International Journal of Grid and Distributed Computing
Vol. 14, No. 1, (2021), pp. 1565 – 1573

Efficient Load Forecasting Approach Using Historical Data of Customer Behaviour

Bingi Manorama Devi[1], Vanteru Sudha[2]

[1]Assistant Professor, Department of CSE, K. S. R. M College of Engineering, Kadapa

[2]Assistant Professor, Department of CSE, K. S. R. M College of Engineering, Kadapa

ABSTRACT:

Cloud computing is coming up into view as a new computing standard that is receiving great attention in both academic as well as business community. It provides pay-as-you-use model for accessing different services over the web that can be accessed from anywhere and at any time. Despite of so much of merits it also faces some challenges. One of the main key issues that needed to be taken care of is load balancing. Load balancing is basically about distributing the workload among all the nodes in an even manner such that it will have positive effect on the factors like resource utilization, scalability, fault tolerant etc. Many algorithms and methods have been proposed for this purpose. Due to advancement in technology and growth in human society, it is necessary to work in an environment that reduces cost, utilizes resources effectively, reduces man power and minimizes space utilization. This led to the development of Cloud Computing technology. Cloud computing is a kind of distributed computing with a collection of computing resources located in distributed data centers. It provides massively scalable IT related capabilities to multiple external customers on "pay per use" concept using internet technologies. The increase in the web traffic and different services day by day makes load balancing a critical research topic. Load balancing is one of the central issues in cloud computing. It is the process of distributing the load optimally and evenly among various servers. Proper load balancing in cloud improves the performance factors such as resource utilization, job response time, scalability, throughput, system stability and energy consumption. Many researchers have proposed various load balancing techniques Here, in this paper we are going to investigate some of these load balancing techniques and the latest approaches used for load balancing in order to provide efficient resource utilization, overall cost minimization etc.

Index Terms- cloud computing, load balancing, virtualization, energy aware, load balancing algorithms.

1. INTRODUCTION

Cloud computing has received a great attention in both
academic as well as industrial community as a computing paradigm that provides dynamically scalable and virtualized resources as a service over the web. It has emerged as a next generation platform that has moved computing and data away from desktop and portable PCs to large data centres. Cloud computing provides access to different services on pay-as-you-go basis i.e. users have to pay for on the basis of usage which makes it easy to adjust capacity quickly [1].

Thus, Cloud Computing is a framework that provides on demand network access to a shared pool of computing resources (e.g. networks, servers, storage, and applications). These resources can be provisioned and de-provisioned quickly with minimal management effort or service provider interaction [2].

Currently, there are a lot of multiple vendors offering
cloud services such as Amazon, Microsoft, IBM, Google, HP, Oracle, Citrix, EMC etc. Many researchers have attempted to define the cloud computing. Buyya et al. [3] have defined it as follows: ―Cloud is a parallel and distributed computing system consisting of a collection of inter-connected and virtualised computers that are dynamically provisioned and presented as one or more unified computing resources based on service-level agreements (SLA) established through negotiation between the service provider and consumers. The National Institute of Standards and Technology (NIST) [4] characterizes cloud computing as ―a pay-peruse model for enabling convenient, available, on-demand network access to a shared pool of configurable computing resources (e.g. networks, servers, storage, applications etc.) that can be provisioned and released with minimum management effort or service provider interaction. A. Characteristics According to the national institute of

ISSN: 2005-4262 IJGDC

standards and technology‘s definition of cloud computing, there are following five essential characteristics of cloud computing [4]:

- On demand service- Cloud computing provides consumers with on demand access to different resources over the web.
- Broad Network Access- In cloud computing all the capabilities available over the network are accessed through different mechanisms.
- Resource Pooling- Service providers use different models to pool the resources to make them available to their consumers. All the resources are assigned dynamically and reassigned according to demand.
- Rapid Elasticity- Quantity of resources can be increased or decreased at any time according to the customer‘s requirements
- Measured Service- In cloud computing

environment, usage of different types of resources can be monitored, controlled for both consumers and the provider B. Cloud computing architecture

Service means different types of applications provided by different servers across the cloud. Over cloud many services are delivered to the users which can be divided mainly into three types: Software as a service, Infrastructure as a service and Platform as a service [5].
These services are available to the consumers on subscription basis under the pay-as-you-go model [6]. On the basis of services provided, generally a cloud
computing architecture can be coarsely divided into four layers: the hardware/datacenter layer, the infrastructure as a layer, the platform layer and the application layer as shown in the figure given below [7]:

