Review on Load Balancing Techniques in Cloud Computing Environment

Sukhvir Kaur[1], Supriya Kinger[2]

[1]Student, Computer Science and Engineering
Sri Guru Granth Sahib World University, Fatehgarh Sahib, Punjab, 140406, India

[2]Assistant Professor, Computer Science and Engineering
Sri Guru Granth Sahib World University, Fatehgarh Sahib, Punjab, 140406, India

Abstract: ***Cloud computing is a fast growing area in computing research and industry today. Three main services provided by the cloud are IaaS, SaaS, and PaaS. With the advancement of the Cloud, there are new possibilities opening up on how applications can be built and how different services can be offered to the end user through Virtualization, on the internet. There are the cloud services providers who provide large scaled computing infrastructures defined on usage, and provide the infrastructure services in a very flexible manner. The establishment of an effective load balancing algorithm and how to use Cloud computing resources efficiently for effective and efficient cloud computing is one of the Clouds computing service provider's ultimate goals. Cloud computing has critical issue like security, load manager, fault tolerance. In this paper we are discussing the Load Balancing Approach. Many types of load concern with cloud like memory load, CPU load, and network load. Load balancing is the process of distributing the load over the different nodes which provides the good resource utilization when nodes are overloaded with job. Load balancing has to handle the load when the one node is overloaded. When the node is overloaded at that time load is distributed over the other ideal nodes. Many algorithms are available for load balancing like static load balancing and dynamic load balancing This paper presents a review of a few load balancing algorithms or technique in cloud computing and their corresponding advantages, disadvantages and performance metrics are studied in detail.***

Keywords: Cloud Computing, Load Balancing, Distributed Virtual Environments, Power Management.

1. Introduction

Cloud Computing is a new era which aims to have shared data over a one platform. As the technology is booming fast, so does the requirements of the clients. This new paradigm of cloud computing is appealing vendors which increases its popularity. As the definition of NIST [1] says "Cloud Computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g. network, server, storage and applications and services) that can be rapidly provisioned and released with minimal management efforts or service provider interaction.

1.1 Cloud Perspectives

Cloud has different meaning to different stakeholders. There are three main stakeholders of cloud:

1.1.1 End users

These are the customers or consumers of cloud. They use the various services (Infrastructure/ Software/Platform) provided by the cloud. Before using the cloud services, the users of cloud must agree to the Service Level Agreement (SLA) specified by the Cloud Provider. They use the services on demand basis and have to pay for the services availed depending upon their usage. Cloud provides its users flexibility in availing its services by incorporating utility computing. Hence, for end user, the cloud computing is the scenario where the user can have access to any kind of infrastructure, software or platform in a secure manner- at reduced cost on demand basis in an easy to use manner.

1.1.2 Cloud Provider

Cloud provider can offer either public or private or hybrid cloud. They are responsible for building of the cloud. Private clouds are owned by enterprises or business for their internal use. They may use it to store and manage Big-Data of their organization or to provide enough resources on demand basis to its team of employees or clients. They offer greatest level of security. Open Stack, VMware and Cloud Stack are private clouds. Public clouds may be used by individuals or an organization based upon their requirements and necessities. They offer greatest level of efficiency in shared resources. Confidentiality is the major security issue in using public cloud. They are more vulnerable than private clouds. Amazon web services, Google Compute Engine, Microsoft Azure, HP cloud are some of the public clouds. A hybrid cloud is a combination of public and private cloud. It allows businesses to manage some resources internally within organization and some externally. The downside is that the complexity of overall management increases along with security concerns. To optimize the use of one or more combination of private or public clouds allows the businesses to accommodate changing needs of users. Cloud provider must accomplish its job of "resource provisioning". Resource provisioning includes two main tasks. These include managing of huge bundle of resources that make up cloud and providing these resources to the end users. Several provisioning related issues are mentioned in Table 1.

