange their solutions and obtain the optimal solution as the current optimal solution for the next iteration process, in order to accelerate the algorithm convergence and enhance the accuracy of convergence.

Limitations

- Ant colony optimization suffers from trade-off between makespan and load
- Algorithmic complexity as it requires more than one iterations for task scheduling

Paper 2

Title – A multi-model estimation of distribution algorithm for energy efficient scheduling under cloud computing system

Concept

In this paper, the author proposes a multi-model estimation of distribution (mEDA) algorithm to determine both task processing permutation and voltage supply levels (VSLs). The primary performance goal is to reduce the execution time (makespan) of the application. As the need to cloud computing grows, the environmental influence of data centers attracts much attention. This paper aims at the scheduling of the precedence-constrained parallel application to

minimize time and energy consumption efficiently. Specific operators to decrease execution time and energy consumption are designed. An improvement operator is also designed to enhance the diversity of the non-dominated solutions.

Limitations

- Tasks are scheduled based on arrival time. So, when a large task arrives first it will be executed for a long time
- Waiting time of the tasks is increased

Paper 3

Title – Improving Cloud Computing Performance Using Task Scheduling Method Based on VMs Grouping

Concept

In this paper, the author proposes an approach for improving cloud computing performance using task scheduling method based on VMs grouping. In this paper, a method was introduced for scheduling workload based on VM grouping in cloud environments. The aim of the proposed method is improving cloud computing performance by reducing makespan and response time, and also through increasing VMs utilization. The incredible rise of virtualization technology in cloud environments results the fostering workload which needs services provided by cloud resources. Task scheduling and Load balancing amongst the VMs and minimizing the makespan of the tasks are stimulating research concerns.

Limitations

- Frequent grouping of VM as the VM status changes dynamically.
- Increased in delay for scheduling and task processing

Paper 4

Title – Bi-objective decision support system for task-scheduling based on genetic algorithm in cloud computing

Concept

In this paper, the author proposes a bi-objective decision support system for task scheduling based on genetic algorithm in cloud computing. This paper addresses the task-scheduling in cloud computing. This problem is known to be NP-hard due to its combinatorial aspect. The main role of the proposed model is to estimate the time needed to run a set of tasks in cloud and in turn reduces the processing cost. A genetic approach for modeling and optimizing a task-scheduling problem in cloud computing is proposed. The experimental results demonstrate that the proposed solution successfully competes with previous task-scheduling algorithms

Limitations

- Though decision making is fast, it is not stable as only two parameters are considered for task scheduling
- Scheduling is not effective

Paper 5

Title – A Load Balancing Task Scheduling Algorithm based on Feedback Mechanism for Cloud Computing

Concept

In this paper, load is balanced using load balancing task scheduling algorithm. Load balancing and task scheduling is performed based on weighted random and feedback mechanism. Resources in cloud are shorted by weight randomly and then it acquired corresponding information to make load filter. Resource attributes are divided into two namely static and dynamic. At last, it achieved self adaptively to system load through feedback mechanism. Feedback mechanism yields better results than other existing algorithm.

Limitations

- Feedback mechanism increases the computational time of task scheduling.

Paper 6

Title – TTSA: An effective scheduling approach for delay bounded task in hybrid clouds.

Concept

In this paper, task scheduling is performed by temporal task scheduling algorithm (TTSA) which intern dispatches the entire task into both private CDC and public cloud. In each iteration of TTSA, cost minimization problem is solved by a hybrid Simulated Annealing-Particle Optimization Algorithm (SA-PSO). TTSA can efficiently increases the throughput and reduce the cost of private CDC while meeting the delay bounds of all the tasks. This paper mainly considers the cost minimization problem for private CDC in hybrid cloud and solved it using TTSA algorithm.

Limitations

- Simulated annealing convergence speed is very slow and computation time is also more. Performance metrics can be more; it takes only throughput as metrics.

Paper 7

Title – A Hybrid Strategy for Resource Allocation and Load Balancing in Virtualized Data Centers Using BSO Algorithms

Concept

In this paper, BSO algorithm is used to allocate proper resources and balance the load in job scheduling environment of cloud. BSO is combination of both Bacteria Foraging Optimization (BFO) and Particle Swarm Optimization (PSO) algorithm. In this process, jobs are collected as batches and for each job it allocates specific resources. After resource allocation process, Makespan, operational cost and utilizations are calculated. Using these values jobs are scheduled to proper virtual machines.

Limitations

- Allocation of resources to the jobs are taking more time, because if proper resources are not obtained or allocated resources are busy then it will lead to increase in waiting time. Bacteria Foraging Optimization algorithm takes more time to computation.

