I.J. Modern Education and Computer Science, **2017, 2, 27-33**
Published Online February 2017 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2017.02.03

Resource Allocation Strategy with Lease Policy and Dynamic Load Balancing

Pooja S. Kshirsagar
CSE dept., Walchand Institute of Technology, Solapur, 413006, India
Email: 5407pujak@gmail.com

Prof Anita M. Pujar
CSE dept., Walchand Institute of Technology, Solapur, 413006, India
Email: arkulkarni10@gmail.com

Abstract**—**Cloud Computing has managed to attract the entire buzz in the growing era of technology due to its on-demand services for resource request. Despite of the enormous growth of cloud computing, there are many problems related to resource allocation in cloud that are still unaddressed. Current work for resource allocation strategy focuses on various methods to place Virtual Machine per appropriate requests. The current state of art focuses on the dynamic nature of the work load on cloud. But there is still scope of improvement in the resource allocation strategies that have been proposed in terms of well-balanced network even at the resource contention.
This study proposes a hybrid model composed of lease methodology and dynamic load balancing algorithm, with an attempt to overcome the problems of resource contention and starvation and a well-balanced network even at the input of varying loads. An attempt to increase the CPU utilization and throughput along with no request rejection is taken. The work also retains the lease options for its clients thus maintaining the anti-starvation for pre-emptible requests.

Index Terms**—**Cloud computing, ACWN, haizea, load balancing, VM scheduling.

I. Introduction

Cloud computing is a popular computational model to process applications that are computationally intensive and data, allowing a pay-as-use pattern. The difficulty of efficiently allocating resources according to the user requests has increased tremendously due to the increasing demand for cloud based applications which compelled to satisfy the service level agreements between the consumers and the service providers.[3] Furthermore, heterogeneous nature of the cloud resources, the ever changing nature of workload, and the different objectives of different cloud actors further complicate resource allocation in the cloud computing environment. Meeting both producers and consumers demands make resource allocation strategy even more non-trivial. This paper gives a brief description about the resource allocation in cloud. In this paper, we first go through the basic idea of what resource allocation is and its importance in the cloud environment. Further different types of load balancing methodologies are discussed and this is followed by the various types of resource allocation strategies. This paper also discusses various methodologies given by various authors for resource allocation strategy. Finally, we elaborate our proposed system in last section which is a hybrid model of lease methodology as well as dynamic load balancing. [1, 2]

A. Background

This section elaborates the basic concepts and roles involved in resource allocation strategy, load balancing algorithms and its need in cloud computing.

1. Resource allocation in cloud computing:

Cloud resource allocation is the combined process of discovering resource, selecting, resource provisioning and job scheduling, and managing of resources. The current state of art focuses on the dynamic nature of the work load on cloud. But there is still scope of improvement in the resource allocation strategies been proposed in terms of well-balanced network even at the resource contention. Furthermore, cloud resource allocation consists of decision making with respect to how much of resources, which type, when to allocate, and where to allocate the existing resources in response to the user's request. [3, 10]

2. Need for resource allocation in cloud computing:

In cloud environment, there are often situations of peak demands and no-demand for resources. Hence to match with these uncertain demands for resources, good Resource Allocation Strategy is required. Resource Allocation Strategy must satisfy both user's perspective as well as the provider's perspective, i.e. meet the user demands and maximize the provider's profit respectively. Thus, to balance the level of supply and demand of resources, Resource Allocation Strategy must be able to handle the following issues regarding the resources: Contention, Fragmentation, Over-provisioning and under-provisioning. [3, 10, 11]

3. Load balancing:

Load balancing is the operation of distributing the work load among the various nodes of a system with the aim of optimizing the response time and resource utilization in the weak and peak times. Improvising the execution of various distributed applications in most type of distributed architectures many of load balancing algorithms are proposed. Resource Allocation Strategy along with a good load balancing algorithm results in a well-balanced system for handling resource requests. To cope up with the unstable growth and decrement of load in dynamic environments, various dynamic algorithms are proposed. [4]

