Various Load Balancing Approaches in Cloud Computing

J.Ruby Elizabeth
Assistant Professor, Department of Computer Science and Engineering, VV College of Engineering, Tisaiyanvilai, Tamil Nadu, India.

Ru.Va.Gayathri
Assistant Professor, Department of Computer Science and Engineering, VV College of Engineering, Tisaiyanvilai, Tamil Nadu, India.

Abstract – Cloud computing is the most recent emerging paradigm promising to turn the vision of "computing utilities" into a reality. Cloud computing provides the capability to use computing and storage resources on a metered basis and reduce the investments in an organization's computing infrastructure. The spawning and deletion of virtual machines running on physical hardware and being controlled by hypervisors is a cost-efficient and flexible computing paradigm. Load balancing distributes workloads across multiple computing resources, such as computers, a computer cluster, network links, central processing units or disk drives. Load balancing aims to optimize resource use, maximize throughput, minimize response time, and avoid overload of any single resource. In this paper the various load balancing approaches in cloud computing are compared.

Index Terms – Cloud computing, Load balancing.

1. INTRODUCTION

A cloud is a pool of virtualized computer resources. A cloud can host a variety of different workloads, including batch-style backend jobs and interactive and user-facing applications. Computing is being transformed by a new model, cloud computing. In this model, data and computation are operated somewhere in a "cloud," which is some collection of data centers owned and maintained by a third party[1]. Cloud computing is a technological advancement that focuses on the way we design computing systems, develop applications, and leverage existing services for building software. It is based on the concept of dynamic provisioning, which is applied not only to services but also to compute capability, storage, networking, and information technology (IT) infrastructure in general. Resources are made available through the Internet and offered on a pay-per-use basis from cloud computing vendors. Load balancing is a technique to enhance resources, utilizing parallelism, exploiting throughput improvisation, and to cut response time through an appropriate distribution of the application.[2]

2. CLOUD MODELS

2.1. Infrastructure as a Service (IaaS)

IaaS offerings are computing resources such as processing or storage which can be obtained as a service. Examples are Amazon Web Services with the Elastic Compute Cloud (EC2) for processing and Simple Storage Service (S3) for storage and Joyent who provide a highly scalable on demand infrastructure for running web sites and rich web applications. PaaS and SaaS providers can draw upon IaaS offerings based on standardized interfaces. Instead of selling raw hardware infrastructure, iaas providers typically offer virtualized Infrastructure As A Service. Virtual resources provisioned by the users are billed on a pay per use paradigm. Common metering metrics used are the number of virtual machine hours used and/or the amount of storage space provisioned [3].

2.2. Platform as a Service(PaaS):

Platforms are an abstraction layer between the software applications (SaaS) and the virtualized infrastructure (IaaS). PaaS offreings are targeted at software developers. Developers can write their applications according to the specifications of a particular platform without needing to worry about the underlying hardware infrastructure (IaaS). Developers upload their application code to a platform, which then typically manages the automatic upscaling when the usage of the application grows. Examples are Google App Engine, which allows application to be run on Goole's infrastructure, and salesforce's Force.com platform. The PaaS layer of a cloud relies on the standardized interface of the IaaS layer.[3]

2.3. Software as a Service(SaaS):

SaaS is asoftware that is owned, delivered and managed remotely by one or more providers and that is offered in a pay per use manner. SaaS is the most visble layer of cloud computing for end users, because it is about the actual software applications that are accessed and used. From the prespective of the user, obtaining software as a service is mainly motivated by cost advantages due to utility based payment model. The examples are salesforce.com and Google Apps.The typical user of a SaaS offering usually has neither knowledge nor control about the underlying infrastructure.SaaS applications are platform independent and can be accessed from various client devices such as workstations, laptop, tablets and smartphones, running different operating systems [3].