Table 1: Stakeholders of Cloud

Types of Stakeholder	*Requirements/Issues*
End User	-Security -Provenance -Privacy -High Availability -Reduced Cost -Ease-of-use
Cloud Provider	-Managing Resources -Outsourcing -Resource Utilization -Energy Efficiency -Metering -Providing Resources -Cost Efficiency -Meet end user requirements -Utility Computing
Cloud Developer	-Elasticity/ Scalability -Virtualization -Agility and Adaptability -Availability -Data Management -Reliability -Programmability

1.1.3 Cloud developer

This entity lies between end user and cloud provider. Cloud developer has the responsibility of taking into consideration both the perspectives of the cloud (i.e. view of end user and cloud provider). The developer of cloud must adhere to all the technical details of the cloud which are essential to meet the requirements of both, the cloud user as well as the cloud provider. Some of the basic issues that cloud developer must focus on are given in Table 1. Main motive of the developer is to bridge the gap between the end user of the cloud and the cloud provider.

2. Load Balancing

Load balancing is the process of improving the performance of the system by shifting of workload among the processors. Workload of a machine means the total processing time it requires to execute all the tasks assigned to the machine [2]. To provide massive computing and storage resources in cloud environment here needs load balancing. Load balancing is the important concept in network. The load balancer accepts multiple requests from the client and distributing each of them across multiple computers or network devices based on how busy the computer or network device is. Load balancing helps to prevent a server or network device from getting overwhelmed with requests and helps to distribute the work. For example the client can send application request to the server at that time the server over loaded in another process the current process is wait for some time till the serve is idle. Here the client can wait. To avoid this first we check the utilization of the server and process the client request. The CPU utilization can properly do by load balancing algorithm. Load balancing is the mechanism that decides which requesting nodes/client will use the virtual machine and which requesting machines will be put on hold. Load balancing is also required to minimize the power consumption and maximize the user satisfaction forth service being offered. The load balancing algorithm which is dynamic in nature does not consider the previous state or behavior of the system, that is, it depends on the present behavior of the system [3].

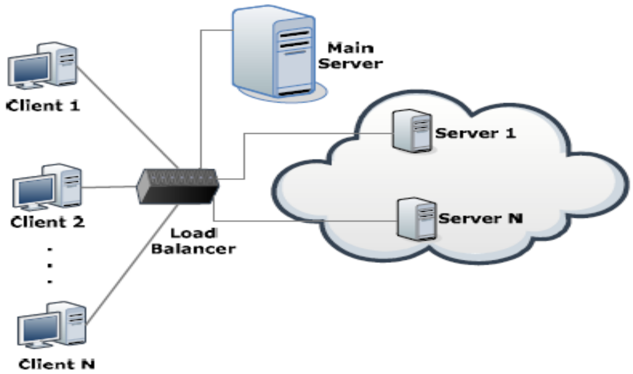


Figure: 1 Load Balancing in Cloud Computing

2.1 Load Balancing on the basis of Cloud Environment

Cloud computing can have either static or dynamic environment based upon how developer configures the cloud demanded by the cloud provider.

2.1.1 Sender Initiated

In this type of load balancing algorithm the client sends request until a receiver is assigned to him to receive his workload i.e. the sender initiates the process.

2.1.2 Receiver Initiated

In this type of load balancing algorithm the receiver sends a request to acknowledge a sender who is ready to share the workload i.e. the receiver initiates the process.

2.1.3 Symmetric

It is a combination of both sender and receiver initiated type of load balancing algorithm. Based on the current state of the system there are two other types of load balancing algorithms.

2.1.4 Static Environment

This approach is generally defined in the design or implementation of the system. This algorithm has a drawback that the task is assigned to the processors or machines only after it is created and that task cannot be shifted during its execution to any other machine for balancing the load.

2.1.5 Dynamic Environment

This approach takes into account the current state of the system during load balancing decisions. This approach is more suitable for widely distributed systems such as cloud computing. An important advantage of this approach is that its decision for balancing the load is based on the current state of the system which helps in improving the overall performance of the system by migrating the load dynamically.

