ISSN 2277-3061
Volume 15 Number 12
International Journal of Computer and Technology

CREDIT BASED LOAD BALANCING IN CLOUD ENVIRONMENT: A REVIEW

Jagdeep Singh [(1)], Mr. Pawan Luthra [(2)]
[(1)] Research Scholar, Department of Computer Science & Engineering, SBSSTC, Ferozepur, Punjab.
jagdeep9417@gmail.com
[(2)] Assistant Professor, Department of Computer Science & Engineering, SBSSTC, Ferozepur, Punjab.
pawanluthra81@gmail.com

ABSTRACT

In present days cloud computing is one of the greatest platform which provides storage of data in very lower cost and available for all time over the internet.But it has more critical issue like security, load management and fault tolerance. In this paper we are discussing 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 load over the different nodes which provides good resource utilization when nodes are overloaded with job. Load balancing has to handle the load when one node is overloaded. When 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.

Keywords

Cloud Computing, Load Balacing, Virtual Machine, Data Center, Data Center Broker.

INTRODUCTION

Cloud computing is the use of the pooled computing resources accessible over Internet. Computing resources can be hardware or software. Cloud derives its name from the cloud shaped symbol representing Internet, as it is used as an abstraction for its complex infrastructure. It provides services as per requirement. It allows user to customize, configure, and deploy cloud services. It offers services as per payment. Cloud provides resources over Internet using virtualizat ion technology, multi-tenancy, web services, etc. Virtualizat ion provides abstraction of independent hardware access to each virtual machine. Multi-tenancy allows the same software platform to be shared by multiple applications. Multi-tenancy is important for developing software as a service application. Applications communicate over the Internet using web services [1]. Load balancing is one of the central issues in cloud computing. It is a mechanis m that distributes the dynamic local workload evenly across all the nodes in the whole cloud to avoid a situation where some nodes are heavily loaded while others are idle or doing little work. It helps to achieve a high user satisfaction and resource utilizat ion ratio, hence improving the overall performance and resource utility of the system. There are four deployment models of cloud. Cloud Computing is the biggest technology advancement now a days. It has taken computing in initial to the next level. Cloud computing provides the information technology as a service. Cloud computing uses the internet and the central remote servers to support different data and applications. It is an internet based technology. It allows the users to find their personal files at any computer with internet access. Cloud computing is flexible in nature. It allocates the resources on the authority request. [2] Cloud computing provides the act of uniting. It is an emerging technology, that is used to provide various computing and storage services over the Internet. In cloud computing, the internet is viewed as a cloud. By the use of cloud computing, the capital and operational costs can be cut.

In the older days every company was to license their software through CDs DVDs and when it was to come on upgrading, they were to face lots of problems. When cloud computing comes as a service part like rental the cost of supplying and vendor system could be reduced, where the software comes to any organization directly. Cloud computing incorporates the infrastructure, platform, and software as services. These service providers rent data center hardware and software to deliver storage and computing services through the Internet. Internet users can receive services from a cloud as if they were employing a super computer which be using cloud computing. To storing data in the cloud instead of on their own devices and it making ubiquitous data access possible. They can run their applications on much more powerful cloud computing platforms with software deployed in the cloud which mitigating the users burden of full software installation and continual upgrade on their local devices.

SERVICE MODELS OF CLOUD COMPUTING

Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. It is composed of five essential characteristics, three service models, and four deployment models [1] :

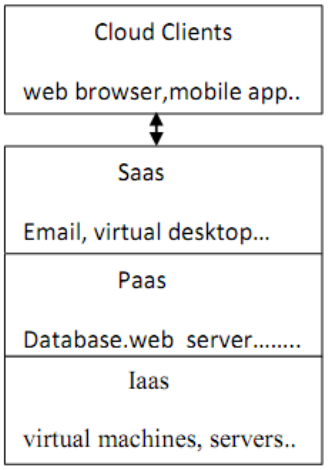

Figure 1. Service Models

1. Cloud Software as a Service (SaaS) : The capability provided to the consumer is to use the provider's applications running on a cloud infrastructure. The applications are accessible from various client devices through a thin client interface such as a web browser (e.g., web-based email). The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user-specific application configuration settings.
2. Cloud Platform as a Service (PaaS) : The capability provided to the consumer is to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages and tools supported by the provider. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage, but has control over the deployed applications and possibly application hosting environment configurations.
3. Cloud Infrastructure as a Service (IaaS) : The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, deployed applications, and possibly limited control of select networking components (e.g., host firewalls).

CLOUD CHARACTERISITCS

Cloud computing possess the following key characteristics [1] :

1. On-demand self-service : A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service's provider.

2. Broad network access : Cloud computing provide the users with various capabilities over the network which are accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, laptops etc.)