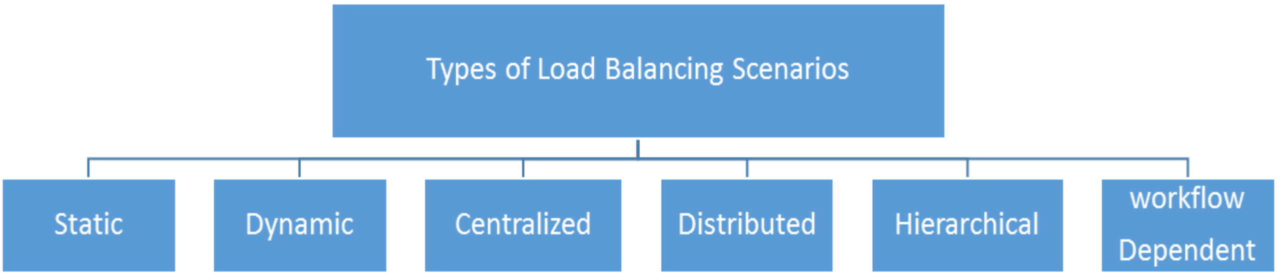

Fig.1. Load Balancing types in different environment[14]

In fig. 1. Different Load balancing types of algorithms are given which vary accrding to the requirement of the environment, usage and scenarios. Description of these types is as follow:

3.1 Static

This type of algorithm requires a prior knowledge about the environment and status of each node and the user requirements in advance. Such algorithms are best suited in homogeneous environment where there is no frequent fluctuation in the user requirements and work load.

3.2 Dynamic

In this type of algorithms run time changes of the node and their load status are monitored and taken into consideration for balancing the load at every moment. It is used in the heterogeneous environments but at the same time such algorithms are complex and time consuming.

3.3 Centralized

Centralized algorithms execute with the help of a central node or server which is responsible to maintain and update the load of the entire network. It is apt for smaller networks with minimum load. But its drawbacks are that it is not fault tolerant. Since the decision making is on the central node, load at peak times make it under-functioning.

3.4 Distributed

All the nodes in the network maintain the load locally and every processor participates in the task of effective load balancing. It is suitable in large and heterogenous scenarios. But it is difficult to construct such algorithms since many attributes are involved in it. To keep all the nodes in connection and work harmoniously communication primitives are used. But at peak time communication overhead slowdowns the system.

3.5 Hierarchical

In this type of algorithm the nodes form hierarchy and higher level nodes communicate with lower level nodes to get the information about the network load. It is suitable for large or medium sized networks which are heterogenous in nature. Despite of its heterogenous nature it is power in fault tolerance and complex in structure.

3.6 Workflow Dependent

In this DAG is used for representing network dependencies and can be further used to make decision for balancing the network. It is suitable for either homogeneous environment or even heterogeneous environment. Such models are difficult and complex to build due to overhead of collection of the database required. [14]

4. Resource Allocation Strategies

Resource allocation strategies can be classified according to the input parameters that are given t the RAS, the services , infrastructure and type of the application demanding the resources. Table 1. depicts the classification of RAS based upon the diifferent parameters required by the cloud infrastructure. Following section dicusses these different strategies of the resource allocation in cloud environment.[2]

4.1. Execution Time

Such strategies consider the execution time of the tasks and their preemption while allocating resources. This model overcomes the issues of resource contention and improves the resource utilization capabilities with the help of different leasing methods for computation. But while doing so errors often occur, due to improper estimation of the execution time of a job by user.