Paper 8

Title – A novel hybrid of Shortest job first and round Robin with dynamic variable quantum time task scheduling technique

Concept

In this paper, task scheduling is completed using hybrid of shortest job first and round Robin schedulers considered only dynamic variable quantum task. This method splits the ready queue into two types namely Q1 and Q2. Q1 queue is used to store short task whereas Q2 queue is used to store long task. Short and long task are found by obtaining burst time for each task. Using this burst time median value is founded. Median value is used to separate the task into short and long. This process reduces the waiting time and response time.

Limitations

- In this paper, dynamic quantum computation is used to calculate median value which increases the computation time for larger tasks.

Paper 9

Title – Multi-Rumen Anti-Grazing approach of load balancing in cloud network

Concept

In this paper, load is balanced using multi rumen anti grazing approach which consists of three phases. In phase 1, task completion time (TCT) and task completion time with communication delay matrix is calculated using VM capacity and task size. In Phase 2, tasks are assigned to the respective VM, Makespan and TCT values are updated for every iteration. Finally in phase 3 tasks are rescheduled into new VM according to the new TCT value and Makespan value. Using these three phases loads are balanced in cloud network.

Limitations

- In this paper, scheduled tasks are rescheduled again in the virtual machine which increases the computational time and complexity. It works only for non-preemptive independent task.

Paper 10

Title – Proposed Load Balancing Algorithm To Reduce Response Time And Processing Time On Cloud Computing

Concept

In this paper, response time and processing time is reduced using Throttled Modified Algorithm (TMA). TMA load balancer performs load balancing by updating, maintaining two index variable namely busy index and available index. In available index status of VM is '0' whereas in busy index status of VM is '1'. The Data Center Controller sends the request to the specified VM by that ID. The Data Center Controller informs the TMA Load Balancer for a new allocation. If DC receives response from VM, it will notify to TMA controller to update the Available index table.

Paper 11

Title – Scheduling Live Migration of Virtual Machines

Concept

In this paper, mVM new and extensible scheduler introduced which minimizes the completion time effectually. a migration scheduler that relies on realistic migration and network models to compute the best moment to start each migration and the amount of bandwidth to allocate. It also decides which migrations are executed in parallel to provide fast migrations and short completion times. Experiments on a real testbed show mVM outperforms state-of-the-art migration schedulers.

Paper 12

Title – A Load Balancing Algorithm for Resource Allocation in Cloud Computing

Concept

In this paper, Load balancing in virtual machine performed successfully using hybrid combination of Teaching-Learning-Based Optimization (TLBO) and Grey Wolves Optimization algorithms (GWO) which increases the throughput. Exploitation and exploration of GWP is integrated with Scheduler is responsible to allocate jobs on virtual machine in distributed system and provides scheduling for resource allocation. This method increases throughput greater than PSO and BBO algorithms.

Paper 13

Title – A Dynamical and Load-Balanced Flow Scheduling Approach for Big Data Centers in Clouds

Concept

In this paper, dynamical load balanced scheduling approach is proposed for Big data centers in clouds. This process increases the network throughput while balancing workload dynamically. Two representative openflow architectures are implemented namely FPN and FTN. These openflows are dynamically migrates flows which occupies more bandwidth in congested server. Experimentation results shows proposed method outperforms the representatives of Round Robin and LOBUS method.

Paper 14

Title – RIAL: Resource Intensity Aware Load Balancing in Clouds

Concept

In this process, Resource Intensity Aware Load (RIAL) algorithm is used which dynamically allocates different weights to different resources based on their usage intensity which significantly reduces the time and cost to avoid load balance. It avoids unnecessary migrations using strict migration algorithm. Finally, it conducts destination PM selection in decentralized manner to improve scalability. Simulations results show that proposed method outperformed in real world applications.

Paper 15

Title – Stochastic Load Balancing for Virtual Resource Management in Datacenters

Concept

In this paper, stochastic load balancing method is used to improve the virtual resource management in datacenters. Virtual machine migration is minimized using stochastic load balancing scheme. In this process, migration of virtual machine considers distance between source and destination physical machines which interns minimizes the VM migration overhead in load balancing. Proposed method effectively addresses the prediction of distribution of resource demand and multi-dimensional resource requirement.

Paper 16

Title – Cloud Task Scheduling for Load Balancing based on Intelligent Strategy

Concept

In this paper, intelligent strategy is used for load balancing and task scheduling. Minimization of Makespan is performed using Ant Colony Optimization algorithm (ACO) which find the optimum resources for batch of tasks in dynamic environment of cloud computing. It balances the system load while trying to minimizing the Makespan of a given task set. Implementation results from simulation showed that proposed algorithm produces the better results than existing algorithm.

BIBILIOGRAPHY

F. U. Xiao, ``Task scheduling scheme based on sharing mechanism and swarm intelligence optimization algorithm in cloud computing," Comput. Sci., vol. 45, no. 1, pp. 303307, 2018.

C.-G. Wu and L. Wang, ``A multi-model estimation of distribution algorithm for energy efficient scheduling under cloud computing system," J. Parallel Distrib. Comput., vol. 117, pp. 6372, Jul. 2018.