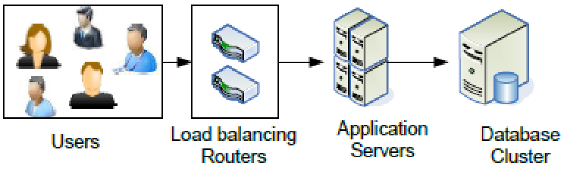


Figure: 2 Schematics of typical high-availability computer System with hardware load balancers

2.1.6 Centralized Load Balancing

In centralized load balancing technique all the allocation and scheduling decision are made by a single node. This node is responsible for storing knowledge base of entire cloud network and can apply static or dynamic approach for load balancing. This technique reduces the time required to analyze different cloud resources but creates a great overhead on the centralized node. Also the network is no longer fault tolerant in this scenario as failure intensity of the overloaded centralized node is high and recovery might not be easy in case of node failure.

2.1.7 Distributed Load Balancing

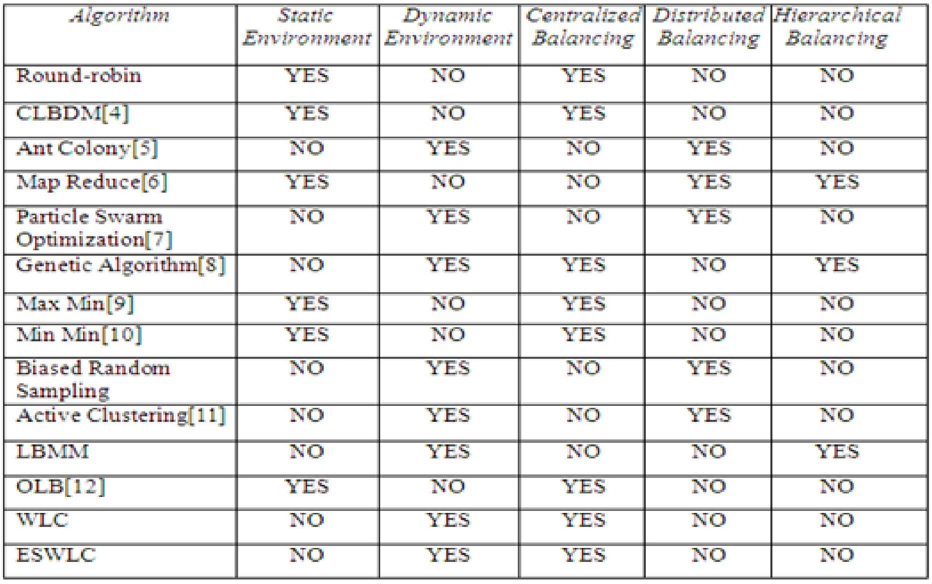
In distributed load balancing technique, no single node is responsible for making resource provisioning or task scheduling decision. There is no single domain responsible for monitoring the cloud network instead multiple domains monitor the network to make accurate load balancing decision. Every node in the network maintains local knowledge base to ensure efficient distribution of tasks in static environment and re-distribution in dynamic environment. In distributed scenario, failure intensity of a node is not neglected. Hence, the system is fault tolerant and balanced as well as no single node is overloaded to make load balancing decision.

2.1.8 Hierarchical Load Balancing

Hierarchical load balancing involves different levels of the cloud in load balancing decision. Such load balancing techniques mostly operate in master slave mode. These can be modeled using tree data structure wherein every node in the tree is balanced under the supervision of its parent node. Master or manager can use light weight agent process to get statistics of slave nodes or child nodes. Based upon the information gathered by the parent node provisioning or scheduling decision is made. Three-phase hierarchical scheduling proposed in paper [19] has multiple phases of scheduling. Request monitor acts as a head of the network and is responsible for monitoring service manager which in turn monitor service nodes. First phase uses BTO (Best Task Order) scheduling, second phase uses EOLB (Enhanced Opportunistic Load Balancing) scheduling and third phase uses EMM (Enhanced Min-Min) scheduling.