3. Resource pooling : The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. Examples of resources include storage, processing, memory, network bandwidth, and virtual machines.

4. Rapid elasticity : Capabilities can be rapidly and elastically provisioned, in some cases automatically, to quickly scale out, and rapidly released to quickly scale in. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be purchased in any quantity at any time.

5. Measured Service : Cloud systems automatically control and optimize resource use by leveraging a metering capability1 at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts). Resource usage can be monitored, controlled, and reported, providing transparency for both the provider and consumer of the utilized service.

TYPES OF CLOUD

There are four types of clouds [1]

Private cloud: The cloud infrastructure is operated solely for an organization. It may be managed by the organization or a third party and may exist on premise or off premise.

Community cloud: The cloud infrastructure is shared by several organizations and supports a specific community that has shared concerns (e.g., mission, security requirements, policy, and compliance considerations). It may be managed by the organizations or a third party and may exist on premise or off premise.

Public cloud: The cloud infrastructure is made available to the general public or a large industry group and is owned by an organization selling cloud services.

Hybrid cloud: The cloud infrastructure is a composition of two or more clouds(private, community or public) that remain unique entities but are bound together by standardized or proprietary technology that enables data and application portability.

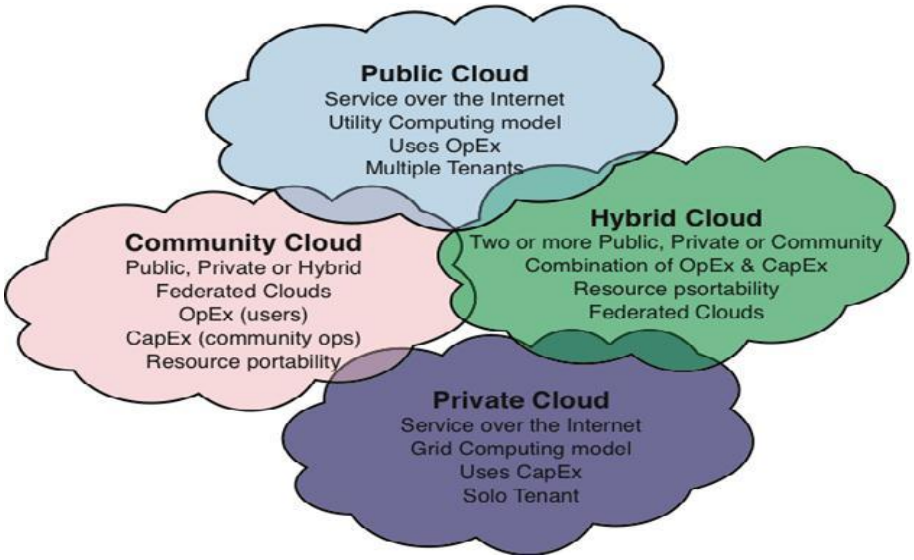

Figure 2. Types of Cloud

Load Balancing

It is a process of reassigning the total load to the individual nodes of the collective system to make resource utilization effective and to improve the response time of the job, simultaneously removing a condition in which some of the nodes are over loaded while some others are under loaded. A 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. The important things to consider while developing such algorithm are : estimation of load, comparison of load, stability of different system, performance of system, interaction between the nodes, nature of work to be transferred, selecting of nodes and many other ones [4] . This load considered can be in terms of CPU load, amount of memory used, delay or Network load.

Goals of Load balancing

As given in [4], the goals of load balancing are :

- To improve the performance substantially
- To have a backup plan in case the system fails even partially
- To maintain the system stability
- To accommodate future modification in the system

Policies or Strategies in dynamic load balancing

There are 4 policies [4]:

- Transfer Policy: The part of the dynamic load balancing algorithm which selects a job for transferring from a local node to a remote node is reffered to as Transfer policy or Transfer strategy.
- Selection Policy: It specifies the processors involved in the load exchange (processor matching)
- Location Policy: The part of the load balancing algorithm which selects a destination node for a transferred task is reffered to as location policy or Location strategy.
- Information Policy: The part of the dynamic load balancing algorithm responsible for collecting information about the nodes in the system is reffered to as Information policy or Information strategy.

METRICS FOR LOAD BALANCING IN CLOUD

Various metrics considered in existing load balancing techniques in cloud computing are discussed below-

• **Scalability** is the ability of an algorithm to perform load balancing for a system with any finite number of nodes. This metric should be improved.

• **Resource Utilization** is used to check the utilization of re-sources. It should be optimized for an efficient load balancing.

• **Performance** 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.

• **Response Time** is the amount of time taken to respond by a particular load balancing algorithm in a distributed system. This parameter should be minimized.

• **Overhead Associated** 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 interprocess communication. This should be minimized so that a load balancing technique can work efficiently.