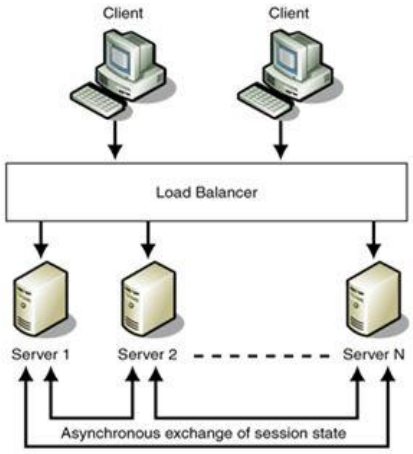

Fig 1.1. System Architecture showing Process of Load Balancing

The hardware layer: At this layer, the physical resources of the cloud including physical servers, routers, switches and cooling systems are managed. It is usually implemented in data centers. Some of the issues handled at this layer include fault tolerance, traffic management, hardware configuration, power and cooling resource management.

- The infrastructure layer: can be referred as the virtualization layer, the infrastructure layer creates a pool of storage and computing resources by partitioning the physical resources using virtualization echnologies such as Xen [8] and VMware [9]etc.
- The platform layer: Consists of operating systems and application frameworks. The platform layer aims at minimizing the burden of deploying application requests directly into virtual machine containers. For example, Google App Engine works at the platform layer for providing the API support to implement storage, database and business logic of web applications.
- The application layer: At the highest level of the hierarchy, lies the application layer that is responsible for providing the actual cloud applications. Cloud applications differs from the traditional ones as they have the advantage of the automatic-scaling feature to achieve availability, better performance, and

ISSN: 2005-4262 IJGDC

lower operating cost.

C. Compelling features:

Cloud computing has many compelling features that attracts both business and technical users. Some of these benefits are following [6, 10]:

- Almost Zero Upfront Infrastructure Investment
- More Efficient Resource Utilization
- More Efficient Development Life Cycle
- Autonomic software updates
- Just-in-Time Infrastructure
- Reduced Time to Market
- Usage-Based paying
- Auto-scaling
- Proactive Scaling
- Disaster Recovery and Business Continuity
- Improved Testability

D. Research Challenges:

As we have already discussed cloud computing has so many benefits but despite of these merits it also faces various challenges that are following:

- Energy management
- Virtual machine migration
- Automated service provisioning
- Traffic management and analysis
- Server consolidation
- Storage technologies and data management
- Data security
- Novel cloud architectures
- Software frameworks
- Load balancing

In remaining paper is organised as follows. In second section, load balancing is described. Third section contains the literature survey of some existing energy aware load balancing techniques. Fourth section describes the discussion and comparison between the algorithms. Fifth one contains the conclusion.

II. RELATED WORK

Virtualization is a rapidly evolving technology that provides a range of benefits to computing systems, such as improved resource management and utilization, application portability and isolation and system reliability etc [14]. Also it provides with features like live migration that helps in moving virtual machines on run. Using this capability we can optimize the resource utilization and thereby energy consumption can be improved. In the field of cloud computing, many works have been done by using this aspect of virtualization to improve the energy efficiency of a cloud. Some existing techniques are discussed below:

• Bo Li et al. [15] have proposed a novel approach called Ena-Cloud which improves the energy consumption using application live placement dynamically. In Ena-Cloud each application is encapsulated in a Virtual Machine, which helps in live migration and application scheduling to minimize the number of running machines and there by helps in saving energy. The placement of application is considered as a bin packaging problem and an energy aware algorithm is designed to tackle the problem.

Two main goals of Ena-Cloud are:

• Minimizing the number of running server nodes
• Minimizing the number of migrations

ISSN: 2005-4262 IJGDC

Above goals are achieved in the paper by aggregating workloads so tightly as to reduce the number of running servers by performing application live migration and filling small workloads in the resource gaps available. Also an Over-provision approach is introduced to deal with dynamic resizing of requirements for resources. One demerit of this technique is that this technique result causes migration overhead.