2.2 Comparison of Load Balancing Algorithms in Cloud Computing Environment

Table 2: Comparison of Load Balancing Algorithms in Cloud Computing is:



Algorithm	*Static Environment*	*Dynamic Environment*	*Centralized Balancing*	*Distributed Balancing*	*Hierarchical Balancing*
Round-robin	YES	NO	YES	NO	NO
CLBDM[4]	YES	NO	YES	NO	NO
Ant Colony[5]	NO	YES	NO	YES	NO
Map Reduce[6]	YES	NO	NO	YES	YES
Particle Swarm Optimization[7]	NO	YES	NO	YES	NO
Genetic Algorithm[8]	NO	YES	YES	NO	YES
Max Min[9]	YES	NO	YES	NO	NO
Min Min[10]	YES	NO	YES	NO	NO
Biased Random Sampling	NO	YES	NO	YES	NO
Active Clustering[11]	NO	YES	NO	YES	NO
LBMM	NO	YES	NO	NO	YES
OLB[12]	YES	NO	YES	NO	NO
WLC	NO	YES	YES	NO	NO
ESWLC	NO	YES	YES	NO	NO

2.3 Metrics in Existing Load Balancing Techniques in Cloud Computing

Table 3: Metrics in existing Load Balancing techniques in cloud computing

Metric	*Illustration*
Throughput	Throughput It is used to calculate the no. of tasks whose execution has been completed. It should be high to improve the performance of the system
Overhead	It determines the amount of overhead involved while implementing a load balancing algorithm. It is composed of overhead due to movement of tasks, Inter-processor and inter-process communication. This should be minimized so that a load balancing Technique can work efficiently.
Fault Tolerance	It is the time to migrate the jobs or resources from one node to other. It should be minimized in order to enhance the performance of the system.
Response Time	It is the amount of time taken to respond by a particular load balancing algorithm in a distributed system. This parameter should be minimized.
Resource Utilization	It is used to check the utilization of resources. It should be optimized for an efficient load balancing.
Scalability	It is the ability of an algorithm to perform load balancing for a system with any finite number of nodes. This metric should be improved.
Performance	It is used to check the efficiency of the system. This has to be improved at a reasonable cost, e.g., reduce task response time while keeping acceptable delays

3. Existing Load Balancing Techniques in Distributed System

3.1 A Fast adaptive load balancing method

D. Zhang et al. [13] proposed a binary tree structure that is used to partition the simulation region into sub-domains. The characteristics of this fast adaptive balancing method are to be adjusted the workload between the processors from local areas to global areas. According to the difference of workload, the arrangements of the cells are obtained. But the main workload concentrates on certain cells so that the procedure of adjusting the vertices of the grid can be very long because of the local workload can be considered. This problem can be avoided by the fast load balancing adaptive method. Here the region should be partitioned by using the binary tree mode, so that it contains leaf nodes, child nodes, parent nodes etc. There were partition line between the binary tree and the indexes of the cells on the left are smaller that of right and the indexes on the top are smaller than the bottom. Calculate the workload based on the balancing algorithm. This algorithm has a faster balancing speed, less elapsed time and less communication time cost of the simulation procedure. Advantages are Relative smaller communication overhead relative smaller communication overhead, faster balancing speed, and high efficiency and the disadvantage is it cannot maintain the topology that is neighboring cells cannot be maintained.

3.2 Honey Bee Behavior Inspired Load Balancing

Dhinesh et al. [14] proposed an algorithm named honeybee behavior inspired load balancing algorithm. Here in this session well load balance across the virtual machines for maximizing the throughput. The load balancing cloud computing can be achieved by modeling the foraging behavior of honey bees. This algorithm is derived from the behavior of honey bees that uses the method to find and reap food. In bee hives, there is a class of bees called the scout bees and the another type was forager bees .The scout bee which forage for food sources, when they find the food, they come back to the beehive to advertise this news by using a dance called waggle/tremble/vibration dance. The purpose of this dance, gives the idea of the quality and/or quantity of food and also its distance from the beehive. Forager bees then follow the Scout Bees to the location that they found food and then begin to reap it. After that they return to the beehive and do a tremble or vibration dance to other bees in the hive giving an idea of how much food is left. The tasks removed from the overloaded VMs act as Honey Bees. Upon submission to the under load VM, it will update the number of various priority tasks and load of tasks assigned to that VM. This information will be helpful for other tasks , i.e., whenever a high priority has to be submitted to VMs, it should consider the VM that has a minimum number of high priority tasks so that the particular task will be executed earlier. Since all VMs are sorted in an ascending order, the task removed will be submitted to under loaded VMs. Current workload of all available VMs can be calculated based on the information received from the data center. Advantages are maximizing the throughput; waiting time on task is minimum and overhead become minimum. The disadvantage is if more priority based queues are there then the lower priority load can be stay continuously in the queue.

