ISSN:0975-9646

A Review of Load Balancing Technique of Cloud Computing Using Swarm Intelligence

Abhishek Kumar Tiwari
M.Tech Scholar, CSE ,OIST,Bhopal, India

Sreeja Nair
Department of CSE,OIST,Bhopal, India

Abstract- The process of load balancing increases the performance of cloud based services. Cloud based service provide hardware, software and platform as service. For the balancing of public cloud used two types of approach one is traditional approach and other is dynamic approach. The traditional approach follows the concept of CPU scheduling technique such as round robin and first come first service. The dynamic based technique uses swarm intelligence for balancing of load. The swarm intelligence offers various types of algorithm such as particle of swarm optimization, ant colony optimization, glowworm optimization algorithm and many more derived algorithm for optimization. In this paper we present the review of load balancing technique using different swarm based algorithm.

Keywords: - cloud computing, load balancing, swarm intelligence.

INTRODUCTION

Load balancing is major issue in public cloud computing. The public cloud computing infrastructure consists of hardware, software and platform for the execution of public demand and request. For the handling of multiple request of user's cloud computing process uses job scheduling and task scheduling process [4]. The job and task scheduling process is performed by job scheduler, for the selection of resource and job scheduler uses scheduling algorithm such as first come first serve and round robin. But this algorithm is not sufficient for the process of large task on the demand of cloud infrastructure. Currently various authors have used swarm based searching technique for the scheduling of task for the proper execution of task. The family of swarm intelligence gives verity of algorithm such as ant colony optimization, particle swarm optimization and glowworm swarm optimization. It is assumed that the total load considered here is of the arbitrarily divisible of kind that can be partitioned into fractions of loads to be assigned to all the master and slave computers in the cloud. In this case each master computer first assigns a load share to be measured to each of the corresponding N slave computers and then receives the measured data from each slave. Each slave then begins to measure its share of the load once the measurement instructions from the respective master have been completely received by each slave. We also assume that computation time is negligible as compared with communication and measurement time [6]. Cloud computing provides much more effective computing by centralized memory, processing, storage and bandwidth. It should make sure that the tasks are not loaded heavily on one VM and also ensure that some VMs do not remain idle and/or under loaded. In cloud computing technology the data and applications are maintained using the internet and central remote servers. Over the last few years Cloud Computing has been gaining immense popularity where user can pay (as you use) for software, hardware. Load balancing methods aims to speed up the execution of applications by removing tasks from over loaded VMs and assigning them to under loaded VMs and execution of applications of resources whose workload varies at run time in an unpredictable manner [13]. Load balancing is the process of improving the performance of a parallel and distributed system through a redistribution of load among the processors or nodes As Load Balancing is one of the major issues related to cloud computing, the load may represent a CPU capacity, memory, network load etc. It is necessary to distribute the load equally among the nodes in a network. Load balancing can affect the overall performance of a system. Load balancing is dividing the traffic between all servers, so data can be sent and received without any delay with load balancing. In cloud environment many algorithms are available that helps in proper traffic Load between all available servers .Most of them can be applied in the cloud environment with suitable verifications. In cloud computing environment load balancing algorithms can be divided into two main groups. first algorithm type is Batch mode heuristic scheduling algorithms (BMHA) and second is online mode heuristic algorithms. Honey Bee Foraging Algorithm Achieves global load balancing through local server action. Honey Bee Behavior inspired Load Balancing [HBBLB] a technique which helps to achieve even load balancing across virtual machine to maximize throughput. The honey bee behavior load balancing algorithm achieves well balanced load across virtual machines by maximizing the throughput and minimizing the response time [16]. In BMHA Jobs are combined together when they are arriving in the system. The BMHA scheduling algorithm will start after a fixed time period. The examples of BMHA based algorithms are: First Come First Served Scheduling algorithm (FCFS), Round Robin scheduling algorithm (RR), Min Min algorithm and Max Min algorithm. In On-line mode heuristic scheduling algorithm, all Jobs are scheduled when they are arriving in the system. The cloud environment is a heterogeneous system and in this speed of each processor varies quickly and easily [5]. The online mode heuristic scheduling algorithms are more appropriate and better for a cloud environment.

