**International Journal of Electrical and Computer Engineering (IJECE)**
Vol.9, No.3, June2019, pp. 1637~1644
ISSN: 2088-8708, DOI: 10.11591/ijece.v9i3.pp1637-1644 

# Demand-driven Gaussian window optimization for executing preferred population of jobs in cloud clusters

**Vaidehi M[1], T.R.Gopalakrishnan[2]**
[1]Department of Information Science and Engineering, Dayanandasagar College of Engineering, India
[2]Rajarajeshwari College of Engineering, India

**Article Info**

***Article history:***

Received Jun 6, 2018
Revised Dec 15, 2018
Accepted Dec 29, 2018

***Keywords:***

Cloud computing
Demand-driven
Efficiency
Instantaneous utilization
Jobs
Resource utilization
Scheduling

**ABSTRACT**

Scheduling is one of the essential enabling technique for Cloud computing which facilitates efficient resource utilization among the jobs scheduled for processing. However, it experiences performance overheads due to the inappropriate provisioning of resources to requesting jobs. It is very much essential that the performance of Cloud is accomplished through intelligent scheduling and allocation of resources. In this paper, we propose the application of Gaussian window where jobs of heterogeneous in nature are scheduled in the round-robin fashion on different Cloud clusters. The clusters are heterogeneous in nature having datacenters with varying sever capacity. Performance evaluation results show that the proposed algorithm has enhanced the QoS of the computing model. Allocation of Jobs to specific Clusters has improved the system throughput and has reduced the latency.



***Corresponding Author:***

Vaidehi M,
Departement of Information Science and Engineering,
Dayanandasagar College of Engineering,
Kumaraswamy layout, Bangalore, India.
Email: vaidehim-ise@dayanandasagar.edu

## 1. INTRODUCTION

Generation of a huge volume of data and requirement of high-speed computation with less investment has paved way to fast development of the Cloud technology. There is a complete transformation of computing when compared to the traditional method. This successful model is built using the Grid technology, Virtualization and Distributed computing. The Cloud provides high-speed processors for computation and storage as a service. Though this model is highly used for computation and storage there are several issues to be addressed as Infrastructure-as-service. Due to non-uniform and time-varying workload, the resources required to sustain the workload is also variable [1]. Amazon, Google, etc., invested a considerable amount of money in their data centers as they have to maintain the serversto sustain their peak workload. The average utilization of servers was only 10% [1]. They then realized that merging different workloads with the complimentary usage patterns will enhance the server efficiency and it would be a cost-effective economic model to rent the resources to the public [2]. Amazon launched AWS (Amazon Web Services) utility computing, and after the launch, several IT industries opted for Cloud computing than investing on costly servers [2]. As the demand grew, the computing model started encountering severe challenges like job scheduling and resource allocation [3] exclusively to compute real data.

The service provider has to cater to heterogeneous jobs and not just a cluster of clients whose requests are homogeneous in nature. Challenges are related to flexibility in IaaS. This paper presents a novel approach using the Gaussian window to optimize the execution of preferred population of jobs in cloud clusters. Here the computing model comprises of clusters of four different capacities. Figure 1 presents a

conceptual model for application of Gaussian window. The model includes of two entities, a Job dispatcher and Clusters. The model estimates the required amount of resources for computation of jobs to avoid inappropriate use of the available resources.

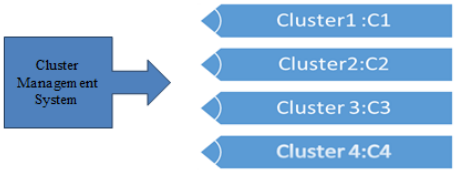


Figure 1. Conceptual model of clusters with a cluster management system

The paper is structured as follows, Section I presents the evolution of the computing model, features and the challenge addressed in this paper. Section II provides a summary of the related literature. Section III provides the details of the conceptual model and the algorithm for allocating the resources to requesting jobs. Large amount of jobs arriving are classified based on properties like size, resource utilization period and computing cost. Performance evaluation of the proposed algorithm is discussed in Section IV. Finally, Section V concludes the paper.

