CERTAIN INVESTIGATIONS ON LOAD BALANCING AND RESOURCE MANAGEMENT IN CLOUD COMPUTING ENVIRONMENT

R. Dharani[1], Dr. M. Kalaiarasu[2]

[1]PG Scholar, Computer Science and Engineering,

[2]Associate Professor, Department of Information Technology,

Sri Ramakrishna Engineering College, Coimbatore, Tamilnadu, (India)

ABSTRACT

Cloud computing is a recent advancement technology in IT infrastructure and applications are provided as 'services' to end-users under a payment based on the usage model. It can leverage virtualized services based on requirements of workload patterns varying with time. To evaluate the performance of Cloud provisioning methods, application workload models, and resources performance, the proposed system uses CloudSim an extensible simulation tool kit that enables modeling and simulation of Cloud computing systems. To eliminate the system bottlenecks and to optimize the resource utilization, there is a demand for distributed storage system to employ a workload balancing and efficient resource management framework. The proposed framework implements the workload balancing and resource management, a widely used and typical distributed storage system on cloud. The framework was designed to monitor the workload and analysis for discovering overloaded which reduces the system performance and under loaded nodes in the cluster. To balance the workload among the nodes Split, Merge algorithm and replication methods are implemented to regulate virtual machines on cloud. To demonstrate its effectiveness, experiments are conducted and evaluated. And the experimental results and performance evaluation show that the framework can achieve the goals of efficient management of resources and balancing the workloads.

Keywords: Distributed storage system, Load balancing, Resource management, Resource utilization, Virtualization.

I. INTRODUCTION

Cloud computing can transform the business by offering new options for businesses to increase efficiencies while reducing costs. Cloud lets user can access all applications and documents from anywhere in the world, freeing from the enclosing of desktop and making it easier for group members to collaborate in different locations. Load balancing is one of the issues which cloud computing faces [1]. With the increasing popularity which was gained by cloud computing systems, cloud providers have built several ultra scale data centers at a variety of geographical locations, including hundreds of thousands of computing servers [2]. It is a model for enabling convenient, on-demand network access to configure shared pool of resource and reliable computing resources (e.g., networks, servers, storage) that can be provisioned rapidly and released with minimal consumer management.

The server-centric vision was the traditional approach to IT infrastructure, where IT is responsible for managing, maintaining, procuring, designing, deploying, and troubleshooting servers hosted on the company's premises or located at the organization's central datacenter. A data center can also be referenced to a centralized repository, either for physical or virtual environment, for the storage, management, and dissemination of data and also for information organized around a particular body of knowledge or pertaining to a particular business.

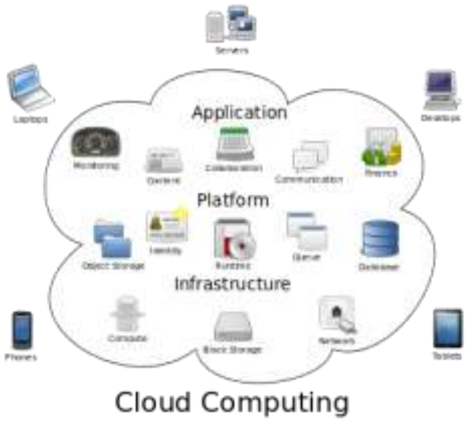

Fig. 1. Cloud Architecture

As the number of concurrent user's increases, scalability becomes one of the major challenges in designing an interactive system [4]. The efficiency can be increased by Virtualization approach by allowing consolidation of server workloads to reduce cost and increase system utilization, but even a virtualized datacenter still has a server-centric infrastructure which requires high degree of management overhead. The bottom line here is that businesses considering making the transition to the cloud need to have an understanding of IT from two perspectives: the type of sourcing and the kinds of services being consumed.