Load balancing methods can be classified in two different ways Static Load Balancing Algorithms and Dynamic Load Balancing Algorithms:- Static Load Balancing Algorithms This approach is mainly defined in the design or implementation of the system. Static load balancing algorithms divide the traffic equivalently between all servers. It distribute the work among processors prior to the execution of the algorithm i.e., the distribution of work load is done at compile time when resource requirements are estimated. Dynamic Load Balancing Algorithms This approach considers only the current state of the system during load balancing decisions. Dynamic approach is more suitable for widely distributed systems such as cloud computing. It searches for the lightest server in the network and then designated appropriate weights on it. Dynamic load balancing approaches have two types .They are distributed approach and non-distributed (centralized) approach. It is defined as following: Centralized approach, In centralized approach, only a single node is responsible for managing and distribution within the whole system. All other nodes are not responsible for this. Distributed approach: In distributed approach, each node independently builds its own load vector. Vector collects the load information of other nodes. All decisions are made locally using local load vectors. Distributed approach is more suitable for widely distributed systems such as cloud computing

II Cloud Load Balancing

In cloud, Load Balancing is a technique to allocate workload over one or more servers, network boundary, hard drives, or other total resources. Representative datacenter implementations depends on massive, significant computing hardware and network communications, which are subject to the common risks linked with any physical device, including hardware failure, power interruptions and resource limits in times of high demand [8]. High-quality of load balance will increase the performance of the entire cloud. Though, there is no general method that can adapt to all possible different circumstances. Various methods have been developed in improving existing solutions to resolve new problems. Each particular method has advantage in a particular area but not in all situations. Therefore, the current model integrates several methods and switches between the load balance methods based on the system status. When the cloud partition is idle, several computing resources are presented and comparatively few jobs are receiving. In these circumstances, this cloud partition has the capability to process jobs as fast as possible so an effortless load balancing method can be used. Zadeh [12] proposed a fuzzy set theory in which the set boundaries were not precisely defined, but in fact boundaries were gradational. Such a set is characterized by continuum of grades of membership (characteristic) function which assigns to each object a grade of membership ranging between zero and one [14]. These days, techniques in artificial neural networks, fuzzy set and fuzzy system have been combined together which is termed as soft computing or intelligence technique.

III Related Work

Siva Theja Maguluri, R. Srikant Et al. [2] In This paper author describe about the cloud computing environment for data storage in a centralized manner and maintain the resources on remote basis. Load balancing has become one of the key issues with the rapid growth of web traffic and the services available under cloud environments. To attain the maximum throughput and minimum time various researchers throughout the world proposed many algorithms and approaches. In this paper they propose a solution for the existing load balancing algorithm “Honey Bee Behavior Inspired Load Balancing Algorithm” to provide optimum balancing of tasks in cloud environments.

Mauro Andreolini, Sara Casolari, Michele Colajanni, Michele Messor Et al. [3] Author here discuss on a new Bee Swarm optimization algorithm called Bees Life Algorithm (BLA) applied to efficiently schedule computation jobs among processing resources onto the cloud datacenters. It is considered as NP-Complete problem and it aims at spreading the workloads among the processing resources in an optimal fashion to reduce the total execution time of jobs and then, to improve the effectiveness of the whole cloud computing services. BLA has been inspired by bees’ life in nature represented in their most important behaviors which are reproduction and food source searching.

Philipp Hoenisch, Stefan Schulte, Schahram Dustdar Et al. [5] Author in this paper define the Although smartphones are increasingly becoming more and more powerful, enabling pervasiveness is severely hindered by the resource limitations of mobile devices. The combination of social interactions and mobile devices in the form of ‘crowd computing’ has the potential to surpass these limitations. In this paper, they introduce Honeybee a crowd computing framework for mobile devices. Honeybee enables mobile devices to share work, utilize local resources and human collaboration in the mobile context. It employs ‘work stealing’ to effectively load balance tasks across nodes that are a priori unknown. They describe the design of Honeybee, and report initial experimental data from applications implemented using Honeybee.

Yossi Azar, Naama Ben-Aroya ,Nikhil R. Devanur Et al. [6] In this paper author shows the interest on cloud computing environment and load balancing mechanism: load balancing is required to achieve evenly distribute load among the nodes and to efficiently make use of the resources Load balancing ensures that all the processor in the system or every node in the network does approximately the equal amount of work at any instant of time. This technique can be sender initiated, receiver initiated or symmetric type (combination of sender initiated and receiver initiated types). This paper presents the features and discusses about the pros and cons of various load balancing algorithm in the authors style. Various consideration of the algorithm like throughput, performance, fault tolerance, migration time, response time etc has been discussed.

Borja Sotomayor, Ruben S. Montero, Ignacio M. Llorente, Ian Foster Et al. [7] In this paper, they define the honey bee foraging mechanism for load balancing is improved by random stealing technique. For finding the state of a virtual machine, the deviation of virtual machine load is calculated and checked for confinement within a threshold condition set. With random stealing method, tasks are stolen from a random Virtual machine when a VM is idle. It thus saves the idle time of the processing elements in the Virtual machine. The performance evaluation is done by using cloudSim. Simulation results show that the modified honey bee foraging technique with random stealing reduces the makespan of algorithm execution together with balancing system load and reduces the idle time of Virtual machine.