Table 1. Different Resource Allocation Strategies in Cloud Environment [2]

Resource Allocation Strategies	Execution Time	Matchmaking
	Policy	Security, Processor
	VM	Load on the node, cost of the resource, speed of service, type of resource
	Gossip	Peer information, peer resources, expert's knowledge
	Utility	Response Time, Profit, Application Satisfaction
	H/W Resource Dependency	CPU dependent, I/O dependent, Storage dependent, Communication dependent
	Auction	Market bid
	Application	Large scale application, Real time Applications, Data sensitive applications and Shared database application
	SLA	Response time, throughput, Quality of service

4.2. Policy based Resource allocation

When centralized server and the resource mangement startegy fails to manage the resources, user requests and the organization level operations, alternative policy based models can be aopted. One such model is the most-fit policy. It allocates a job to cluster, which later produces the distribution of a leftover processor, thus leading immediate job allocations subsequently every time. To find the target cluster, many complex searching processes are taken up. It requires that clusters need to be homogeneous and distributed. The number of processors in every cluster is also predefined. Migration of job takes place when load sharing activities occur. But from the Experimental results, it is depicted that this policy has higher time complexities but are negligible compared to the overall system long time operations. This policy can be practically used in real systems.

4.3. VM

With availability of VM in an infrastructure, the system is capable of live migration of jobs across the physical infrastructure of multiple domains. Due to the dynamic demands of both the infrastructure as well as the user resource requirements, a virtual computation environment is capable of automatically relocating across the infrastructure and thus scale its resources.

4.4. Gossip

A protocol for resource allocation based on this policy is well suited in largescale cloud environments. In this a key function is performed within distributed architecture for big clouds. It is assumed that every node has a specific CPU and a memory capacity. This protocol applies a distributed method that allocates resources in the cloud to a group of applications that demand time dependent memory space and it also maximizes a global cloud utility function dynamically. Gossip based protocol can also be induced with co-operative VM management and cost management. With the help of this method, the organizations can cooperate with each other to share the existing resources to reduce the cost. For such systems, public and private cloud environments are considered.

4.5. Utility

In a system, by optimizing some functions such as minimizing cost function, increasing system performance function and incrementing the QoS, we can improve the working of a dynamically dependent system. This optimizing function is defined as Utility property, which is determined as per the response time, profit, number of QoS, targets met etc. This utility function may vary from system to system as per the demand of the environment and user needs.

4.6. Hardware Resource Dependency

To enhance the use of the system hardware, this model has a scheduler named Multiple job scheduler which allocates multiple job to the existing resources efficiently. The classification factors for scheduling of jobs may be disk I/O, network I/O bound, memory bound and CPU bound jobs. The scheduler is responsible to detect the type of jobs into different categories. Based on their categories, resource allocation will be done.

4.7. Auction

This methodology addresses resource allocation in cloud with the help of auction mechanism. One such mechanism is sealed-bid auction. In this the cloud service provider is responsible for collecting all users' bids and determines their price. The resource is divided to the 1st kth highest bidders with the price of the (k+1)th highest bid. Such system converts the resource problem into the ordering problem thus simplifying the service provider's decision and allocation. The focus of this allocation strategy is to leverage the profits of both resource provider and the customer in large clouds by maintaining the balance in the demand and supply in the market.

4.8. Application

This methodology allocates resources depending upon the nature of the application. For the application, which are workflow based, Virtual machine allocation strategies are designed where resource allocation is based on the workflow of the application. this methodology helps to estimate the exact amount of resources required by the user. For allocating resources and scheduling strategies such as Naive, services group optimization, FIFO, Optimized are designed.

4.9. SLA

In SaaS providing clouds, SLA providers are not that highly developed. Thus, to satisfy the objectives of these providers, specialised resource allocation strategies are introduced to satisfy the constraints of SaaS Clouds. One of the biggest advantage of emergence of SaaS is that, applications are more web based than pc based. The focus of this RAS for SaaS is to provide customer benefits. [2]