1.1. load balancing in cloud computing

Load balancing in cloud computing provides an efficient solution to various issues residing in set-up and usage in cloud computing environment. Cloud computing is deployed in the data centre where virtualized physical machine are available. Cloud computing being the new technology has both advantages and disadvantages; load balancing is one of the issues which cloud computing faces. Load balancing must take into account two major tasks; one is the resource provisioning and resource allocation in distributed environment. Cloud computing providers are unable to predict geographic location for distribution of users consuming their services, hence the automatic load coordination must happen, and there must be a change for distribution of services in response to the changes in the load [3]. Efficient provisioning of resources as well as well balanced tasks will ensure:

- On demand Resources are easily available.
- Resources are efficiently utilized under condition of load.
- Energy is saved in case of low load (i.e. cloud resources is below certain threshold while the usage).
- Cost of using resources is reduced.

For measuring the efficiency and effectiveness the simulation environment of Load Balancing algorithms are required. CloudSim is the most efficient tool that can be used for cloud modeling. During the cloud lifecycle, CloudSim allows VMs to be managed by hosts which in turn are managed by datacenters.

1.2. Resource Management

Resource management and provisioning is the task of mapping of the resources to different entities of cloud on demand basis [9]. Resources must be allocated in such a manner that no node in the cloud is overloaded and do not undergo any kind of wastage of the available resources in the cloud (wastage of bandwidth or processing core or memory etc.).

Distributed storage systems become more and more popular because of their horizontal scalability and high robustness. cloud computing is not a completely new concept; it has intricate connection to the relatively new grid computing paradigm but thirteen-year of establishment, and other relevant technologies such as utility computing, cluster computing, and distributed systems[5][14]. Cloud computing involves engineers and researchers from various backgrounds, e.g., Grid computing, database and software engineering [8]. Generally, in distributed storage systems there are numbers of storage nodes. Where data can be equally distributed to those nodes by some techniques, such as hashing. By distributing data to storage nodes at different locations, the load of data being accessed is also distributed to those nodes, which improve the ability of processing large amount of data in distributed storage systems. The load balancing mechanism based on data amount balancing is often satisfied in practice. The adaptive data migration model can improve overall performance of the system and resource utilization while meeting workload deadlines [6].

According to the traditional balancing mechanisms, if access load for different data is similar, then the workload will be well balanced for all storage nodes. However, there is a large difference in access the load of different data. Hotspot data are frequently accessed than non-hotspot data. So to balance the workload among those nodes, Split, Merge Algorithms and Replication technique is used. The increasing demand for large amount of data processing, traditional centralized storage systems cannot meet the requirement. Storage nodes storing the large amount of hotspot data could become system bottleneck while the storage nodes with lot of ideal space could cause low computation resource utilization.

A number of open research challenges, addressing which can bring benefits to both resource providers and consumers substantially [11]. And the traditional mechanisms for workload balancing are difficult to achieve the goal. Since the hotspot data are application-related, and it is impossible for the original distributed storage systems to predict or identify that which data will become hotspot data and which will not. To eliminate the system bottleneck, optimize the resource utilization; employ a workload balancing and resource management framework.

II. RELATED WORKS

In this section, a general survey is conducted of the recent works related to load balancing and resource management in cloud. Replication Methods for Load Balancing on Distributed Storages in P2P Networks was used to achieve load balancing and to improve the search performance, replicas of original data are created and distributed over the Internet. In [13], the methods of replication which have been proposed so far focus only on the improvement of search performance. The advantage is to improve the search performance and to achieve load balancing. But the disadvantage is limiting the performance only within the acceptable level.

DLBS: Duplex Loading Balancing Strategy on Object Storage System was the typical load balancing strategies in traditional storage system have been proved to be effective in that they succeeded in providing specific algorithms on data partition and workload distribution. In [12], most of strategies are hard to comply with the new feature device for smarter storage named Object Storage Device (OSD) that is capable of expanding with upper burden to release the storage aware jobs from distributed file systems. So load balance within replica technique emerges as a hot issue in nowadays storage application. The framework proposes a replica-based duplex load balancing strategy (DLBS) to better load balancing. The advantage is can be utilized in the real OBS system to provide more effective and efficient load balancing. And the disadvantage is it does not consider resource utilisation.