3.3 Heat Diffusion Based Dynamic Load Balancing

Yunhua.et al. [15] proposed an efficient cell selection scheme and two heat diffusion based algorithm called global and local diffusion. Considered the distributed virtual environments there were various numbers of users and the load accessing by the concurrent users can cause problem. This can be avoided by this algorithm. According to the heat diffusion algorithm; the virtual environment is divided in two large numbers of square cells and each square cell having objects. The working of the heat diffusion algorithm is in such a way that every node in the cell sends load to its neighboring nodes in every iteration and the transfer was the difference between the current node to that of neighboring node. So it was related to heat diffusion process. That is the transfer of heat from high to low object, when they were placed adjacently in local diffusion algorithm, there were local decision making and efficient cell selection schemes are used. Here they simply compared the neighboring node loads to the adjacent node loads. If load is small then the transfer of load becomes possible. When global diffusion algorithm considered, it has two stages that is global scheduling stage and local load migration stage. From various experimental results the global diffusion algorithm becomes the better one. Advantages are communication overhead is less, high speed and require little amount of calculations. Disadvantages are network delay is high and several iterations are taken so there was a waste of time.

3.4 Decentralized Scale-Free Network Construction and Load Balancing in Massive Multiuser Virtual Environments

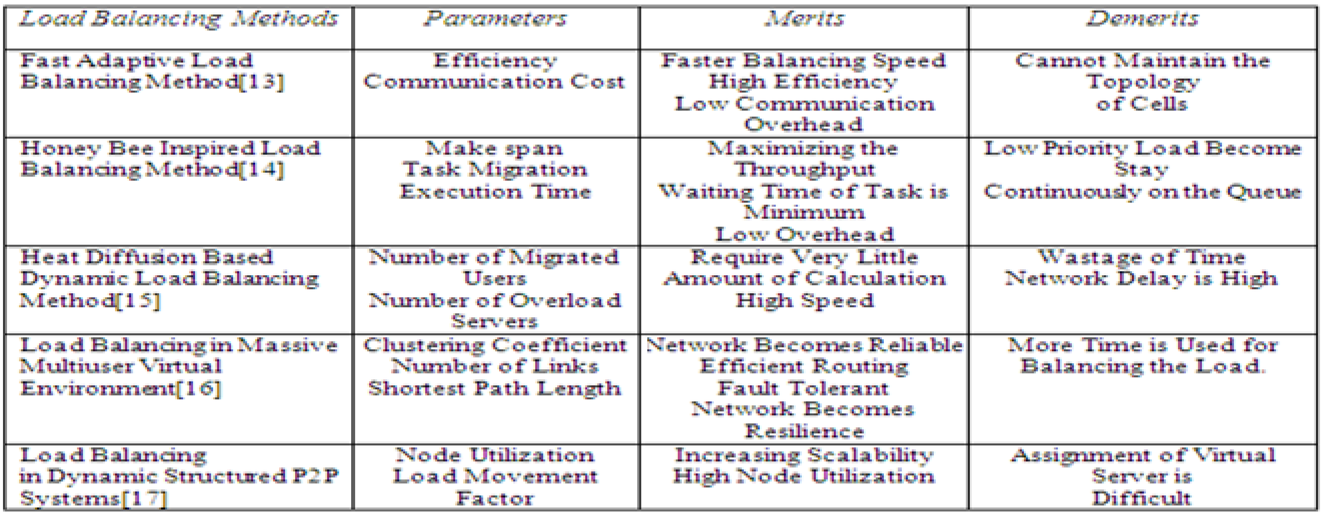
Markus et al. [16] addressed the concept of overlay networks for the interconnection of machines that makes the backbone of an online environment. Virtual online world that makes the opportunities to the world for better technological advancements and developments. So the proposed network that makes better feasibility and load balancing to the dynamic virtual environments. This proposed system developed hyper verse architecture that can be responsible for the proper hosting of the virtual world. There were self organized load balancing methods by which the world surface is subdivided in to small cells, and it is managed by a public server. In this cells various hotspots so that the absolute mass of the object in the cell can be calculated by the public server. Hotspot accuracy is better when increasing the network load. The proposed algorithm cannot avoid the overloaded nodes but find out the number of links that assigned to each node while joining the network. The advantages are the network becomes reliable; the network becomes resilience, efficient routing, and fault tolerant. The disadvantage is the overload ratio at the beginning is higher so that public servers are initially placed randomly so some time is used for balancing the load.