Aarti Singha, Dimple Junejab, Manisha Malhotra Et al. [8] In order to provide valuable information and influence the decision-making process of a load balancer, thus maintaining optimal load balancing in hosted (or cloud) environments, it is not enough just to provide information from networking part of the computer system or from external load balancer. Load balancing models and algorithms proposed in the literature or applied in open-source or commercial load balancers rely either on session-switching at the application layer, packet-switching mode at the network layer or processor load balancing mode. The analysis of detected issues for those load balancing algorithms is presented in this paper, as a preparation phase for a new load balancing model (algorithm) proposition.

Sasmita Parida, Suvendu Chandan Nayak Et al. [9] In this paper author discussed about the cloud computing deployment models, In cloud computing, fault tolerance is a major problem and one of the metric which consider being most important since the resource failure affects job execution, throughput, response time and performance of system and network. Fault tolerance in load balancing is one of the main challenges in cloud computing, which is required to distribute the workload equally across all the nodes, detect the fault and remove fault from the network and share workload to all the nodes to increase the performance of cloud network.

Suguna R, Divya Mohandass, Ranjani R Et al. [10] In this paper author define A Kind of two-level centralized scheduling model is proposed with the Global Centralized Scheduler (GCS) at higher level and the Local Centralized Scheduler (LCS) at next level to overcome high communication cost of distributed algorithms and single point of failure problem of centralized algorithms. An energy-efficient load balancing technique can be used to improve the performance of cloud computing by balancing the workload across all the nodes in the cloud with maximum resource utilization, in turn reducing energy consumption and carbon emission to an extent which will help to achieve Green computing.

IV Load Balancing In Cloud Computing Environment

In the event of processing many jobs, the load balancing becomes essential for efficient operation and to improve user satisfaction [9].The purpose of load balancing is to make every processor or machine to perform the same amount of work throughout the network which helps in increasing the throughput, minimizing the response time and reducing the number of job rejection. The scalability of network is generally marked by efficient usage of the resources available [17]. This could only be achieved Load balancing in cloud computing provides an efficient solution to various issues residing in cloud computing environment set-up and usage. Load balancing must take into account two major tasks, one is the resource provisioning or resource allocation and other is task scheduling in distributed environment [3]. Efficient provisioning of resources and scheduling of resources as well as tasks will ensure:

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

For measuring the efficiency and effectiveness of Load Balancing algorithms simulation environment are required [11]. Cloud Sim is the most efficient tool that can be used for modeling of Cloud. During the lifecycle of a Cloud, Cloud Sim allows VMs to be managed by hosts which in turn are managed by datacenters. Cloud sim provides architecture with four basic entities. These entities allow user to set-up a basic cloud computing environment and measure the effectiveness of Load Balancing algorithms [10].

Datacenters entity has the responsibility of providing Infrastructure level Services to the Cloud Users. They act as a home to several Host Entities or several instances hosts' entities aggregate to form a single Datacenter entity. Hosts in Cloud are Physical Servers that have pre-configured processing capabilities. Host is responsible for providing Software level service to the Cloud Users. Hosts have their own storage and memory. Processing capabilities of hosts is expressed in MIPS (million instructions per second). They act as a home to Virtual Machines or several instances of Virtual machine entity aggregate to form a Host entity. Virtual Machine allows development as well as deployment of custom application service models. They are mapped to a host that matches their critical characteristics like storage, processing, memory, software and availability requirements. Thus, similar instances of Virtual Machine are mapped to same instance of a Host based upon its availability. Application and System software are executed on Virtual Machine on-demand [8]. Load Balancing is an important aspect of cloud computing environment. Efficient load balancing scheme ensures efficient resource utilization by provisioning of resources to cloud user's on-demand basis in pay-as-you-say-manner. Load Balancing may even support prioritizing users by applying appropriate scheduling criteria.

Load Balancing is an essential task in Cloud Computing environment to achieve maximum utilization of resources. Various load balancing schemes, each having some pros and cons. On one hand static load balancing scheme provide easiest simulation and monitoring of environment but fail to model heterogeneous nature of cloud. On the other hand, dynamic load balancing algorithm are difficult to simulate but are best suited in heterogeneous environment of cloud computing. Also the level at node which implements this static and dynamic algorithm plays a vital role in deciding the effectiveness of algorithm. Unlike centralized algorithm, distributed nature of algorithm provides better fault tolerance but requires higher degree of replication and on the other hand, hierarchical algorithm divide the load at different levels of hierarchy with upper level nodes requesting for services of lower level nodes in balanced manner. Hence, dynamic load balancing techniques in distributed or hierarchical environment provide better performance. However, performance of the cloud computing environment can be further maximized if dependencies between tasks are modeled using workflows [8].