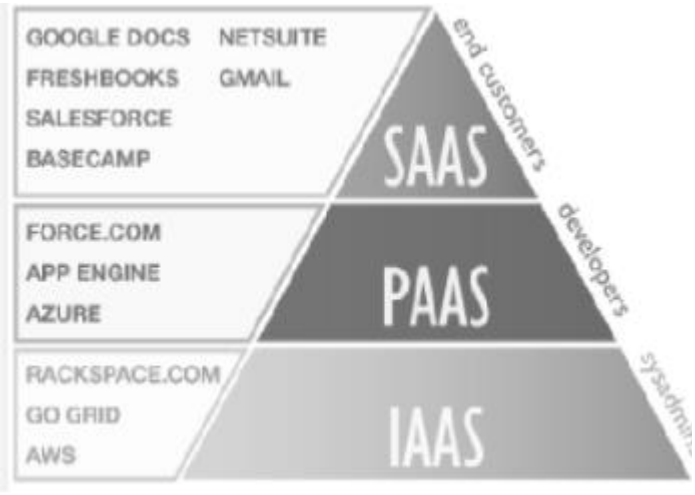

Figure 1: Cloud Models

3. CLOUD DEPLOYMENT MODELS

3.1. Public Cloud

A public cloud is data center hardware and software run by third parties.Example: Google and Amazon, which expose their services to companies and consumers via the internet. A public cloud is not restricted to a limited user, it is made available in a pay as you go manner to the general public.The cloud resorces are shared among different users. Public clouds are best suited for usres who want to use cloud infrastructure for development and testing of applications and host applications in the cloud to serve large workloads, without upfront invesments in IT infrastructure.

3.2. Private Cloud

Private clouds refer to such internal data centers of a company or other organization. A private cloud is fully owned by a single company who has total control over the applications run on the infrastructure, the place where they run and the people or organizations using it. Public clouds are best suited for applications where security is very important and organizations that want to have tight control over their data. Examples: Sun, Hp Data Center, Oracle, IBM, 3tera.

3.3. Hybrid Cloud

A hybrid cloud is an integrated cloud service utilizing both private and public clouds to perform distinct functions within the same organization. The public and private cloud infrastructures, which operate independently of each other, communicate over an encrypted connection, using technology that allows for the portability of data and applications.

4. LOAD BALANCING APPROACHES

Load balancing divides the amount of work that a computer has to do between two or more computers so that more work gets done in the same amount of time and in general, all users get served faster. The metrics [4] for Load Balancing are

1. Throughput: - Throughput is the rate of production or the rate at which something can be processed. 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 fault tolerant technique.

3. Migration time: It is the time to migrate the jobs or resources from one node to other nodes. The migration time should be minimum.

4. Response Time: Response time is the time taken by a load balancing algorithm to response a task in a system. The response time should be minimum.

5. Scalability: It is the ability of an algorithm to perform Load balancing for any finite number of nodes of a system.

6.Performance: It is used to check the efficiency of the system.

The Existing Load Balancing Algorithm are as follows:

1. Task Scheduling Based On LB:

This algorithm consists of two levels for task scheduling mechanism. This algorithm achieves load balancing by first mapping tasks to virtual machines and then all virtual machines to host resources. [5],[9]

2. Opportunistic Load Balancing Algorithm:

Opportunistic Load Balancing (OLB) is the algorithm that assigns workloads to nodes in free order. It do not consider the present workload of the VM. It keeps each and every node busy.[5]

3. Active clustering Algorithm:

Active Clustering is an improved method of random sampling. In active clustering, same type of system nodes are grouped together to form a cluster. It is used in large scale cloud system. A method called match-maker is introduced in this algorithm .While an execution starts, the first node selects the neighbour node. The neighbour node is taken as match make node, which connects the neighbour node that is same as initial node. At last the match maker node gets disconnected. And this process is done iteratively to balance the load equally.

4. Ant colony optimization Algorithm:

Ant algorithm is suitable for complex combinatorial optimization problem. It uses multi-agent approach. In this optimization algorithm, the movement of the ant is triggered when the request is made.[5] There are two ways in the movement of ants. They are

Forward Movement- The ant moves continuously in the forward direction from one overloaded node to another node

and the algorithm performs a check to determine whether the node is over-loaded or underloaded. If it finds an over-loaded node, the process is repeated.

