International Journal of Engineering Technology and Applied Science

(ISSN: 2395 3853), Vol. 3 Issue 5 May 2017

A Survey on various Load Balancing Techniques in Cloud Computing

Kiran thakur.
Computer Science & Engg.
Lakshmi Narain College of Technology.
Bhopal ,M.p, India.
thakurkiran271@gmail.com.

Dr .Suneel Phulre.
Computer Science Department.
Lakshmi Narain College of Technology.
Bhopal ,M.p, India.
Suneel.lnct@gmail.com.

Dr.Vineet Richhariya.
Head of Department .
Lakshmi Narain College of Technology.
Bhopal, M.p, India.
Vineetrich100@gmail.com.

***Abstract*— Cloud computing (CC) is known as a well-liked and important term in the IT society these days. It has relieve as a huge scale distributed computing paradigm which is driven thru economies of scale and confer the circumstances which services can be dynamically configured and delivered on demand. Load balancing is a procedure of reassigning the aggregate load to the individual nodes of the aggregate framework to make resource usage powerful and to improve the reaction time of the employment, all the while evacuating a condition in that a portion of the nodes are over stacked while some others are under stacked. PSO is a population-based optimization method which search explanation to a difficulty in a search space thru predicting and modeling insect social behavior in the presence of objectives. The general term "particle" is exploited to represent birds, bees or any other individuals who exhibit social behavior as group and interact with each other.**

Keywords*— *Cloud Computing, Load Balancing, Particle Swarm Optimization, Data center, Virtual Machine.

I. INTRODUCTION

Cloud computing is everywhere. Pick up any tech magazine or visit almost any IT website or blog and you'll be sure to see talk about cloud computing. The only problem is that not everyone agrees on what it is. Ask ten different professionals what cloud computing is, and you'll get ten different answers.
And is cloud computing even worth all the hype.

Figure 1. Cloud Computing Architecture

Some people don't think so. In fact, in 2008 Oracle CEO Larry Ellison chastised the whole issue of cloud computing, saying that the term was overused and being applied to everything in the computer world. Berkeley RAD Lab defines Cloud Computing as follows: Cloud Computing refers to both the applications delivered as services over the Internet and the hardware and systems software in the datacenters that provide those services. The services themselves have long been referred to as Software as a Service (SaaS). The datacenter hardware and software is what we will call a Cloud. When a Cloud is made available in a pay-as-you-go manner to the general public, we call it a Public Cloud; the service being sold is Utility Computing. We use the term Private Cloud to refer to internal datacenters of a business or other organization, not made available to the general public. Thus, Cloud Computing is the sum of SaaS and Utility Computing, but does not include Private Clouds. People can be users or providers of SaaS, or users or providers of Utility Computing [1].

II. LOAD BALANCING

Load balancing is used to distributing a larger processing load to smaller processing nodes for enhancing the overall performance of system. In cloud computing environment load balancing is required distribute the dynamic local workload evenly between all the nodes [2].

• Load balancing helps in fair allocation of computing resource to achieve a high User satisfaction and proper Resource utilization .High resource utilization and Proper load balancing helps in minimizing resource consumption. It helps in implementing fail over, scalability, and avoiding bottlenecks.

• Load balancing is a technique that helped networks and resources by providing a Maximum throughput with minimum response time. Load balancing is dividing the traffic between all servers, so data can be sent and received without any delay with load balancing.

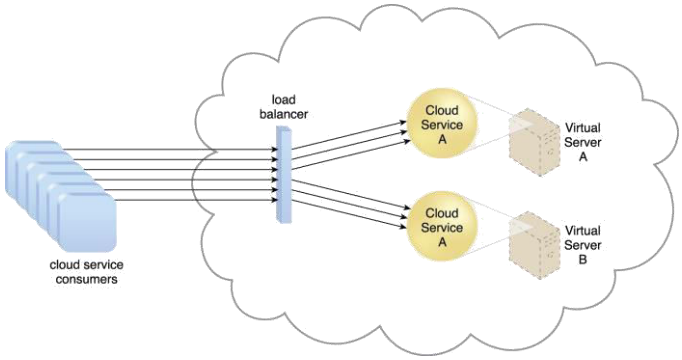

Figure 2. Load Balancing in Cloud Computing

• In cloud environment many algorithms are available that helps in proper traffic Loaded 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. 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. The online mode heuristic scheduling algorithms are more appropriate and better for a cloud environment.

• It is very important to estimate proper load , need to do comparison of all load, stability of all different systems, performance of purposed system, interaction between all the nodes and nature of work to be transferred while developing a load balancing algorithm . The most important thing is selecting the nodes and its also include many other ones. CPU load, amount of memory required combine together to calculate the load of machine.