- Dzmitry Kliazovich et al. [16] have presented a scheduling solution named as e-STAB which takes care of traffic requirements of cloud applications and focuses on the role of communication fabric providing optimized energy efficient traffic load balancing and job allocation in data center networks. Effective distribution of traffic helps in improving quality of service of running cloud applications by reducing the number of congestion hotspots, packet losses and communication related delays. This improvement comes in without sacrificing the energy efficiency.

Main goals of e-STAB are:

- To achieve load balanced network traffic
- Prevent network congestion
- Achieving above goals while optimizing the energy

consumption of datacenter IT equipments.

- Li He [17] has described a multi-objective decision making method of virtual machine placement based on grey correlation degree in order to maintain the energy consumption reduction while improving the resource utilization. In this paper, Li has used three factors like the energy consumption, server level agreement (SLA) violation and server load as the evaluation indexes. Functions for these three factors are built using there evalution indexes and a multi objective decision making model is established for VM placement. The proposed method analyses the influences of the CPU utilization on SLA violation, energy consumption. It also used to analyse the influence of the server load on the number of migrations. It reduces average SLA violation and energy consumption.

It also decreases the total number migrations as it has used server load as an evaluation index and therefor it leads to reduce the loss of CPU utilization in the VM migration.

- T. Kokilavani [18] has explained the Min-Min algorithm for load balancing. It begins with the set of tasks that are not assigned to any server. Firstly, in this algorithm minimum completion time of all the unassigned tasks is calculated. Then from these calculated minimum completion times, a minimum completion time is selected which has the most minimum value among them all. Then the task having the selected minimum completion time is assigned to the corresponding machine required and this task is removed from set of unassigned tasks. After this the execution time for all the tasks running on that machine is calculated and updated. This process will continue until all the unassigned tasks are assigned with the required resource.

This algorithm achieves better performance where the numbers of tasks with smaller execution time are more than the larger ones. This algorithm has drawback that it can lead to starvation. Also it does not consider the task heterogeneity.

- T. Kokilavani [18] has also discussed a technique similar to the Min-Min load balancing algorithm and named that technique as Max-min load balancing algorithm. It also begins with the making a set of unassigned tasks. At first step, calculation of minimum completion time for each unassigned task.

After that from these minimum completion times that minimum completion time is selected which is maximum of them all. Now task corresponding to the maximum minimum time chosen is assigned to the processor required by that task and this task is removed from the set of unassigned tasks. Then execution time of assigned task is added and execution times for all other tasks present on that machine are updated. It has the merit that all the requirements are known already which helps algorithm in performing well.

- Zhong Xu et al. [19] have explained a simplest load balancing technique i.e. Round Robin load balancing algorithm. In this, all the processes are divided into the processors. The round robin scheme is used for allocating the jobs. It allocates the job to first processor randomly and then to other processors in a round robin fashion. Here job is assigned to the processors in a circular order without considering the priority. Although distribution pattern of workloads is same among all processors but execution time is different. As a result, at any

ISSN: 2005-4262 IJGDC

instant of time, some processors tend to heavily loaded while some remain idle. Therefore, this algorithm is mostly used on those web servers where http requests are not only similar in nature as well as equally distributed.

• Dynamic round robin [20] load balancing algorithm is an advanced form of round robin load balancing algorithm. For consolidating the VMs placement it has two rules: the first one states that the machine in retiring state will not be given any load and it will be powered off after all the VMs running on it finish their execution. According to the second rule if any machine that is in use for a very long period of time is in its _retiring' state then it is forced to shut down by transferring all the VMs running on it to another machine. It prevents any machine from reaching to its saturation point

• Qi Zhang et al. [21], in their paper have provided a control theoretic solution to the dynamic capacity provisioning problem. The given solution claims to minimize the total energy cost while meeting some performance objectives. The dynamic capacity provisioning problem is an approach for energy saving in data centers by adjusting the data center capacity dynamically i.e. done by turning off the unused machines. In this paper, this problem is modelled as a constrained discrete-time optimal control problem and also it uses model predictive control to find the optimal control. The proposed dynamic capacity provisioning system controls the number of running servers while keeping in mind factors like demand fluctuations, the cost of dynamic capacity reconfiguration and variance in energy prices. In this paper, through simulations and analysis, it is shown that the given model can provide significant reduction in cost while maintaining an acceptable average scheduling delay for individual tasks. Despite of having several advantages it lacks behind at one point i.e. it assumes all the machines to be homogenous in nature.
• S. Usmin et al. [22] presented an approach named as EVISBP i.e. Enhanced Variable Item Size Bin Packing. It uses virtualization technology to allocate data centers resources dynamically based on application demands and it also supports green computing by optimizing the number of servers actively used. Here, the resource allocation problem has been modelled as the bin packing problem where each server is a bin and each VM is item to be packed. It has used live migration as a variant of the relaxed classical online bin packing problem and a practical algorithm is developed that works well in a real system according to SLAs. In the given method, adjustment of available resources is done to each VM within and across physical servers with memory deduplication technologies. Some of the benefits of this approach are that it can be used where multiple resource constraints are considered. It is a relaxed online method as it does not assume any knowledge of the future while managing the current event and a little movement of already packed items is allowed. In the paper, performance of the given algorithm is not checked corresponding to heavy network load.