Backward Movement: The ant moves in the backward direction to the previous node when it finds an over-loaded node. If it finds the target node, the algorithm terminates.[2],[10]

5. Honeybee Foraging Behavior:

This algorithm is based on the behaviour of honey bees. Honey bees have been classified into two types. They are: finders and reapers. The finder honeybee helps in finding the honey source. Once honey source is found, they do the waggle dance to indicate the quality and quantity of available honey. After that, the reapers gather the honey from the sources. Then, again they go for the waggle dance to specify the honey that is left. In load balancing, the servers are combined together as virtual servers, where each and every virtual server has a process queue. Once the request is received from the queue, it calculates the profit quality as the bee does in waggle dance [4]. The server stays only when the profit is high, or else proceeds to forage by indicating that whether the state is loaded, overloaded, under loaded or balanced. Based on this, the current virtual machines are combined. It needs to maintain a separate queue for each and every node. [5],[8]

6. A Fast Adaptive Load Balancing Method:

The algorithm partition the simulation region into sub- domains using binary tree structure. The indexes of the cells follows the binary tree structure ie, the cells with less index is placed on the left and remaining on the right and the root of the tree always holds the cell with the smaller index. When there is a change in workload between local areas and global areas, the characteristics need to be adjusted. Then the workload is calculated based on the balancing algorithm. This algorithm has a faster balancing speed, high efficiency, less elapsed time and less communication time cost of the simulation procedure.[2]

7. Biased Random Sampling:

This algorithm uses random sampling to self-organize the balancing nodes. The virtual graph is constructed based on the connectivity between all nodes. Each node of the graph corresponds to the node computer of the cloud system. Edges are used for considering the load of particular system and also for the of the resources of the node.[5]

8. Max-Min Algorithm:

This algorithm chooses the maximum value based on their minimum execution times. Then the task is assigned to the selected machine. The execution time of the assigned tasks is added with the execution times of other tasks on the selected machine, them the execution times of all tasks are updated. After execution all the assigned tasks are removed from the list. [5]

9. Lock-free multiprocessing solution for LB:

This algorithm proposes a lock-free multiprocessing load balancing solution which avoids the use of shared memory but other multiprocessing load balancing solutions uses shared memory and lock to maintain a user session. It is achieved by modifying the kernel.[5]

10. Heat Diffusion Based Dynamic Load Balancing:

In this algorithm, proposed an efficient cell selection scheme and two diffusion based algorithm called global and local diffusion. According to heat diffusion algorithm, the virtual environment is divided into a large no of square cells and each square cells having object and every node in the cell send load to its neighboring nodes in every iteration and the transfer was the difference between the current node to that of neighboring node. it is related to heat diffusion process.[2],[7].

11. Compare and Balance Algorithm:

This algorithm manages unbalanced systems load using an equilibrium condition. The current host randomly selects a host based on the probability and their load is compared. If the current host's load is more then extra load is transferred to the selected host. This procedure is repeated until the load gets balanced.[2]

12. Load Balancing Using Firefly Algorithm:

This algorithm is inspired from firefly algorithm. The proposed approach deals with a simulated cloud network with set of requests and servers. The servers are associated with nodes and each node is supplied with some attributes. The attributes are assigned to control the load in each node.

It has three levels.

1.Initially a population is generated from the cloud network.

2.Scheduling index calculation

3.The scheduling list is optimized by firefly algorithm. It is efficient in optimizing the scheduling by balancing the load.[5]

13. Load Balancing Particle Swarm Optimization (PSO) Algorithm:

Particle swarm optimization (PSO) is a group based searching algorithm. PSO is simple and effective with low computational cost and fast speed that is the reason PSO becomes popular but it has low convergence accuracy.