Xen and the Art of Virtualization is about numerous systems have been designed which use virtualization to subdivide the ample resources of a modern computer. In [10], presents Xen, an x86 virtual machine monitor which allows multiple commodity operating systems of conventional hardware to share in a safe and resource managed fashion, but without sacrificing either the functionality or the performance. This is achieved by providing an idealized virtual machine abstraction to which operating systems such as Windows XP, Linux and BSD, can be ported with minimal effort. The advantage is virtualization approach taken by Xen is extremely efficient. And the disadvantage is that it does not consider load balancing strategy.

For decentralized workload Balancing Heuristic Neighbor Selection Algorithm in Heterogeneous clustered computational environment is used to execute applications in parallel that require significant amount of computing resources either in the form of computational processing of resources or data storage [7], workload for each system in the cluster cannot be equally or evenly distributed but has to be taken as an important parameter for the workload balancing strategy. A centralized workload balancer would require a global load balancer that will be overwhelmed with message communication in a large clustered environment.

The overhead which reduces the system performance can be addressed by a decentralized approach to load balancing where operations are formulated by the nodes performing scheduling algorithm vis-a-vis communicating with each member node in the cluster. The framework proposes a heuristic based neighbor selection algorithm which selects a neighbor for job distribution whenever overload occurs. Overall improvement for load balancing can be achieved by reducing the average response time was one of the advantages, and the disadvantage is only neighboring node was considered for balancing the load.

Static data placement strategy towards perfect load-balancing for distributed storage is used for applications like cluster-based video-on-demand (VOD) systems which are inherently data-intensive because clients retrieve data frequently stored on a distributed storage subsystem that are interconnected by a local network with high-speed [9], to meet the client request imposed for quality-of-service (QoS), quick responses to access requests are fundamental for these applications.

There are numerous ways to reduce response time but data placement, has attracted much attention from researchers due to its low cost and effectiveness. The framework called perfect balancing (PB), propose a novel load-balancing and performance oriented strategy for static data placement, which can be applied to noticeably improve system responsiveness of distributed storage subsystems in clusters.

The basic idea of PB is to balance the workload across local disks and to minimize simultaneously the discrepancy of service time of data on each disk. An experimental study shows that PB reduces mean response time up to 19.04% and 8.67% using two well-known algorithms such as data placement algorithms Greedy and SP respectively. Quick responses to access requests was an advantage and the disadvantage is only local storage nodes are considered.

III. PROPOSED SYSTEM

Data are distributed and stored into different storage nodes. Each storage node was considered as a virtual node. Proposed system maintains physical machine's computation where the resource can be separated into several parts and allocated to several virtual machines as given in architecture diagram shown in the figure. In traditional distributed storage systems, there is only mapping of data to storage location. The mapping rule is static and difficult to change. To achieve the goal of dynamic workload balancing in the framework, in addition there is a mapping from physical machines to storage nodes, which is dynamic and easy to modify. In this model, storage nodes are deployed in virtual machines and these virtual machines reside in physical machines. To balance the workload among those virtual machines split, merge algorithm and replication technique are implemented.

3.1. Architecture for proposed framework

The architectural or design framework focuses on workload balancing among virtual nodes. As in virtual layer, the framework is responsible for monitoring the work-load of a virtual node and scheduling the computation resource allocated to it was according to its workload status through interacting with the framework deployed in a physical node. To achieve the optimization goal, implement algorithms for virtual layer. First step is resource allocation process where the resources allocated to a virtual node according to its workload. Split, Merge and Replication algorithms are implemented for virtual node migration to achieve the objective of load balancing and efficient resource management.