• In our daily life example of load balancing is websites. Users could experience many Problems without Load balancing like delays, timeouts and long system responses.

A. Load balancing classification:

Fig.3 represents different load balancing algorithms. This is mainly divided into two categories: static load balancing algorithm and dynamic load balancing algorithm:

1) Static approach: - This approach is mainly defined in the design or implementation of the system. Static load balancing algorithms divide the traffic equivalently between all servers.

2) Dynamic approach:- This approach considered only the current state of the system during load balancing decisions. Dynamic approach is more suitable for widely distributed systems such as cloud computing . Dynamic load balancing approaches have two types .They are distributed approach and non-distributed (centralized) approach. It is defined as following:

a) Centralized approach: - In centralized approach, only a single node is responsible for managing and distribution within the whole system. Other all nodes are not responsible for this.

b) Distributed approach: - In distributed approach, each node independently builds its own load vector. Vector collecting 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.

B. Metrics for Load Balancing:

1. Throughput: - It is used to calculate the all tasks whose execution has been completed. The performance of any system is improved if throughput is high.

2. Fault Tolerance: -It means recovery from failure. The load balancing should be a good faulttolerant technique.

3. Migration time: -It is the time to migrate the jobs or resources from one node to other nodes. It should be minimized in order to enhance the performance of the system.

4. Response Time: - It is the amount of time that is taken by a particular load balancing algorithm to response a task in a system. This parameter should be minimized for better performance of a system.

5. Scalability: - It is the ability of an algorithm to perform Load balancing for any finite number of nodes of a system. This metric should be improved for a good system.

C. Policies of load balancing algorithm

There are many policies are used in load balancing algorithms:

• Information policy: It defined that what information is required and how this information is collected. This is also defined that when this information is collected

• Triggering policy: This policy defined that time period when the load balancing operation is starting to manage the load.

• Resource type policy: This policy defined the all types of resources which are available during the load balancing.

• Location policy: This uses all the results of the resource type policy. It is used to find a partner for a server or receiver.

• Selection policy:

This policy is used to find out the task which transfers from overloaded node to free node.

A. Major goals of load balancing algorithms 1. Cost effectiveness:

Load balancing help in provide better system performance at lower cost.

2. Scalability and flexibility:

The system for which load balancing algorithms are implemented may be change in size after some time. So the algorithm must handle these types' situations. So algorithm must be flexible and scalable.

3. Priority:

Prioritization of the resources or jobs needs to be done. So higher priority jobs get better chance to execute.

III. LITERATURE SURVEY

Josué Pagán (2017) et al present that The huge amount of data generated for these systems require high computing capabilities only available in Data Centers. This paper presents a real case of prediction in the eHealth scenario, devoted to neurological disorders. The presented case study focuses on the migraine headache, a disease that affects around 15% of the European population. This paper extrapolates results from real data and simulations in a study where migraine patients are monitored using an unobtrusive Wireless Body Sensor Network. Low-power techniques are applied in monitorization nodes. Techniques such us: on-node signal processing and radio policies to make node's autonomy longer and save energy, have been applied. Workload balancing policies are carried out in the coordinator nodes and Data Centers to reduce the computational burden in these facilities and minimize its energy consumption. Our results draw average savings of e 288 million in this eHealth scenario applied only to 2% of European migraine sufferers; in addition to savings of e 1272 million due to the benefits of the migraine prediction [4].

Alireza Sadeghi Milani (2016) et al present that The optimal selection of a resource for a particular job does not mean that the selected resource persists optimized for the whole execution of the job. The resource overloading/under-loading must be avoided which could be gained by appropriate load balancing mechanisms. However, to the best of our knowledge, despite the importance of load balancing techniques and mechanisms, there is not any comprehensive and systematic review about studying and analyzing its important techniques. Hence, this paper presents a systematic literature review of the existing load balancing techniques proposed so far. Detailed classifications have also been included based on different parameters which are relying upon the analysis of the existing techniques. Also, the advantages and disadvantages associated with several load balancing algorithms have been discussed and the important challenges of these algorithms are addressed so that more efficient load balancing techniques can be developed in future [5].

Ranesh Kumar Naha (2016)et al present that A cloud-based service broker provides intermediation to seek appropriate service providers in terms a suitable trade-off between price and performance. On the other hand, load balancing among cloud resources ensures efficient use of a physical infrastructure, and at the same time, minimizes execution time. This makes service brokers and load balancing among the most important issues in cloud computing systems. This paper aims to propose three different cloud brokering algorithms, and a load balancing algorithm. A simulation-based deployment confirms that our proposed algorithms minimized the cost, and at the same time, witnessed gains in service performance [6].