3.5 Load Balancing in Dynamic Structured P2P Systems

Brighten et al. [17] proposed an algorithm for load balancing in dynamic peer-to-peer system and other hybrid environments. In most peer-to-peer system the non uniform of objects in the space and algorithm, the load information of the peer nodes are stored in different directories. These directories help to schedule reassignment of the virtual servers to develop a better balance. Greedy heuristic algorithm used to find out a better solution for the proper utilization of the nodes. The huge number of virtual servers in the system helps to increase the utilization. The various load information in to the corresponding pool and then the virtual server assignments are to be done. This proposed algorithm should be applied to different types of resources like storage, bandwidth etc, It was designed to handle the various situations like varying load of the node, node capacity, entering and leaving of nodes and also insertion and deletion of the nodes. Advantages are high node utilization and increasing scalability. Disadvantage is the reassignment of the virtual server is difficult.

4. Comparative Analysis of Existing Load Balancing Algorithms

Table 4: Comparison Analysis of Existing Load Balancing Techniques is:



Load Balancing Methods	Parameters	Merits	Demerits
Fast Adaptive Load Balancing Method[13]	Efficiency Communication Cost	Faster Balancing Speed High Efficiency Low Communication Overhead	Cannot Maintain the Topology of Cells
Honey Bee Inspired Load Balancing Method[14]	Make span Task Migration Execution Time	Maximizing the Throughput Waiting Time of Task is Minimum Low Overhead	Low Priority Load Become Stay Continuously on the Queue
Heat Diffusion Based Dynamic Load Balancing Method[15]	Number of Migrated Users Number of Overload Servers	Require Very Little Amount of Calculation High Speed	Wastage of Time Network Delay is High
Load Balancing in Massive Multiuser Virtual Environment[16]	Clustering Coefficient Number of Links Shortest Path Length	Network Becomes Reliable Efficient Routing Fault Tolerant Network Becomes Resilience	More Time is Used for Balancing the Load.
Load Balancing in Dynamic Structured P2P Systems[17]	Node Utilization Load Movement Factor	Increasing Scalability High Node Utilization	Assignment of Virtual Server is Difficult

5. Conclusion and Future Work

The purpose of this paper is to focus on one of the major concerns of cloud computing that is Load balancing and Power Consumption. The goal of load balancing is to increase client satisfaction and maximize resource utilization and increase the performance of the cloud system thereby reducing the energy consumed and the carbon emission rate. Also the purpose of load balancing is to make every processor or machine perform same amount of work throughout which helps in increasing the throughput, minimizing the response time and reducing the number of job rejection.

Reference

[1] P. Mell,T.Grance, “The NIST Definition of Cloud Computing (Draft)”

[2] Sandeep Sharma, Sarbjeet Singh, Meenaksshi Sharma, “Performance Analysis of Load Balancing Algorithms, World Academy of Science, Engineering and Technology, 2008

[3] STATEN, J., Is Cloud Computing Ready For The Enterprise? 2008.

[4] Naghibzadeh, M. (2007). A min-min max-min selective algorithm for grid task scheduling. 1-42440-1007-X/07/$25.00, 2007 IEEE. Dept. of Computer Engineering Ferdowsi University of Mashad.

[5] Li, K., Xu, G., Zhao, G., Dong, Y. & Wang, D. (2011). Cloud task scheduling based on load balancing ant colony optimization. 2011 Sixth Annual ChinaGridConference 978-0-7695-4472-4/11 $26.00 © 2011 IEEE. DOI 10.1109/ChinaGrid.2011.17.