• Anton Beloglazov et al. [23] have proposed an energy management system for virtualized Cloud data centers that reduces not only operational costs as well as provides Quality of Service (Qos). The main concept used here is live migration of VMs. In the paper a decentralised architecture of the resource management system is presented and continuous consolidation of VMs is done on the basis of current utilization of resources, thermal state of computing nodes and virtual network topologies established between VMs.

Three stages of continuous optimization of VM placement have been proposed but heuristics for a simplified version of first stage have been presented in the paper. This supports heterogeneity of both VMs and hardware. Also no knowledge about particular applications running on the VMs is required and it is independent of workload.

• Jeffrey M. Galloway et al. [24] have presented an energy aware load balancing approach for IaaS cloud architecture and named it as PALB. The approach considers the heterogeneous nature of local organization's cloud. This proposed algorithm keep track of state of all computing nodes and decide the number of computing nodes need to be in operating state on the basis of their respective utilization percentage. This algorithm could be applied to power aware cluster controller of a local cloud. The presented algorithm consist of three sections. The first section is responsible for finding where to place the new VMs on the basis of utilization percentage of all running nodes, that's why it is called as the balancing section.

The next section i.e. the upscale section is used to power on some extra computing nodes if all the running nodes have reached to utilization over 75%. The third section is downscale section which is responsible for shutting down the idle nodes to save power. It reduces the energy consumption more as compared to some existing load balancing algorithm. It balances the load between the computing nodes. Also it works well in heterogeneous environment. • Kyong Hoon Kim et al. [25], in this paper have investigated power aware virtual machine placement scheme for real time environment. The main goal of the given method is:

ISSN: 2005-4262 IJGDC

- To model a real time service in the form of a real time virtual machine request.
- To manage the placement of the virtual machines using DVFS (Dynamic Voltage Frequency Scaling) schemes.
- Zhen Xiao et al. [26] have introduced an energy aware approach for all allocating data center resources dynamically according to application requirements using virtualization that minimizes the number of servers used and hence supports green computing. A predictive algorithm has been given that can predict the future resource needs without looking inside the vms.

III.PROPOSED SYSTEM

Service capacities are usually regarded to be unlimited in cloud computing, which can be used at any time. In our Proposed System is detailed control of energy-aware operation model used for load balancing with Cloud Work Flow Scheduling using Multiple Servers for application scaling and other Cloud Activities. Extending Scheduling Activities with multiple servers provides Energy Saving, Huge Load Controlling, Sharing Un Used or Idle servers. Idle and lightly-loaded servers are switched to one of the sleep states to save energy. The load balancing and scaling algorithms also exploit some of the most desirable features of server consolidation mechanisms discussed in the literature.

- Cloud servers that are based on different operating administrations with various degrees of processing efficiency.
- Load balancing and application scaling to maximize the number of servers.
- Provides Server Resource Utilization in a Effective way.

V. LOAD BALANCING

In cloud computing environment, requests for different datacenter resources comes at variable time instances. These dynamic workloads tend to make some systems overloaded while some of the systems remain unused. Therefore, it is very necessary to distribute this dynamically coming load effectively so as to prevent such uneven resource utilization. Here comes load balancing concept in use. Load balancing is basically a method of distributing load across different computer clusters, disk drives, CPUs, network links and some other resources, to achieve improved optimal resource utilization, minimize response time, maximize throughput and avoid overloading [11]. Following are some of the aims of Load balancing [12]:

- To maximize the utilization of resources,
- Enhance the performance
- Maintain system stability
- Make a system fault tolerant achieve the user satisfaction minimize the execution time and waiting time of task coming that are from different location while balancing the load ,we need to take care of some factors that can be used to measure the performance of a load balancing algorithm.