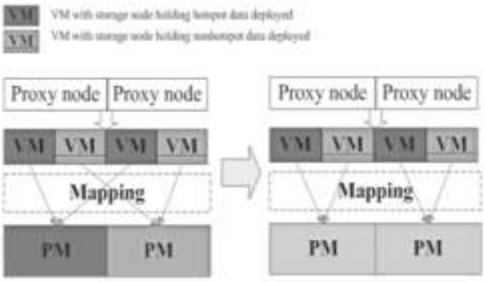
VM
VM with storage node holding hotspot data deployed
VM
VM with storage node holding nonhotspot data deployed
Proxy node
Proxy node
VM
VM
VM
VM
Mapping
PM
PM
Proxy node
Proxy node
VM
VM
VM
VM
Mapping
PM
PM

Fig. 2. Architecture for proposed framework

3.2. Flow chart for the proposed framework

- As the first step cloud users are entered for requesting various numbers of tasks with their respective required memory size.
- The available virtual machine memory segment was monitored and analysed.

- The load balancer checks whether the input memory segment can be allocate to each virtual node.
- The node will reach the overload state when there is no memory space to allocate the receiving new task at that situation the user request was migrated to next virtual machine where split algorithm is used.
- The node will reach the under load state when there is very less amount of task was allocated to a single node. If under loaded means the load balancer will check each node whether the under loaded node can be allocated or not. If true means then those two nodes are merged together and under loaded node was migrated to the other node where merge algorithm was implemented.
- So the above step by step procedure was used to achieve the objective of load balancing and efficient resource management.

3.3. Advantages of proposed system

The advantages of proposed system is to

- Eliminate system bottle neck
- Avoid wastage of underutilized resources
- Increased performance by efficient resource management

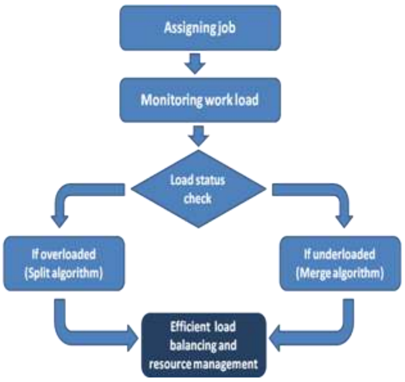
Assigning job
Monitoring work load
Load status check
If overloaded (Split algorithm)
If underloaded (Merge algorithm)
Efficient load balancing and resource management

Fig. 3. Flow chart for effective load balancing and resource management

IV. SYSTEM IMPLEMENTATION

4.1. Monitoring and analysis of workload

To achieve the goal of workload balancing and adaptive resource management, workload monitor is necessary. It provides a foundation for the algorithms proposed. Through monitoring the workload states of virtual nodes are collected and saved for further workload analysis [16]. In the framework, Workload monitor intermittently collects computation resource utilization information collected by workload monitor for virtual node. The goal of workload analysis is to learn the workload status of each node. To solve the problem, the framework defines two types of workload of a node; they are under loaded and overloaded. The state of workload of a node can be represented by a vector known as computation resource utilization vector[15], which consists of CPU utilization, memory utilization, etc. the following functions are used to determine the type of workload of the node.

$L(n)=\text{over loaded if } \square\ i, u_i > opt_i$

L(n)=under loaded if $\sum_{i=1}^{n} \mu_i(\mathrm{opt}_i - u_i) > t$

Where, u_i represents the utilization of Resource i

μ_i represents the weight of Resource i

opt_i represents the defined ideal utilization of Resource i

t represents the threshold of under loaded.

Since excessive utilization of any resource can lead to poor system performance, the rule for determining overloaded is rational [17]. And the rule for determining under loaded takes all kinds of resource utilization into consideration and also provides weight for elastic configuration. Therefore, it is rational, too. Apparently, excessive resource utilization can cause poor system performance. Based on this principle, the framework optimizes the performance through the computation resource utilization regulation. However, low resource utilization does not mean high performance, since critical resources may tends to low resource utilization.

4.2. Splitting and merging the workload

To regulate the virtual node two types of algorithms are used

- Split Algorithm(SA)
- Merge Algorithm(MA)

4.2.1. Split Algorithm

Split Algorithm (SA) is invoked in the virtual node when it is determined to be overloaded which may be caused by the following reason. The node will reach the overload state when there is no memory space to allocate the receiving new task at that situation the user request was migrated to next virtual machine. SA is designed to achieve this goal. SA relieves an overloaded virtual node’s workload through virtual node migration. Before virtual node migration, SA decides to which virtual node the requested task is to be migrated.