Negar Chitgar ; Hamid Jazayeriy ; Milad Rabiei, “Improving Cloud Computing Performance Using Task Scheduling Method Based on VMs Grouping”, 2019 27th Iranian Conference on Electrical Engineering (ICEE), IEEE, 2019.

H. Aziza and S. Krichen, ``Bi-objective decision support system for task-scheduling based on genetic algorithm in cloud computing," Computing, vol. 100, no. 2, pp. 6591, Feb. 2018.

Zhang Qian, Ge Yufei, Liang Hong, Shi Jin, " A Load Balancing Task Scheduling Algorithm based on Feedback Mechanism for Cloud Computing", International Journal of Grid and Distributed Computing, Vol. 9, No. 4 , pp.41-52, 2016.

Haitao Yuan, Student Member, IEEE, Jing Bi, Member, IEEE, Wei Tan, MengChu Zhou, Fellow, IEEE, Bo Hu Li, and Jianqiang Li, Senior Member, IEEE, "TTSA: An Effective Scheduling Approach for Delay Bounded Tasks in Hybrid Clouds", Volume 47, Issue 11, PP-3658 - 3668 ,2016.

V. Jeyakrishnan, P. Sengottuvelan, "A Hybrid Strategy for Resource Allocation and LoadBalancing in Virtualized Data Centers Using BSO Algorithms", Wireless Personal Communications, Volume 94, Issue 4, pp 2363–2375, 2017.

Samir Elmougy, Shahenda Sarhan and Manar Joundy, "A novel hybrid of Shortest job first and round Robin with dynamic variable quantum time task scheduling technique", Journal of Cloud Computing: Advances, Systems and Applications, Volume 2017, Issue 6, 2017.

Sumanta Chandra Mishra Sharma1 Amiya Kumar Rath, "Multi-Rumen Anti-Grazing approach of load balancing in cloud network", International Journal of Information Technology, Volume 9, Isssue 2, PP- 129-138,2017.

Nguyen Xuan Phi, Cao Trung Tin, Luu Nguyen Ky Thu and Tran Cong Hung, "Proposed Load Balancing Algorithm To Reduce Response Time And Processing Time On Cloud Computing", International Journal of Computer Networks & Communications (IJCNC) Vol.10, No.3, 2018.

Vincent Kherbache, Member, IEEE, E´ ric Madelaine, Member, IEEE, and Fabien Hermenier, Member, IEEE,"Scheduling Live Migration of Virtual Machines", IEEE Transactions on Cloud Computing

Seyedmajid Mousavi(&), Amir Mosavi,and Annamária R. Varkonyi-Koczy, "A Load Balancing Algorithm for Resource Allocation in Cloud Computing", Recent Advances in Technology Research and Education, pp 289-296, 2017.

Feilong Tang Member, IEEE, Laurence T. Yang Senior Member, IEEE, Can Tang, Jie Li Senior Member, IEEE, and Minyi Guo Senior Member, IEEE, "A Dynamical and Load-Balanced Flow Scheduling Approach for Big Data Centers in Clouds", Ieee Transactions On Cloud Computing 2016

Haiying Shen*, Senior Member, IEEE, "RIAL: Resource Intensity Aware Load Balancing in Clouds", IEEE Transactions on Cloud Computing, 2017.

Lei Yu, Liuhua Chen, Zhipeng Cai, Haiying Shen, Yi Liang, Yi Pan, "Stochastic Load Balancing for Virtual Resource Management in Datacenters", Ieee Transactions On Cloud Computing, 2014.

Arabi E. keshk, Ashraf B. El-Sisi, and Medhat A. Tawfeek. "Cloud Task Scheduling for Load Balancing based on Intelligent Strategy**",** Intelligent Systems and Applications, 2014

P. Rimal, M. Maier, Workflow Scheduling in Multi-Tenant Cloud Computing Environments, IEEE Transactions on Parallel and Distributed Systems, Vol. 28, No. 1, pp. 290-304, January, 2017.

Shi, Z. Zhang, T. Robertazzi, Energy-aware Scheduling of Embarrassingly Parallel Jobs and Resource Allocation in Cloud, IEEE Transactions on Parallel and Distributed Systems, Vol. 28, No. 6, pp. 1607-1620, June, 2017.

. Farahnakian, T. Pahikkala, P. Liljeberg, J. Plosila, N. T. Hieu, H. Tenhunen, Energy Aware VM Consolidation in Cloud Data Centers Using Utilization Prediction Model, IEEE Transactions on Cloud Computing, Vol. 7, No. 2, pp. 381-388, October, 2016.

Ulysse Rugwiro, Chunhua Gu, Weichao Ding, "Task Scheduling and Resource Allocation Based on Ant-Colony Optimization and Deep Reinforcement Learning," *Journal of Internet Technology*, vol. 20, no. 5 , pp. 1463-1475, Sep. 2019