The load balancing problem is the most divergent job to be achieved while dealing with Cloud Computing. Along with load balancing accomplishment, various access Control mechanisms are to be forked with. DAC, MAC, RBAC are such access controls that have already been developed. But as the number of clients are trafficking the network, so do the problems [2].

Load balancing is a way to spread requests out over multiple resources and it helps a network avoid annoying downtime and delivers optimal performance to users. Network Load Balancing (NLB) can use a distributed algorithm to load balance network traffic across a number of hosts, helping to enhance the scalability. The global server load balancing (GSLB) can operate the Web site or another application server farm at multiple data centers and provide continuous availability by directing users to an alternative site when one site fails or the entire data center is down . General server load balancing only limited to a single data center or near conduct, but GSLB can across different regions [17].

V Conclusion & Future Scope

In this paper present the review of load balancing technique for cloud computing. Load balancing technique is very important issue in cloud computing. The proper management of load balancing improves the efficiency of throughput. Swarm intelligence play an important role in load balancing technique. Cloud Computing is a vast concept and load balancing plays a very important role in case of Clouds. There is a huge scope of improvement in this area. We have discussed only two divisible load scheduling algorithms that can be applied to clouds, but there are still other approaches that can be applied to balance the load in clouds. The performance of the given algorithms can also be increased by varying different parameters.

References:-

[1] Gaochao Xu, Junjie Pang, Xiaodong Fu "A Load Balancing Model Based on Cloud Partitioning for the Public Cloud" Tsinghua Science And Technology 2013 PP 34-39

[2] Siva Theja Maguluri, , R. Srikant "Scheduling Jobs With Unknown Duration in Clouds" IEEE 2014 PP 1938- 1951.

[3] Mauro Andreolini, Sara Casolari, Michele Colajanni, Michele Messor "Dynamic load management of virtual machines In a cloud architectures" PP 1-13.

[4] Rajiv Ranjan, Liang Zhao, Xiaomin Wu, Anna Liu "Peer-to-Peer Cloud Provisioning: Service Discovery and Load-Balancing" CSE UNSW PP 1-27. .

[5] Philipp Hoenisch, Stefan Schulte, Schahram Dustdar "Workflow Scheduling and Resource Allocation for Cloud-based Execution of Elastic Processes" 2013 IEEE.

[6] Yossi Azar, Naama Ben-Aroya ,Nikhil R. Devanur "Cloud Scheduling with Setup Cost" SPAA 2013 PP 23–25.

[7] Borja Sotomayor, Ruben S. Montero, Ignacio M. Llorente, Ian Foster "An Open Source Solution for Virtual Infrastructure Management in Private and Hybrid Clouds" IEEE Internet Computing, Special Issue On Cloud Computing 2009 PP 1-11.

[8] Aarti Singha, Dimple Junejab, Manisha Malhotra "Autonomous Agent Based Load Balancing Algorithm in Cloud Computing", International Conference on Advanced Computing Technologies and Applications , 2015 PP 832-841.

[9] Sasmita Parida, Suvendu Chandan Nayak "Study of Deadline Sensitive Resource Allocation Scheduling Policy in Cloud Computing" IJCSMC, 2014, PP 521 – 528.

[10] Suguna R, Divya Mohandass, Ranjani R "A Novel Approach For Dynamic Cloud Partitioning And Load Balancing In Cloud Computing Environment" 2014 JATIT PP 662-667.

[11] Amittai Aviram, Sen Hu, Bryan Ford Yale ,Ramakrishna Gummadi "Determinating Timing Channels in Compute Clouds" 2010 PP1-6

[12] Siva Theja Maguluri , R. Srikant "Heavy Traffic Optimal Resource Allocation Algorithms for Cloud Computing Clusters" 2012.

[13] M. D. Dikaiakos, D. Katsaros, P. Mehra, G. Pallis, A. Vakali "Cloud computing: Distributed internet computing for IT and scientific research" Internet Computing, 2009. PP.10-13.

[14] P. Mell , T. Grance, "The NIST definition of cloud computing" 2012.

[15] N. G. Shivaratri, P. Krueger, M. Singhal, "Load distributing for locally distributed systems" 1992 PP. 33-44.

[16] S. Penmatsa , A. T. Chronopoulos, "Game-theoretic static load balancing for distributed systems", Journal of Parallel and Distributed Computing, 2011 PP. 537-555.

[17] D. Grosu, A. T. Chronopoulos, M. Y. Leung, "Load balancing in distributed systems: An approach using cooperative games", IEEE Intl. Parallel and Distributed Processing Symp., Florida, USA, 2002, PP 52-61.