M. Vanitha (2016) et al present that The cloud service provider undertakes to deliver all the subscribers' requirements as per the service level agreement (SLA). These resources must be well-protected since they are used by many subscribers. There is a constant high level of demand for these resources and services, and it is therefore necessary to balance the loads on the various servers; this is done in order to avoid congestion in the network and to reduce consumption of the capital or resources. This load balancing uses algorithms such as genetic algorithms (GA) in order to distribute the load on the virtual machines (VM) that are available in the network. A novel load balancing method is proposed, involving a well-organized use of resources, which is known as the dynamic well-organized load balancing (DWOLB) algorithm. This is a powerful algorithm for reducing the energy that is consumed in cloud computing [7].

Dinesh Kumar (2015) et al present that Virtual machine scheduling problem can be defined as an allocation of a set of virtual machines (VMs) to a set of physical machines (PMs). The proposed work focuses on PSO based VM schedulin strategy for VM placement in cloud infrastructure. The strategy focuses on efficient VM allocation to physical servers in order to minimize the total resource wastage and the number of servers used. Simulation experiments were conducted to observe the allocation of VMs to the servers and to evaluate the proposed algorithm with respect to performance and scalability [8].

Vahid Abedifar (2013) et al present that Virtualization of optical networks is key enabler of cloud computing. A main part of optical network virtualization is virtual network mapping on the physical infrastructure. One major issue in Routing and Wavelength Assignment, RWA, problem is optimized allocation of optical network resources. In this paper at first a background on Particle Swarm Optimization concept and formulation is presented. Then, an optimization scheme using PSO is proposed for virtual network mapping. Five different cost functions are formulated and a new encoding method for optical networks is proposed. The constraints for solutions of RWA problem are addressed and some heuristics are proposed to satisfy them. Proposed optimization scheme is simulated by finding the map of different virtual networks on a physical infrastructure in order to optimize five different cost functions. Results are presented and discussed for defined cost parameters [9].

IV CONCLUSION

It can be concluded from the various techniques that proposed round robin selection of data centers in service broker policy works efficiently when it comes to resource utilization. It can also be observed that the total cost is same for all data centers when proposed policy using round robin distribution is used in experimentation when compared to conventional data center selection algorithm. From the results of simulation it can be concluded that proposed algorithm works efficiently when it comes to resource utilization, processing time of the data center and response time of user base. In future, we can implement more effectual policy to select data center with maximum number of resources required for processing of user requests and to process requests based on priorities

REFERENCES

[1]. Shyam Patidar, Dheeraj Rane and Pritesh Jain," A Survey Paper on Cloud Computing", 2012 Second International Conference on Advanced Computing & Communication Technologies, pp: 397-401.

[2]. Rajwinder Kaur and Pawan Luthra," Load Balancing in Cloud Computing", Proc. of Int. Conf. on Recent Trends in Information,
Telecommunication and Computing, ITC, Association of Computer Electronics and Electrical Engineers, 2014

[3]. Tao Xiang, Weimin Zhang and Fei Chen," A Verifiable PSO Algorithm in Cloud Computing", 2014 IEEE Congress on
Evolutionary Computation (CEC) July 6-11, 2014, Beijing, China

[4]. Josué Pagán, Marina Zapater and José L. Ayala," Power transmission and workload balancing policies in eHealth mobile cloud computing scenarios", Future Generation Computer Systems
(2017)

[5]. Alireza Sadeghi Milani and Nima Jafari Navimipour," Load balancing mechanisms and techniques in the cloud environments: systematic literature review and future trends", Journal of Network and Computer Applications 2016

[6]. Ranesh Kumar Naha and Mohamed Othman," Cost-aware service brokering and performance sentient load balancing algorithms in the cloud", Journal of Network and Computer Applications 75 (2016) 47–57

[7]. M. Vanithaa and P. Marikkannu," Effective resource utilization in cloud environment through a dynamic well-organized load balancing algorithm for virtual machines", Computers and
Electrical Engineering 000 (2016) 1–10

[8]. Dinesh Kumar and Zahid Raza," A PSO based VM Resource Scheduling Model for Cloud Computing", 2015 IEEE International
Conference on Computational Intelligence & Communication Technology, 2015 IEEE, pp:213-219.

[9]. Vahid Abedifar, Mohammad Eshghi, Seyedali Mirjalili and S. Mohammad Mirjalili," An Optimized Virtual Network Mapping Using PSO in Cloud Computing", 2013 IEEE

.

International Journal of Engineering Technology and Applied Science

(ISSN: 2395 3853), Vol. 3 Issue 5 May 2017

Paper ID: IJETAS/May/2017/81