This section briefs about the background of the proposed sytem emphasizing on significant research work carried out to enhance the quality of sevice in the cloud system. Scheduling and Resource allocation in the heterogeneous cloud environment is an open research challenge where many researchers and academicians are working towards to enhance the performance of the computing model. Form the survey it is observed that the scheduling techniques has an impact on computation cost, resource utilization time, energy utilization and QoS, efficient scheduling also reduces the job rejection ratio. T. R. Gopalakrishnan Nair et al., in their work say that it is essential to identify the trends of different request streams in every category by auto classification and organize pre-allocation of resources in a predictive way to reduce the number of jobs being rejected and also reduction in cost per task completion [4]. M. Mezmaz et al. in their work has investigated the problem of scheduling the problem precedence -constrained parallel applications on the heterogeneous computing system (Cloud computing), in their work they have proposed a new parallel bi-objective hybrid genetic algorithm that takes into account the task completion time and also minimized energy consumption [5]. Mingsong Chen et al., say that due to the existence of resource variations, it is a challenge for cloud workflow resource allocation strategies to guarantee a reliable QoS. They also say that it is hard to predict their performance under variations because of lack of accurate modelling and evaluation methods [6].

Hsu Mon Kyi et al., say that in cloud computing systems scheduling and allocation of virtual resources and virtual machine are challenges. To address this issue, they have proposed an algorithm which provides effective and efficient resource allocation. They have used Stochastic Markov model to measure the scalability and tractability of infrastructure resource in private clouds. Their contribution has focused on enhancing the system performance by enabling the response time [7] Wexin Li et al., have proposed a joint optimization model, this chooses the request allocation policy such that the provider gains high bandwidth utilization at its datacenters, and each user experiences a low delay [8]. Pandaba et al., in their work say that cloud infrastructure comprises of several datacenters, and the customer need a slice of the computational power over a scalable network. They say delivery of resources are done in an elastic way. The challenge investigated by them is the wait time experienced by the customers. The researchers have proposed a modified Round Robin algorithm that reduces the wait time, thereby improving the performance [9] Mubarak et al, have made a study on task scheduling algorithms. The researchers have enhanced the Min-Min algorithm to enhance the task completion period. The authors say that, through the experimental analysis, the proposed algorithm has produced a better Make span and improved resources utilization [10]. Stefano Marrone et.al, have proposed a model-driven approach for the automatic negotiation and resource allocation for availability of critical cloud services. The authors have used Bayesian network to evaluate the availability of resources for critical services [11].

Though many research has been carried out to enhance the resource utilization in the cloud environment, there few areas to be focused to increase the quality of service in cloud environment.
The cureent systems less focus on approaches to enhance

1. Latency of the computing model
2. Throughput

Improving these two parameters would enhance better utilization of resources in cloud environment

This work focuses on enhancing the system performance by optimal utilization of the available resources. To achieve this, we propose a conceptual model as shown in Figure 1. The model comprises of clusters which is a set of machines packed into racks. These clusters are connected with high bandwidth cluster network. Clusters are managed by the Cluster management system which allocates jobs to machines. Jobs generally have a set of resource requirement for scheduling or packing the tasks in the machine either for storage or execution. The Gaussian window model is applied to the proposed model to accomplish best utilization of available resources. The Gaussian window is applied to different Clusters.
The machines are clustered depending on different data handling and processing speed.

***Limitations in the existing Resource Allocation Model***

Many of the researchers working on scheduling and resource allocation have proposed new algorithms and computing model but their work does not emphasis much on clustering the machines according to computation or storage requirements.

***Application and working of the algorithm***

The proposed algorithm is applied to a set of sample data shown Table 1. The table comprises, of JobID, Time stamp assigned, size of the job and Machine ID assigned to each computing machine and finally the clusters formed with an Id assigned.