S.NO	LOAD BALANCING ALGORITHMS	ADVANTAGES	DISADVANTAGES
1.	Task Scheduling Based On LB	Provides high resource utilization and improves task response time.	Doesn't improve response to request ratio
2.	Opportunistic Load Balancing Algorithm	Keeps each and every node busy.	The tasks are processed in a slow manner. Whole completion time is poor.
3.	Active clustering Algorithm	Efficient utilization of resources.	Performance is poor when there is an increase in variety of nodes.
4.	Ant colony optimization Algorithm	Reduces the unnecessary backward movement. Suitable for fault tolerance	Complex Network
5.	Honeybee Foraging Behavior	Best suited for the conditions where the diverse population of service types is required.	Throughput deceases when variety of nodes increases.
6.	A Fast Adaptive Load Balancing Method	faster balancing speed, less elapsed time & less communication time	The topology of the cells cannot be maintained
7.	Biased Random Sampling Max-Min Algorithm	Achieves load balancing across all system nodes using random sampling of the system domain	Smaller jobs have to be waiting for long time
8.	Lock-free multiprocessing solution for LB	Improves the overall performance of load balancer in a multicore environment.	Modifying of kernel is needed
9.	Heat Diffusion Based Dynamic Load Balancing	Communication overhead is less, high speed and require little amount of calculation.	High network delay and waste of time because of several iterations.
10.	Compare and Balance Algorithm	Reduces virtual machines migration time.	Poor performance and throughput
11.	Load Balancing Using Firefly Algorithm	Diminishes unnecessary load movement	trapping into several local optima when solving complex problem
12.	Load Balancing Particle Swarm Optimization (PSO) Algorithm	It increases the overall performance Low convergence accuracy	affinity problem in inertia weight

5. CONCLUSION

Cloud computing is one of the rapidly growing area in the field of computer science. Load balancing is very important since many users prefer cloud computing due to its increasing advantages. Load balancing helps to improve the metrics so that the performance of system is impproved. In this paper the basic concepts of cloud computing and various existing load balancing algorithms that provide better scheduling and resource allocation techniques are discussed. Many researchers have been done in this area inorder to improve the system performance.

REFERENCES

[1] Namrata Swarnkar , Asst. Prof. Atesh Kumar Singh and Dr. R. Shankar A Survey of Load Balancing Techniques in Cloud Computing. Vol. 2 Issue 8, August – 2013

[2] Mell, Peter and Grance, Tim, "The NIST definition of cloud computing", National Institute of Standards and Technology, 2009,vol53, pages50, Mell2009

[3] Kai Hwang, Geoffrey C Fox, Jack G Dongarra, "Distributed and Cloud Computing, From Parallel Processing to the Internet of Things", Morgan Kaufmann Publishers, 2012.

[4] Amandeep, Vandana Yadav, "Different Strategies for Load Balancing in Cloud Computing Environment: a critical Study", International Journal of Scientific Research Engineering & Technology (IJSRET), ISSN 2278 – 0882 Volume 3 Issue 1, April 2014

[5] Rajwinder Kaur and Pawan Luthra, "Load Balancing in Cloud Computing", ACEEE,Proc. of Int. Conf. on Recent Trends in Information, Telecommunication and Computing, ITC.

[6] A. Paulin Florence and V. Shanthi A LOAD BALANCING MODEL USING FIREFLY ALGORITHM IN CLOUD COMPUTING. © 2014 Science Publications

[7] 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

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

[9] Y. Fang, F. Wang, and J. Ge, A Task Scheduling Algorithm Based on Load Balancing in Cloud Computing, Web Information Systems and Mining, Lecture Notes in Computer Science, Vol. 6318, 2010, pages 271-277.

[10] Nishant, K. P. Sharma, V. Krishna, C. Gupta, KP. Singh, N. Nitin and R. Rastogi, Load Balancing of Nodes in Cloud Using Ant Colony Optimization. In proc. 14Th International Conference on Computer Modelling and Simulation (UKSim), IEEE, pp: 3-8, March 2012.