These factors are [13]:

- Associated Overhead
- Fault Tolerant
- Migration Time:
- Migration overhead and other associated overheads
- Response Time:
- Utilization of resources
- Scalability

The increasing demand for computational power has led to development of large scale datacenters consuming a large amount of electrical energy. To decrease this consumption of energy there is a need to balance the load in a way that results in less usage of energy. In this paper, some of the energy aware load balancing approaches are discussed.

ISSN: 2005-4262 IJGDC

VI.CONCLUSION AND FUTURE

Load balancing is one of the main issues of cloud computing and balancing the load energy efficiently is more major task to do. In this paper, some energy aware load balancing algorithms are discussed. These techniques are aimed to allocate the resources to the vm requests in a way to reduce the energy consumption. Each of these have some merits and demerits. In future, we will try to design an algorithm that is able to overcome some of these demrits and can improve the resource utilization energy efficiently while considering other performance factors also.

REFERENCES

[1] G. Ritchie, J. Levine, "A Fast, Effective Local Search for Scheduling Independent Jobs in Heterogeneous Computing Environments", Technical report, Centre for Intelligent Systems and their Applications, School of Informatics, University of Edinburgh, Edinburgh, 2003.

[2] Shanti Swaroop Moharana, Rajadeepan D. Ramesh, Digamber Powar, "Analysis of Load Balancers in Cloud Computing ", International Journal of Computer Science and Engineering (IJCSE), Volume 2, Issue 2, ISSN 2278-9960, pp 101-108, May 2013.

[3] Shu-Ching Wang, Kuo-Qin Yan, Wen-Pin Liao, Shun-Sheng Wang, "Towards a Load Balancing in a Three-level Cloud Computing Network", 3rd IEEE Conference on Computer Science and Information Technology[ICCSIT], Taiwan, Volume 1, pp 108-113, 9-11 July 2010, DOI: 10.1109/ICCSIT. 2010.5563889.

[4] Che-Lun Hung, Hsiao-hsi Wang, Yu-Chen Hu, "Efficient Load Balancing Algorithm for Cloud Computing Network", International Conference on Information Science and Technology (IST 2012), pp 251-253, April 28-30, 2012.

[5] Po-Huei Liang, Jiann-Min Yang, "Evaluation of Cloud Hybrid Load Balancer (CHLB)", International Journal of E-Business Development, Volume 3, Issue 1, pp 38-42, Feb 2013.

[6] Meenakshi Sharma, Pankaj Sharma, "Performance Evaluation of Adaptive Virtual Machine Load Balancing Algorithm", International Journal of

[7] Jasmin James, Bhupendra Verma, "Efficient VM Load Balancing Algorithm for a Cloud Computing Environment", International Journal on Computer Science & Engineering, pp 1658-1663, Volume. 4, ISSN: 0975-3397, September 2012.

[8] Mintu M. Ladani, Vinit Kumar Gupta, "A Framework for Performance Analysis of Computing Clouds", International Journal of Innovative Technology and Exploring Engineering (IJITEE), pp 245-247, Volume 2, Issue 6, ISSN: 2278-3075, May 2013.

[9] Mayank Mishra, Anwesha Das, Purushottam Kulkarni, Anirudha Sahoo, "Dynamic Resource Management using Virtual Machine Migration", IEEE Communications Magazine, pp 34-40,Volume 50, Issue 9, September 2012 , DOI: 10.1109/ MCOM. 2012.6295709.

[10] Shu-Ching Wang, Kuo-Qin Yan, Shun-Sheng, Wang, Ching-Wei, Chen, "A Three-Phases Scheduling in a Hierarchical Cloud Computing Network", Third International Conference on Communications and Mobile Computing [CMC], Taiwan, pp 114-117, 18-20 April 2011, DOI: 10.1109/CMC.2011.28.

[11]. Sidhu, Amandeep Kaur, and Supriya Kinger. "Analysis of load balancing techniques in cloud computing." International Journal of Computers & Technology 4, no. 2 (2013): 737-741.

ISSN: 2005-4262 IJGDC

[12]. Kashyap, Dharmesh, and Jaydeep Viradiya. "A Survey Of Various Load Balancing Algorithms In Cloud Computing." International journal of scientific & technology research 3, no. 11 (2014).