Table 1. Parameters Related to Jobs Scheduled

| Sl.no. | Job ID | Time Stamp | Job Size | Machine Id (Static IP) | Cluster Id |
|---|---|---|---|---|---|
| 1 | $J_1$ | T1 | 600KB | 200.168.2.2 | Cluster02 |
| 2 | $J_2$ | T2 | 500MB | 199.170.5.1 | Cluster01 |
| 3 | $J_3$ | T3 | 600KB | 200.168.2.3 | Cluster02 |
| 4 | $J_4$ | T4 | 200KB | 200.168.2.4 | Cluster02 |
| 5 | $J_5$ | T5 | 700MB | 199.170.5.3 | Cluster01 |
| 6 | $J_6$ | T6 | 299KB | 200.168.2.5 | Cluster01 |

The Cluster management system (CMS) classifies the jobs based on the resource requirement for processing and then distributes to appropriate clusters. The Round-robin scheduling is used for the execution of the jobs in clusters.

Initially, the jobs are classified by CMS as a *free priority, production priorityand monitor priority jobs.*The free priority jobs use minimum resources for computation, and the computational cost is comparatively low, the production priority jobs have the highest priority, the CMS sees to that these jobs are not denied of the requested resources, and they are also not allocated to overloaded machines. This ensures that load balancing is taken care of the proposed model. The free priority jobs are taken care of by the monitor priority jobs to ensure resources.Each job has a time stamp ($t_s$), job Id ($J_i$) and a comparison operator. The comparison operator is greater than or less than.

**Resources and Units**

CPU-number of cores/second
Memory-bytes
Disk Space-bytes
Disk time fraction (I/O in seconds/ seconds) [12]

## 2. ALGORITHM IMPLEMENTATION

a. Assumption: with reference to the proposed model in Figure 1, the clusters 1to 4 are C1, C2, C3 and C4
Assumptions:

1. C1: data handling capability in Petabytes
   C2: data handling capability in Terabytes
   C3: data handling capability in Gigabytes
   C4: data handling capability in Megabytes
2. Job Id=1,2,3, assigned based on their arrival rate as $J_1$, $J_2$, $J_3$ ….. $J_n$
3. Synchronization among the Clusters
4. Cluster failure taken care automatically

*Step 1: $Job_1$ if resource requirement in Petabytes size*
then
*Step2: Check for resources available at that instance of arrival*
*If available schedule*
*Else*
*Step 3: Check for resources in C2 and C3*

*If available*
*Step 4: Split the job in terabytes size and gigabytes size*
*Sync*
*C2 and C3*
*Schedule*
*Step 5: Else block the Job$_1$*
*Step 6: Check wait time of Job$_1$*
*and check resource available*
*If available*
*Step 7: Unblock Job$_1$ and schedule*
Repeat for all jobs

b. Application of Gaussian window

In this paper, we implement the Gaussian window to a set of jobs. The unique property of the algorithm is that just the previous history of the job size [8] is sufficient to predict the resource requirement for the current job. Knowing the Mean "μ" the average required resources, and the standard deviation $\sigma^2$which indicates the actual resource utilized by the job earlier will enable the CMS to allocate resource for the current job in execution without wasting time for computation of resource requirement. Hence this approach also will increase the throughput of the system.

In our research, we use the information of the resource size allocated for execution of the previous job.

c. Gaussian window for prediction

Gaussian distribution is a powerful tool applied expansively to problems like regression and classification. In the Gaussian process a prior probability distribution is assumed initially over the various functions possible to describe the process of generating data, and a posterior probability distribution is obtained after gaining knowledge about the observed values. The posterior improves the knowledge of the observed over the prior.

$$G[z] = a\, e^{-y}$$

Where $y=\frac{[x-b]^2}{2c^2}$

With respect to Gaussian window
The parameters
a= defines the requirement of the resources
x= The actual allocated resource
$c^2$=defines the real resources utilized
b= the average resource required

d. Simulation setup and prediction of results

To implement the proposed algorithm to achieve the results, heterogeneous Clusters along with a communication network using a TCP protocol was built. The Clusters were created using VMware. Wireshark tool was used to monitor the Data transmission among to the Clusters.