B. Haizea

It is an open source, VM scheduler that works on the lease framework. It can be used as a backend for scheduling. Its basic methodology is to allocate resource on complex lease terms instead of just directly allocating a VM to start-up immediately. It provides the user with 4 types of lease options for resource request namely, Best Effort (BE), Advanced Reservation (AR), Immediate and Dead Line Sensitive (DLS). Among which DLS and BE are pre-emptible and AR and immediate are non-preemptible.BE allows First come first service strategy but can be pre-empted at the arrival of higher priority jobs are like BE but with a deadline associated with job within which it must be completed. Immediate type requires resource at the arrival else the request is rejected.AR type of lease is required by multi-level applications requiring number of resources in advance for their execution in later stage. If resources are not available, then the request is dropped. Since BE type requests are pre-empted always at the arrival of AR and scarcity of resources, they face the problem of starvation. Some studies have focused on execution of high priority jobs over low priority ones. Starvation is the major drawback of these studies [11]. Resource allocation strategy must attempt to handle this type of starvation and imbalance in the distribution of the jobs and resources. Further this framework is flexible enough to be plugged with other frameworks for further extensions [1]

II. Related Work

Following section gives a brief description about various studies in the field of resource allocation, various strategies, different algorithms adopted by various authors to handle the issue of resource allocation in the cloud environment:

HebaKurdi, EbtesamAloboud, Sarah Alhassan, Ebtehal T. Alotaibi, Elseveer-2014 [1] have addressed the problem of starvation of BE leases discussed above. They have proposed an anti-starvation algorithm by providing the negotiation constraints along with the threshold limit for the aging counter of the BE leases. Their work has also attempted to minimize the AR rejections. Experimental results show a considerable improvement in the CPU utilization and reduced AR leases as compared to the standard model. Similar approach is attempted by Kumar, Narander, and Swati Saxena (Elsevier 2015) [9] in their work but with preference based approach. Their work involves bidding for resources and options for payment for using cloud services. The work aims at beneficiating both cloud users and providers.

Marwah Hashim Eawna, Salma Hamdy Mohammed, El-Sayed M. El-Horbaty, (2015) [6] have designed a new methodology for resource provisioning in multi-tier clouds. The study incorporates two algorithms for resource handling namely, PSO and SA in its hybrid model. The pitfalls of both the algorithms are compensated in the new hybrid model. The simulation results show that the hybrid algorithm takes less execution time compared to the individual algorithms in the multi-tier architecture.

Endo, P., de Almeida Palhares, A., Pereira, N., Goncalves, G., Sadok, D., Kelner, J., Melander, B. and Mangs, [4] here in this research work authors have discussed various issues in the resource allocation strategy along with the various concepts and tools regarding the cloud paradigm. This study has mentioned various definitions of the cloud concepts along with their examples. Plus, various problems related to the resources allocation strategy along with their solutions is elaborated in this work. Another highlighting sector which helped my thesis, mentioned in this work was of mediation a system that is a framework which is divided into 3 layers for handling the resource allocation. Two major open source architecture of existing mediation systems are also discussed.

In [12], Sudeep R and Guruprasad H S have surveyed on the Cloud Resource Allocation Strategies adopted by various authors. Their study includes the different techniques of resource allocation in cloud including some stochastic models, Market analysing frameworks. In addition to this, different load balancing approach have been described, to balance the load while allocating the resources in effective manner and handle the dynamic and real time load in the cloud environment.

Reducing the computational cost is the major focus of authors in [15]. Selecting appropriate VM for the precise execution is the major focus of the RAS in the work by Shahdi-Pashaki, S. et al. Therefore, to reduce the computational cost, a new mathematical model called group technology is introduced. The large-scale problems are eradicated using the cuckoo algorithm.

In [13], Wolke et al. have described the importance of why reproducibility of experiments is important and what are the difficulties in doing so. They have also added that it is very difficult to replicate a model and reuse it due to the various configuration parameters and detail0ts involved. Such processes are time consuming and complex.

Chandrasekhar S. Pawar, Rajnikant B. Wagh, IEEE, 2013[5], offer a brief work on dynamic resource allocation mechanisms for pre-emptible tasks in cloud environment. They have proposed an algorithm which is based on priority that has considered various SLA objectives of jobs. Two algorithms PBSA and CMMS are compared with the help of simulation in the resource contention scenario.