Split Algorithm (SA)

```
Step 1: Submit the list of tasks T=T1, T2, T3….Tn by the user.
Step 2: Get the available virtual resources from data centre. i.e. VM=VM1, VM2, VM3………..VMn. Where VM represents the Virtual Machines
Step 3: Take each task one by one and allocate it into the VMs.
Step 4: if CurrentLoad = Overloaded then
Step 5: while each searched VirtualNode do
Step 6: if isSuitable() then
Step 7:  doMigration()
Step 8:  isPaired←true
Step 9:  else
Step 10: not isPaired then
Step 11: bootUpNewVirtualNode()
Step 12: doMigration()
Step 13: end if
Step 14: end while
Step 15: end if
```

4.2.2. Merge Algorithm

Merge Algorithm (MA) is invoked when it is determined to be under loaded. Correspondingly, when a physical node is determined to be under loaded, MA is invoked and tries to move all the virtual nodes in it to other physical nodes. However, Merging process is not executed immediately since there may be some overloaded physical nodes searching for under loaded nodes for SA. Therefore, it waits for a specified period and if there are no requests from overloaded nodes, Merge Algorithm continues and invokes Paring process to determine the virtual node migration program. If migration is successful, then that physical node will be ideal. Hence, it can go to sleep for energy saving.

Merge Algorithm (MA)

```
Step 1:  Submit the list of tasks T=T1, T2, T3….Tn
         by the user.
Step 2:  Get the available virtual resources from
         data centre. i.e. VM=VM1, VM2,
         VM3………..VMn. Where VM represents
         the Virtual Machines
Step 3:  Take each task one by one and allocate it
         into the VMs.
Step 4:  if CurrentLoad = Underloaded then
Step 5:  while each VirtualNode do
Step 6:  isPaired←false
Step 7:  while each searched VirtualNode do
Step 8:  if isSuitable() then
Step 9:  isPaired←true
Step 10: else
Step 11:  if not isPaired then
Step 12:  break
Step 13:  end if
Step 14:  end while
Step 15:  if isPaired then
Step 16:  doMigration()
Step 17:  shutdown TheVirtuallNode()
Step 18:  end if
Step 19:  end while
-------------------------------------------------------------
```

4.2.3. Replication Process

There are an increasing number of enterprises adopting distributed storage systems. So the storage nodes holding intensively hotspot data could become system bottlenecks. The storage nodes without hotspot data might result in low utilization of computing resource [15]. Thus the storage node with hotspot data was extremely overloaded which leads to poor performance of computational resource and the response time is also high. To avoid such situation the original hotspot data was replicated that is the copy of that data was taken and distributed to other available storage nodes so that the workload was distributed equally. This results in achieving the goal of load balancing and efficient resource management in cloud computing.

V. PERFORMANCE EVALUATIONS

The experimental tests were conducted among available virtual machines with varying RAM memory segment. The result was produced based virtual machine allocation using split, merge and replication techniques. The proposed split and merge algorithm was compared with the existing algorithm known as resource allocation algorithm as mentioned in table 2 where the entire users task was allocated to each and every available virtual machines which leads to wastage of resources. To avoid this drawback split and merge algorithm was used which will allocate the requested task within less amount of virtual machines which will be in active state and the remaining idle VMs are in sleep state to achieve the power consumption.