## 3. RESULT ANALYSIS

Three different cases were considered to prove the application of the proposed algorithm, here

a. Case I: Best Utilization of available resources by requested Jobs

Figure 2 depicts the best utilization of the allocated resources; here the proposed algorithm classifies the jobs as per the resources requested. It can be observed that the area under the curve shows the utilization of the allocated resources.

The application of the Gaussian window has also enabled to enhance the latency and throughput of the system. The Figure 3 and Figure 5 illustrates the latency and Figure 4 illustrates the throughput. The latency period. With reference to Figure 3 timestamp "0ns" depicts the start of execution of a task, the task waits approximately for "6ns" to load on to the machine and completes the execution in 556ns.

Before scheduling, clustering of the jobs based on their job size based on the proposed algorithm the latency of the system is comparatively improved. Figure 6 represents the throughput of the proposed system. It is observed that scheduling the jobs to appropriate clusters has reduced the job rejections and starvation. The scheduled jobs have fully utilized the allocated resources.

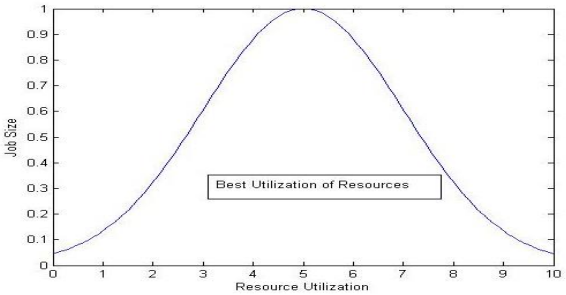


Figure 2. Optimal resource utilization

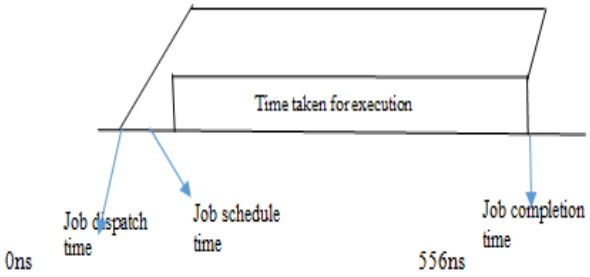


Figure 3. Latency of the proposed system