V. Vinothina, Dr. R. Sridaran, Dr. Padmavathi Ganapathi IJACSA 3.6 (2012) [2], have summarized in detail on various Resource Allocation Strategies for cloud infrastructures, including well defined descriptions of all the techniques and their impacts on cloud. Also, study shows the merits and demerits of Resource Allocation Strategy. An overview on various strategies and their different application in their suitable scenarios is also mentioned. At the end a conclusion is made on the importance of standardization that is required to be incorporated in the various existing methodologies of resource allocation.

III. Proposed Idea

This study proposes a hybrid model composed of lease methodology for VM allocation and dynamic load balancing algorithm, with an attempt to overcome the problems of resource contention and starvation and a well-balanced network even at the input of varying loads. [1] The focus of this research work is to propose a hybrid model for handling resource allocation strategy. Workload is synthetically generated and given to the proposed hybrid algorithm for handling resource allocation in public cloud environment. An attempt to provide improvised version of the lease methodology of haizea framework along with load balancing approach is taken. For Load balancing Adaptive Contracting with the neighbor algorithm is chosen.

A. System Architecture:

This Study is designed for public cloud which provides Infrastructure as a service paradigm and has multiple clusters and each cluster having multiple autonomous nodes i.e. physical machine as depicted in fig. 2. Every Physical Machine has 3 components VM manager, Job allocator and Load estimator. Fig 3 shows the functional structure of each physical node. Similar approach with skewness estimation is studied by Nagpur Mahesh B, et al. in their work [8]. VM Manager will be responsible for maintaining the queue of currently allocated VMs. Job allocator is responsible for allocating the lease to the clients and mapping this lease to the VMs later. Load estimation layer maintains the current load status of itself as well as its neighboring node.

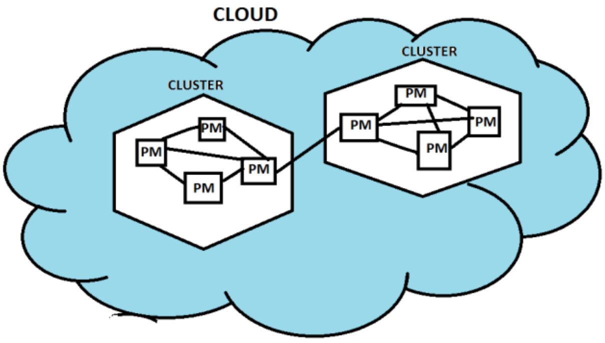

Fig.2. System architecture

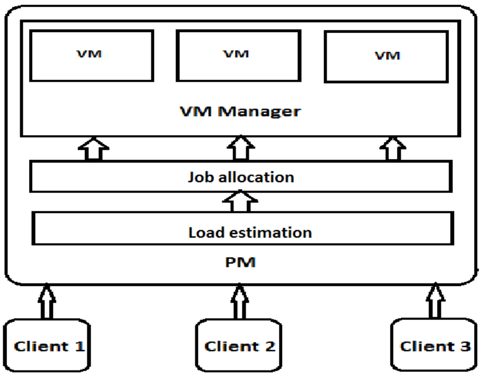

Fig.3. Schematic of every physical machine in the cluster [8]

B. Functional architecture

The Proposed hybrid model composes of two parts of work: one is that of allocating the leases to the clients and two is the load balancing of the system when there is peak time of client requests at a node. Fig 4. Elaborates the methodology of proposed hybrid framework. When client approaches a node with resource request, following methodology will be adopted to solve the request. First check will be made for the availability of VM at that node. If so it allocates it with required lease option to the client. Else it will migrate to that neighbouring node which has load less than threshold value. And then VM for required lease request will be allocated. Once the VM is allocated the load count will be updated at the current node as well as its neighbouring nodes. Here Adaptive Contracting with Neighbours (ACWN) is used for dynamic load balancing. With ACWN every node in the cluster maintains its own current load as well as the load of its neighbours. After applying the load balancing algorithm along with the lease policy; the results are expected to have no AR rejections simultaneously maintaining high CPU utilization and a well-balanced network.