Splitting process will take place using the formula,

$$L(i)=\text{over loaded if } u_i > opt_i$$

TABLE 2. comparison table for Split Algorithm

Parameters	**Algorithm used**	**No of VMs utilised**	**No of migrations**
Before splitting	*Resource Allocation Algorithm*	10	10
After splitting	*Split Algorithm*	6	5

Merge algorithm was compared with split algorithm. After the splitting process the under loaded Virtual Machines are identified and merged together where the usage of number of Virtual Machines are further reduced as given in table 3. Merging technique will take place using the formula,

L(i)=under loaded if $u_i < t$

TABLE 3. comparison table for Merge Algorithm

Parameters	**Algorithm used**	**No of VMs utilised**	**No of migrations**
Before merging	*Split Algorithm*	6	5
After merging	*Merge Algorithm*	4	2

TABLE 4. Comparison table for Resource Allocation, Split and Merge algorithm

Parameters	**Existing Algorithm**	**Proposed Algorithm**	
	Resource Allocation Algorithm	*Split Algorithm*	*Merge Algorithm*
No of VMs utilised	10	6	4
No of migrations	10	5	2

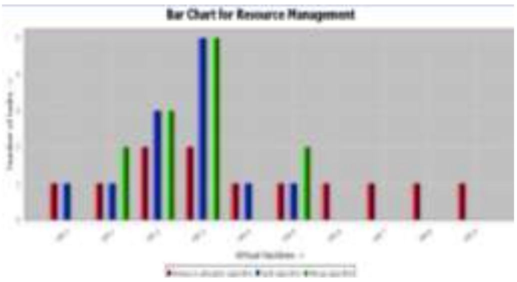

Fig. 4. Graph for Resource management

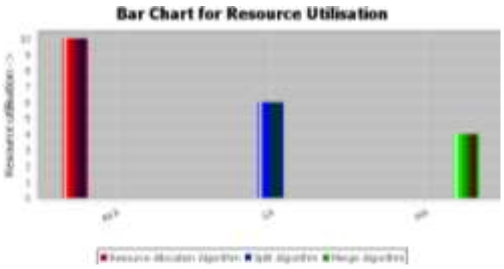

Fig. 5. Comparison graph for Resource Utilization

Through Split algorithm the available resources where managed and allocated efficiently up to 40% when compared with the existing system. And merge algorithm was 33.3% efficient when compared with split algorithm as shown in Fig 5.

The VM which was extremely over loaded contains hot spot data and other VM which does not contain hot spot data have very low utilisation of resources. So in existing system the work load balancing was not achieved. Thus the replication technique was used where the number of extremely overloaded virtual machine was completely reduced as given in the table 5, This technique was used to distribute the work load among the low utilisation resource using the workload distribution formula.

$$\text{Workload distribution}=\frac{\text{Number of users in overloaded VM}}{\text{total number of VMs}}$$

TABLE 5 comparison table for replication process

Parameters	Before Replication	After Replication
No of VMs with hotspot data	1	10
No of extremely overloaded VMs	1	0

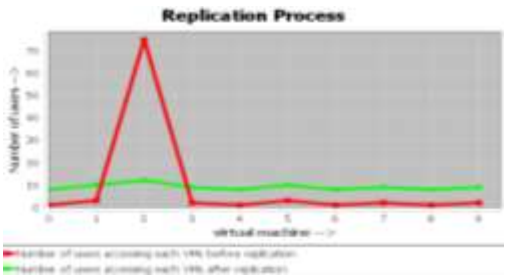

Fig. 6. Graph for Replication process

When compared with the existing system replication technique was 11 percent efficient in balancing the workload among the available virtual machines as shown in Fig 6.

VI. CONCLUSION

Through the experiment result, the proposed framework can achieve the design goal of tuning of system improvement and performance of computation resource utilization and also provides load balancing algorithms for efficient management of available resource and balancing the work load on cloud. The proposed framework can guarantee the reliability of the storage system through live virtual machine migration method as its migration method. During the regulation process of the proposed framework, the service is not interrupted almost. Cloud Computing is a vast concept and load balancing and resource management plays a very important role in case of Clouds. There is a huge scope of improvement in this area. Only few load balancing algorithms are discussed that can be applied to clouds, but there are many other approaches that can be applied to balance the workload in clouds. The performance of the algorithms can also be increased by varying different parameters.

REFERENCES

[1] Ajith Singh. N, HemalathaM., “An approach on distributed load balancing algorithm for cloud computing systems” *International Journal of Computer Applications* Vol-56 No.12, 2012.