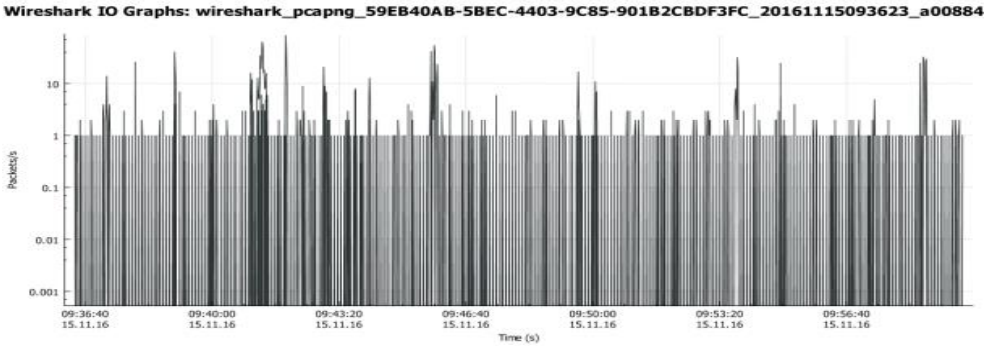


Figure 4. Resource utilization by the allocated jobs in appropriate clusters

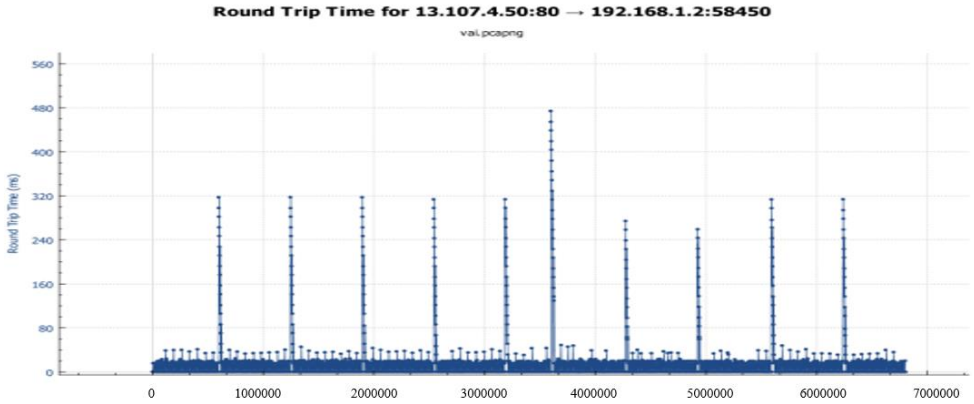


Figure 5. Roundtrip from the proposed system

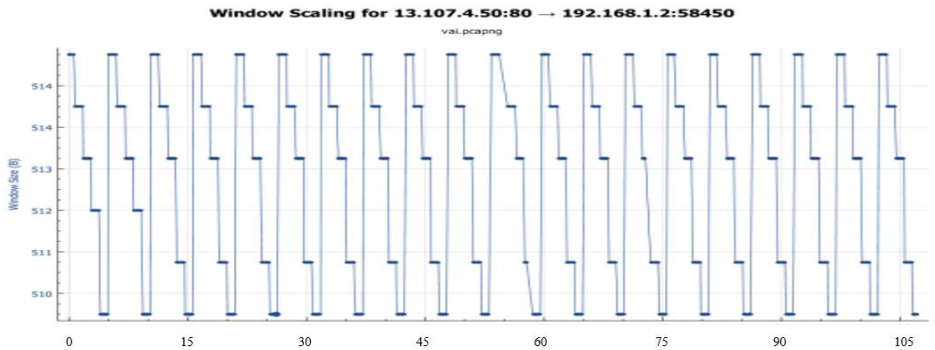


Figure 6. Throughput of the system when job Size is proportionate to the requested resource

b. Case II: Underutilization of Resources when job size is less than the resource allocated

Figure 7 shows the utilization of resources by jobs allocated to machines. In this case the actual resource required for computation is not estimated prior to scheduling. The area under the curve depicts the utilization of allocated resources. Though the computation is complete, the resources are not fully utilized. The throughput of the system is reduced by 15-20%. Figure 8 shows the average reduce in throughput of the system.We can observe that though the computation is completed the allocated resources are not completely utilized.

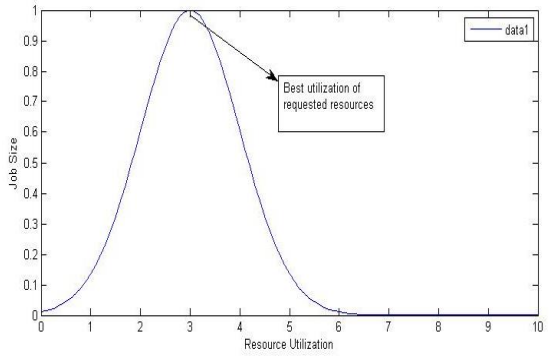


Figure 7. Underutilization of allocated resources

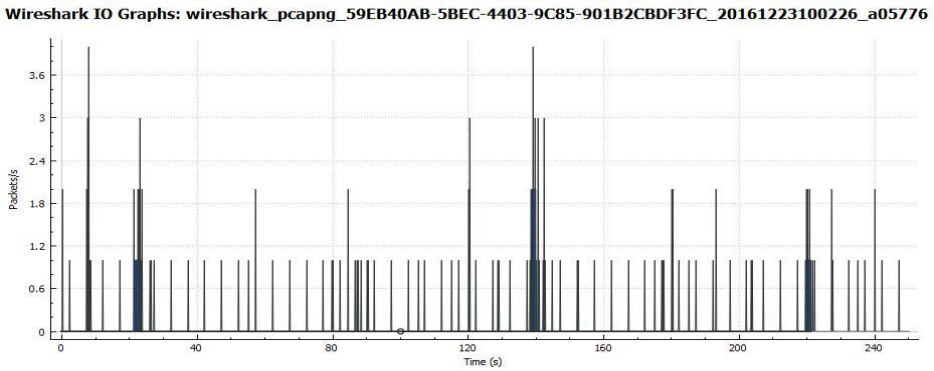


Figure 8. Throughput of the system when job size is less than the allocated resource size

c. Case III: Incompletion of execution due to deficit of required resources: The machines are not clustered and resources are allocated without estimating the required resources for computation