IV. Conclusion

With limited number of resources in a cloud and multiple client requests, VM technology was introduced for efficient allocation strategy [11]. Various approach negotiating for effective resource allocation are attempted. This paper proposes a hybrid framework for resource allocation strategy for clouds with multiple clusters and autonomous nodes. The framework will handle the starvation of pre-emptible requests and will try to maintain a balance in the network by distributing the load using ACWN algorithm for load balancing. The aim is to utilize available resources of the cloud efficiently and allocate the VM to satisfy every client request. Aim is to increase the CPU utilization of the system and avoid the request rejections by providing migration as well as lease options.

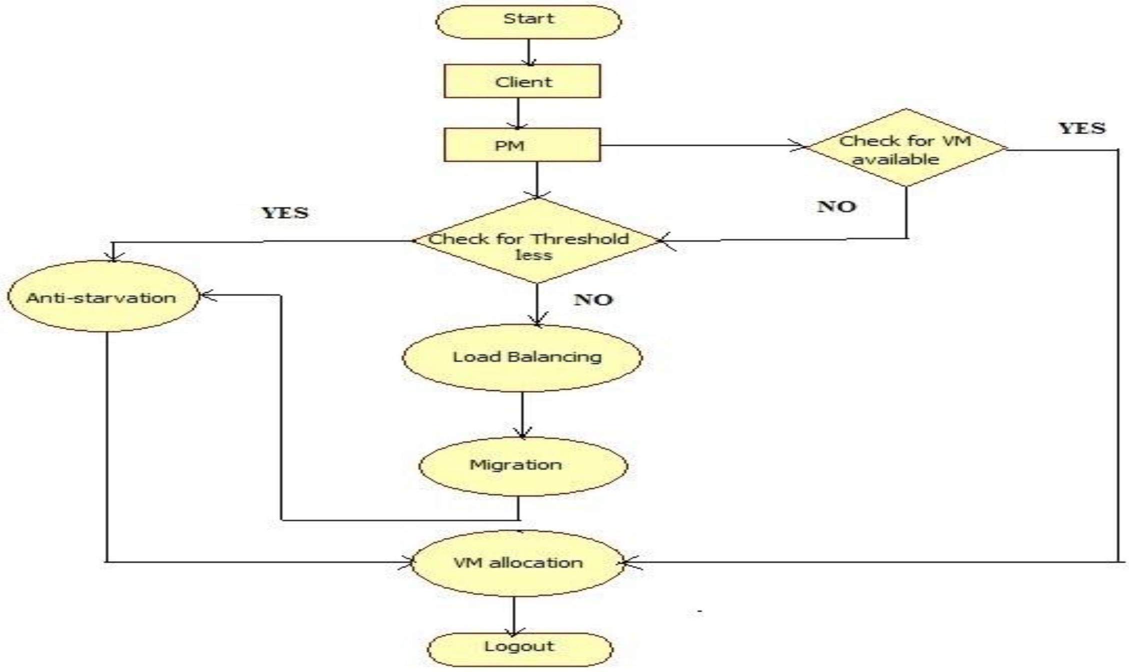

Fig.4. Functional Flow of the hybrid model.

REFERENCES

[1] Kurdi, Heba, Ebtesam Aloboud, Sarah Alhassan, and Ebtehal T. Alotaibi. "An Algorithm for Handling Starvation and Resource Rejection in Public Clouds." Elsevier publication, The 9th International Conference on Future Networks and Communications Procedia Computer Science 34 (2014): 242-48.

[2] V.vinothina, V., Dr. R. sridaran, and Dr. padmavathi Ganapathi. "A Survey on Resource Allocation Strategies in Cloud Computing." International Journal of Advanced Computer Science and Applications IJACSA 3.6 (2012)

[3] Abdullah Yousafzai1, Abdullah Gani, RafidahMd Noor, Mehdi Sookhak, Hamid Talebian, Muhammad Shiraz and Muhammad Khurram Khan. “Cloud resource allocation schemes: review, taxonomy, and opportunities” Springer-Verlag London 2016.