[6] Google App Engine, http://appengine.google.com (April 18, 2010).

[7] Vesna Sesum-Cavic Institute of Computer Languages Vienna University of Technology Wien, Austria, Eva Kühn. Applying swarm intelligence algorithms for dynamic load balancing to a Cloud Based Call Center" 2010 Fourth IEEE International Conference on Self-Adaptive and Self-Organizing Systems. 978-0-7695-4232-4/10 $26.00 © 2010 IEEE DOI 10.1109/SASO.2010.19.

[8] Zhao, C., Zhang, S., Liu, Q., Xie, J. & Hu, J. (2009). Independent Tasks Scheduling Based on Genetic Algorithm in Cloud Computing.

[9] Naghibzadeh, M. (2007). A min-min max-min selective algorithm for grid task scheduling. 1-42440- 1007-X/07/$25.00, 2007 IEEE. Dept. of Computer Engineering Ferdowsi University of Mashad.

[10] Naghibzadeh, M. (2007). A min-min max-min selective algorithm for grid task scheduling. 1-42440- 1007-X/07/$25.00, 2007 IEEE. Dept. of Computer Engineering Ferdowsi University of Mashad.

[11] Randles, M., Lamb, D., Bendiab, A. T. (2010). A Comparative Study into Distributed Load Balancing Algorithms for Cloud Computing. 2010 IEEE 24th International Conference on Advanced Information Networking and Applications Workshops 978-0-7695-4019-1/10 $26.00 © 2010 IEEE. DOI 10.1109/ WAINA.2010.85.

[12] Al Nuaimi, K., Mohamed, N., Al Nuaimi, M. & Al-Jaroodi, J. (2012). A survey of load balancing in cloud computing: challenges and algorithms. 2012 IEEE Second Symposium on Network Cloud Computing and Applications 978-0-7695-4943-9/12 $26.00 © 2012 IEEE DOI 10.1109/NCCA.2012.29. College of Information Technology, UAEU Al Ain, United Arab Emirates.

[13] Dongliang Zhang, Changjun Jiang,Shu Li, "A fast adaptive load balancing method for parallel particle-based simulations", Simulation Modelling Practice and Theory 17 (2009) 1032–1042.

[14] Dhinesh Babu L.D, P. VenkataKrishna, "Honey bee behavior inspired load balancing of tasks in cloud computing environments", Applied Soft Computing 13 (2013) 2292–2303.

[15] Yunhua Deng, Rynson W.H. Lau, "Heat diffusion based dynamic load balancing for distributed virtual environments", in: Proceedings of the17th ACM Symposium on Virtual Reality Software and Technology, ACM, 2010, pp. 203–210.

[16] Markus Esch, Eric Tobias, "Decentralized scale-free network construction and load balancing in Massive Multiuser Virtual Environments", in: Collaborative Computing: Networking, Applications and Worksharing, Collaborate Com, 2010, 6th International Conference on, IEEE, 2010, pp. 1–10.

[17] Godfrey, K. Lakshminarayanan, S. Surana, R. Karp, I. Stoica, "Load balancing in dynamic structured P2P systems", in: INFOCOM 2004. Twenty-third Annual Joint Conference of the IEEE Computer and Communications Societies, vol. 4, IEEE, 2004, pp. 2253–2262.

Author Profile

Sukhvir Kaur is Student of M.Tech in the department of Computer Science at SGGSWU, Fatehgarh Sahib, Punjab. She has done B.Tech from BBSBEC, Fatehgarh Sahib Under Punjab Technical University, Jalandhar

Supriya is currently working with SGGS World University, Fatehgarh Sahib, as Assistant Professor. She received her Master's in Technology degree in field of Computer Science from YMCA, India and Bachelor's in Technology degree in Computer Science from Kurukshetra University, India. Her areas of interest include Parallel and Distributed Processing, Cloud Computing and Software Engineering. Currently she is doing research in field of „Energy Efficient Clouds". She has more than 8 years of experience in UG & PG teaching and research. She has attended and organized a number of workshops and conferences and has presented papers in field of Green Clouds, Component Based Software Engineering, Information Security and Networks. She is a life member of ISTE