[13]. Patel, Mittal, and Chaita Jani. "A Survey on Heterogeneous Load Balancing Techniques in Cloud Computing." International Journal for Innovative Research in Science and Technology 1, no. 10 (2015): 180-185.

[14]. P. Barham, B. Dragovic, K. Fraser, S. Hand, T. Harris, A. Ho, R. Neugebauer, I. Pratt, and A. Warfield. Xen and the Art of Virtualization. In Proceedings of the nineteenth ACM symposium on Operating Systems Principles (SOSP‘03), Lake George, New York, USA, October 19-22, 2003, pp.164-177.

[15]. Li, Bo, Jianxin Li, Jinpeng Huai, Tianyu Wo, Qin Li, and Liang Zhong. "Enacloud: An energy-saving application live placement approach for cloud computing environments." In Cloud Computing, 2009. CLOUD'09. IEEE International Conference on, pp. 17-24. IEEE, 2009.

[16]. Kliazovich, Dzmitry, Sisay T. Arzo, Fabrizio Granelli, Pascal Bouvry, and Samee Ullah Khan. "e-STAB: energy-efficient scheduling for cloud computing applications with traffic load balancing." In Green Computing and Communications (GreenCom), 2013 IEEE and Internet of Things (iThings/CPSCom), IEEE International Conference on and IEEE Cyber, Physical and Social Computing, pp. 7-13. IEEE, 2013.

[17]. He, Li. "A method of virtual machine placement based on gray correlation degree." In Software Engineering and Service Science (ICSESS), 2014 5th IEEE International Conference on, pp. 419- 424. IEEE, 2014.

[18]. Kokilavani, T., and Dr DI George Amalarethinam. "Load balanced min-min algorithm for static meta-task scheduling in grid computing." International Journal of Computer Applications 20, no. 2 (2011): 43-49.

[19]. Xu, Zhong, and Rong Huang. "Performance study of load balancing algorithms in distributed web server systems." CS213 Parallel and Distributed Processing Project Report 1 (2009).

[20]. Lin, Ching-Chi, Pangfeng Liu, and Jan-Jan Wu. "Energy-efficient virtual machine provision algorithms for cloud systems." In Utility and Cloud Computing (UCC), 2011 Fourth IEEE International Conference on, pp. 81-88. IEEE, 2011.

[21]. Zhang, Qi, Mohamed Faten Zhani, Shuo Zhang, Quanyan Zhu, Raouf Boutaba, and Joseph L. Hellerstein. "Dynamic energy-aware capacity provisioning for cloud computing environments." In Proceedings of the 9th international conference on Autonomic computing, pp. 145-154. ACM, 2012.

[22]. Usmin, S., M. Arockia Irudayaraja, and U. Muthaiah. "Dynamic placement of virtualized resources for data centers in cloud." In Information Communication and Embedded Systems (ICICES), 2014 International Conference on, pp. 1-7. IEEE, 2014.

[23]. Beloglazov, Anton, and Rajkumar Buyya. "Energy efficient resource management in virtualized cloud data centers." In Proceedings of the 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, pp. 826-831. IEEE Computer Society, 2010.

[24]. Galloway, Jeffrey M., Karl L. Smith, and Susan S. Vrbsky. "Power aware load balancing for cloud computing." In Proceedings of the World Congress on Engineering and Computer Science, vol. 1, pp. 19-21. 2011.

ISSN: 2005-4262 IJGDC

[25]. Kim, Kyong Hoon, Anton Beloglazov, and Rajkumar Buyya. "Power-aware provisioning of cloud resources for real-time services." In Proceedings of the 7th International Workshop on Middleware for Grids, Clouds and e-Science, p. 1. ACM, 2009.

[26]. Xiao, Zhen, Weijia Song, and Qi Chen. "Dynamic resource allocation using virtual machines for cloud computing environment." Parallel and Distributed Systems, IEEE Transactions on 24, no. 6 (2013): 1107-1117.

[27]. Hieu, Nguyen Trung, Mario Di Francesco, and Antti Yla Jaaski. "A Virtual Machine Placement Algorithm for Balanced Resource Utilization in Cloud Data Centers." In Cloud Computing (CLOUD), 2014 IEEE 7th International Conference on, pp. 474-481. IEEE, 2014.

ISSN: 2005-4262 IJGDC