[4] Endo, P., de Almeida Palhares, A., Pereira, N., Goncalves, G., Sadok, D., Kelner, J., Melander, B. and Mangs, J.-E. (2011) ‘Resource allocation for distributed cloud: Concepts and research challenges’, IEEE Network, 25(4), pp. 42–46. Doi: 10.1109/mnet.2011.5958007.

[5] Pawar, Chandrashekhar S., and Rajnikant B. Wagh. "Priority Based Dynamic Resource Allocation in Cloud Computing." IEEE, 2013 International Symposium on Cloud and Services Computing (2013)

[6] Eawna, MarwahHashim, Salma Hamdy Mohammed, and El-Sayed M. El-Horbaty. "Hybrid Algorithm for Resource Provisioning of Multi-Tier Cloud Computing." Elsevier publication, Procedia Computer Science 65 (2015): 682-90.

[7] Capacity Leasing in Cloud Systems Using the Open Nebula Engine (n.d): n. pag. Web.

[8] Nagpure Mahesh B., PrashantDahiwale, and PunamMarbate. "An Efficient Dynamic Resource Allocation Strategy for VM Environment in Cloud.” IEEE, 2015 International Conference on Pervasive Computing (ICPC) (2015)

[9] Kumar, Narander, and Swati Saxena. "A Preference-based Resource Allocation in Cloud Computing Systems." Elsevier publication, Procedia Computer Science 57 (2015): 104-11.

[10] Lin, Weiwei, James Z. Wang, Chen Liang, and Deyu Qi. "A Threshold-based Dynamic Resource Allocation Scheme for Cloud Computing." Elsevier publication, Procedia Engineering 23 (2011): 695-703.

[11] Saraswathi, A.t., Y. R. A. Kalaashri, and S. Padmavathi. "Dynamic Resource Allocation Scheme in Cloud Computing." Elsevier publication, Procedia Computer Science 47 (2015): 30-3

[12] Sudeepa R, Dr. H S Guruprasad. “Resource Allocation in Cloud Computing.” IJMCTR, ISSN: 2321-0850, Volume-2, Issue-4, April 2014.

[13] Wolke, A., Bichler, M., Chirigati, F. and Steeves, V. (2016) ‘Reproducible experiments on dynamic resource allocation in cloud data centers’, Information Systems, 59, pp. 98–101. doi: 10.1016/j.is.2015.12.004.

[14] M Katyal and A Mishra. “A Comparative Study of Load Balancing Algorithms in Cloud Computing Algorithms” IJDCC, Volume 1, Issue 2, December 2013.

[15] Shahdi-Pashaki, S. et al. "Group Technology-Based Model and Cuckoo Optimization Algorithm for Resource Allocation in Cloud Computing". *IFAC-Papers Online* 48.3 (2015): 1140-1145. Web.

Authors' Profiles

Mrs. A M Pujar is working as Assistant Professor, Computer Science and Engg. Dept., Walchand Institute of Technology, Solapur. She has a teaching experience of 18 years. She has completed ME in CSE dept. and Ph.D. in Computer science and engg. (submitted thesis). Her area of interests are Data Mining, NLP and Cloud.

Miss Pooja S Kshirsagar is pursuing Masters in Engg. From Walchand Institute of Engg., Solapur. She has completed her Bachelors in Computer Science and Engg. Her are of Interests are Distributed Systems and cloud.

How to cite this paper: Pooja S. Kshirsagar, Anita M. Pujar,"Resource Allocation Strategy with Lease Policy and Dynamic Load Balancing", International Journal of Modern Education and Computer Science(IJMECS), Vol.9, No.2, pp.27-33, 2017.DOI: 10.5815/ijmecs.2017.02.03