In this case, Figure 9 represents resource utilization where the job size is greater than the resources allocated. Here the resources are allocated to a traditional computing model where the resources are not clustered.

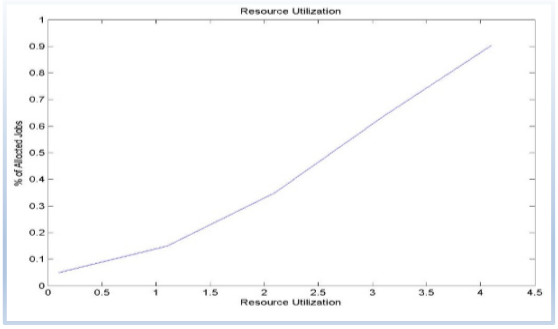


Figure 9. Deficit of resources to allocated jobs

## 4. CONCLUSION

Economic data computing and storage has paved way for IT industries and researchers to improve the existing system by reframing and redesigning the existing technology. In our research work we proposed a Conceptual model of Future Cloud Cluster.The application of the Gaussian window to this model has enhanced the system performance. From the results it can be observed that the allocated resources are best utilized, the proposed algorithm also prevents resource contention.The simulation results show the improvement in throughput and reduce in latency. Further the proposed system can support optimal energy utilization which will be carried as an extension.

## REFERENCES


[1] Saijie Huang, Mingsong Chen, Xiao Liu, Dehui Du, Xiaohong Chen, "Variation-Aware Resource Allocation Evaluation for Cloud Workflows Using Statistical Model Checking," *2014 IEEE International Conference on Big Data and Cloud Computing (BdCloud)*, vol. 01, no. pp. 201-208, 2014, doi:10.1109/BDCloud, 48, 2014.

[2] T. R. Gopalakrishnan Nair, M. Vaidehi, "Efficient Resource Arbitration and Allocation Strategies in Cloud Computing through Virtualization," *2011 International Conference on Cloud Computing and Intelligence Systems (CCIS)*, China, IEEE Xplore 2011.

[3] T. R. Gopalakrishnan Nair, P. Jayarekha, "Pre-allocation Strategies of Computational Resources in Cloud Computing using adaptive Resonance Theory-2," *International Journal on Cloud Computing: Services and Architecture(IJCCSA)*, Vol.1, No.2, August 2011.

[4] M.Mezmaz, N.Melab, Y.Kessaci, Y.C.Lee, E.G. Talbi, A.Y.Zomaya, D.Tuyttens, "A Parallel Bi-objective Hybrid Metaheuristic for Energy-aware Scheduling for Cloud Computing Systems," *Elsevier, Journal of Parallel and Distributed Computing 71(2011)* , pp-1497-1508, 2011.

[5] Mingsong Chen, Saijie Huang, Xin Fu, Xiao Liu, Jifeng He, "Statistical Model Checking -Based Evaluation and Optimization for Cloud Workflow Resource Allocation," *IEEE Transactions on Cloud Computing*, Vol. pp-99, doi: 10.1109/TCC.2016.2586067, 2016.

[6] Hsu Mon Kyi and Thinn Thu Naing "Stochastic Markov Model Approach For Efficient Virtual Machines Scheduling On Private Cloud," *International Journal on Cloud Computing Services and Architecture (IJCCSA)*, Vol.1, No.3, November 2011.

[7] Wenxin, Heng Qi, Kegiu Li, Julong Lan, "Joint optimization of Bandwidth for Provider and Delay for User in Software Defined Data Centers," *IEEE Transactions on Cloud Computing*, Vol 5, No 2, April-June 2017.