[2] Ayyoub, MohammadWardat, “Optimizing expansion strategies ultrascale cloud computing data centers”, http://dx.doi.org/10.1015/ simpat.0021569.190X Elsevier B.V, 2015

[3] Buyya R., R. Ranjan and RN. Calheiros, “InterCloud: Utilityoriented federation of cloud computing environments for scaling of application services,” in proc. 10th International Conference on Algorithms and Architectures for Parallel Processing (ICA3PP), Busan, South Korea,2010.

[4] Deng. L, Lau. A, “Dynamic load balancing for distributed virtual environments”,_in: 17th ACM Symposium on Virtual Reality Software and Technology, ACM Press, New York, pp. 203–210, 2010.

[5] Foster. I, Y. Zhao, I. Raicu and S. Lu, “Cloud Computing and Grid Computing 360-degree compared,” in proc. Grid computing Environments Workshop, pp: 99-106, 2008.

[6] Gong Zhang, L. Chiu, Ling Liu, “Adaptive data migration in multi-tiered storage based cloud environment”, in: Cloud Computing (CLOUD), 2010 IEEE 3rd International Conference on, IEEE Press, New York, pp. 148–155, 2010.

[7] LimJ.W.Y., Poo KuanHoong, Eng-ThiamYeoh, “Heuristic neighbour selection algorithm for decentralized load balancing in clustered heterogeneous computational environment”, in: Advanced Communication Technology (ICACT), 2012 14th International Conference on, IEEE Press, New York, 2012.

[8] LizheWang,Jie Tao, Marcel Kunze “Scientific Cloud Computing: Early Definition and Experience”.Advances International Conference on, IEEE Press, New York, pp. 44–50,2006.

[9] MadathilD.K., ThotaR.B, Paul. P, Tao Xie, A “static data placement strategy towards perfect load-balancing for distributed storage clusters” .in: Software Engineering Advances, International Conference on, IEEE Press, New York, pp. 44–50, 2006.

[10] Paul Barham, Boris ragovic, Keir Fraser, Steven Hand, Tim Harris, XenServer, http://www.citrix.com/products/xenserver/resources-and-support.html IEEE Press, New York, pp. 109–143, 2015.

[11] Shanti Swaroopmoharana, Rajadeepan D. Ramesh Digamber “Powar Aware Analysis of load balancers in cloud computing” *International Journal of Computer Science and Engineering* (IJCSE)ISSN 2278-9960 Vol. 2, Issue 2, May, 101-108,2013.

[12] Tan Zhipeng, Feng Dan, TuXudong, He fei, “DLBS: Duplex loading balancing strategy on object storage system” in: IEEE International Symposium on Parallel and Distributed Processing with Applications, IEEE Press, New York, pp. 45–52, 2013.

[13] Yamamoto. H, D. Maruta, Y. Oie, “Replication methods for load balancing on distributed storages in P2P networks” in: International Symposium on Applications and the Internet, IEEE Press, New York, pp. 264–271, 2005.

[14] Yong Zhao ; Raicu, I. ; ShiyongLu, “Cloud Computing and Grid Computing 360-Degree Compared” Dept. of Comput. Sci., Univ. of Chicago, Chicago, IL, USA ; in proc. Grid computing Environments Workshop, pp: 99-106, 2014.

[15] Zhenhua Wang, Haopeng Chen. Z. Wang et al.“Workload balancing and adaptive resource management for the swift storage system on cloud” in: 51120–131.

[16] W. Zhang X. Qin, , W. Wang, et al., Towards a cost-aware data migration approach for key-value stores, in: 2012 IEEE International Conference on Cluster Computing, IEEE Press, New York, 2012, pp. 551–556.

[17] Z. Liu, M. Lin, A. Wierman, et al., Greening geographical load balancing, in: 2011 ACM SIGMETRICS Joint International Conference Modeling of Computer Systemss, New York, 2011, pp. 233–244.