[8] Pandaba Pradhan, Prafulla Ku. Behera, B.N.B. Ray "Modified Round Robin Algorithm for Resource Allocation in Cloud Computing," Procedia Computer Science, ISSN: 1877-0509, Vol : 85, PP: 878-890, 2016.

[9] Mubarak Haladu, "Optimizing Task Scheduling and Resource allocation in Cloud Data Center using Enhanced Min-Min Algorithm," *IOSR Journal of Computer Engineering, (IOSR-JCE)*, Vol: 18, PP 18-25, 2016.

[10] Stefano Marrone, Roberto Nardone, "Automatic Resource Allocation for High availability Cloud Services," *2nd International Workshop on Computational Antifragility and Antifragile Engineering 2015*, 2015.

[11] Hongbing Wang, Zuling Kang, Lie Wang, "Performance-Aware Cloud Resource Allocation Via Fitness- Enabled Auction," *IEEE Transactions on Parallel and Distributed Systems*, Vol. 27, No.4, April 2016 .

[12] Mohit Dhingra, J.Lakshmi, S.K.Nandy, Chiranjib Bhattacharaya, K.Gopinath, "Elastic Resources Frame Work for IaaS, Preserving Performance SLA's," *Proceedings, 6th International Conference on Cloud Computing*, Santa Clara, California, US, June 2013.

[13] Guilherme Da Cunha Rodrigues, Rodrigo N. Calheiros, Vinicus Tavares Guimaraes, Glederson Lessa dos Santos, Marcio Barbosa de Carvalho, Lisandro Zambenedeti Granville, Liane Margarida Rockenbach Tarouco, Rajkumar Buyya", Symposium on Applied Computing, ACM, Pisa Italy 2016.

[14] Antony Thomas , Krishnalal G, Jagathy Ray V.P, "Credit Based Scheduling Algorithm in Cloud Computing Environment," *International Conference and Communication Technologies (ICICT 2014)* doi:10.1016/j.procs.2015.02.162 Science Direct, 2014.

[15] Narendra Kumar, Swati Saxena, "A Preference-based Resource Allocation in Cloud Computing Systems," *3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015)*, doi: 10.1016/j.procs. 2015 .07.375 Science Direct, 2015.

[16] Alexander Ngenzi , Selvarani R , Suchithra R, "FDMC: Framework for Decision Making in Cloud for Efficient Resource Management," *International Journal of Electrical and Computer Engineering (IJECE)*, Vol. 7, No. 1, , pp. 496~504 ISSN: 2088-8708, DOI: 10.11591/ijece.v7i1.pp496-50, February 2017.

**BIOGRAPHIES OF AUTHORS**

**Mrs.Vaidehi M** obtianed B.E. (Electronics and Communication) and M.E. (Information Technology) degrees from Bangalore University in the year 2001 and 2007 respectively. She is persuing her Ph.D in Visweswaray Technological University, Karnataka, Indai. She is currently working as Asst.Professor in Department of Information Science and Engineering, Dayanandasagar College of EngineeringBangalore. The author has served as Research Associate in Research Industry and Incubation Center, DSI. Her Research Interests are Cloud Computing, Computer Networks.
dm.vaidehi@gmail.com, vaidehim-ise@dayanandasagar.edu

**Dr. T.R. Gopalakrishnan Nair**, obtained his M.Tech degree from Indian Institute Science, Bangalore and his Doctorate in Computer Science and Engineering. He is a member in several professional bodies like IEEE, ACM, CSI, etc. Currently he is serving as a Rector in RRGroup of Institutions , Bangalore. He has served in various capacities as Senior Scientist, Indian Space Research ( PARAM Awardee), Vice President Research, DS Institutions, Bangalore, Visiting Research Professor, University of Ulster, UK. Former Saudi Aramco Endowed Chair, Technology and Information Management, PMU. KSA- 2015
trgnair@gmail.com, www.trgnair.org