CLOUDSIM

Use of cloud computing is increasing at a very fast pace everywhere because it turns the capital expenditure cost into operational cost. In addition to that, use of simulation tools is considered a better option in spite of being on the real cloud as performing experiments in a controlled and dependent environment is difficult and costly to handle [2].Moreover, effective resource utilization is not possible in case of Cloud. So, we just shift towards cloud simulation tools. Following are the advantages of running simulation tools in cloud:

a. No capital cost involved: As we discussed earlier, that cloud computing makes a shift from capital expenditure cost to operational cost. Having a cloud simulation tool also involves having no installation cost or maintenance cost as well.

b. Leads to better results: Using such tools helps to change inputs and other parameters as well very easily which results in better and efficient output

c. Evaluation of risks at an early stage: Because simulation tools involve no cost while running as is in case of being on cloud, so user can identify and solve any risk that is associated with the design or with any parameter.

Analysing the performance, policies in real cloud is difficult to achieve because of its altering nature, so in such a situation, we can opt for CloudSim. CloudSim is a famous tool that is actually a toolkit for simulation of cloud scenarios [4].CloudSim has been developed as a CloudBus project in Australia [4]. CloudSim actually enables the users to have a proper insight into cloud scenarios without worrying about the low level implementation details [5]. All components in CloudSim communicate through the process of message passing. The lowermost layer is responsible for managing the communication between various components. The second layer has all the sub layers in it that have the main cloud components [6]. In case of CloudSim, user can model the data center,virtual machines allocation, power consumption, network behaviour as well [7].

RESEARCH GAP

The random arrival of load in such an environment can cause some server to be heavily loaded while other server is idle or only lightly loaded. Equally load distributing improves performance by transferring load from heavily loaded server. Efficient scheduling and resource allocation is a critical characteristic of cloud computing based on which the performance of the system is estimated. The considered characteristics have an impact on cost optimization, which can be obtained by improved response time and processing time.

PROBLEM DESCRIPTION

After reading the credit based research papers, we have found the below listed problems :

- The capacity of the VM should be considered before allocating the cloudlet to the virtual machine. There can be a scenario where a cloudlet with high credit can be assigned to a VM of lower capacity.
- The existing papers specify that the system will work only in homogeneous environment where all the virtual machines will carry the similar configurations.
- No sorting mechanism is applied on the virtual machines.

CONCLUSION

Cost and time are the key challenge of every IT engineer to develop products that can enhance the business performance in the cloud based IT sectors. Current strategies lack efficient scheduling and resource allocation techniques leading to increased operational cost and time. This research has a wide scope in point of reducing the load over the server to give out the optimized performance.Cloud Computing has widely been adopted by the industry, though there are many existing issues like Load Balancing, Virtual Machine Migration, Server Consolidation, Energy Management, etc. which have not been fully addressed. Central to these issues is the issue of load balancing, that is required to distribute the excess dynamic local workload evenly to al l the nodes in the whole Cloud to achieve a high user satisfaction and resource utilization ratio. It also ensures that every computing resource is distributed efficiently and fairly. This paper presents a concept of Cloud Computing along with research challenges in load balancing. Major thrust is given on the study of load balancing algorithm, followed by a comparative survey of these above mentioned algorithms in cloud computing with respect to scalability, resource utilization, performance, response time and overhead associated.

REFERENCES

[1] Wu, H.-S., Wang, C.-J., and Xie, J.-Y. (2013a). Terascaler elb-an algorithm of predictionbased elastic load balancing resource management in cloud computing. In Advanced Information Networking and Applications Workshops (WAINA), 2013 27th International Conference on, pages 649-654. IEEE.

[2] Wu, X., Deng, M., Zhang, R., Zeng, B., and Zhou, S. (2013b). A task scheduling algorithm based on qos-driven in cloud computing. Procedia Computer Science, 17:1162-1169.

[3] Sharma, A. and Peddoju, S. K. (2014). Response time based load balancing in cloud computing. In Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2014 International Conference on, pages 1287-1293. IEEE.

[4] Ren, H., Lan, Y., and Yin, C. (2012). The load balancing algorithm in cloud computing environment. In Computer Science and Network Technology (ICCSNT), 2012 2nd International Conference on, pages 925-928. IEEE.

[5] Raju, R., Amudhavel, J., Kannan, N., and Monisha, M. (2014). A bio inspired energy-aware multi objective chiropteran algorithm (eamoca) for hybrid cloud computing environment. In Green Computing Communication and Electrical Engineering (ICGCCEE), 2014 International Conference on, pages 1-5. IEEE.

[6] Mesbahi, M., Rahmani, A. M., and Chronopoulos, A. T. (2014). Cloud light weight: A new solution for load balancing in cloud computing. In Data Science & Engineering (ICDSE), 2014 International Conference on, pages 44-50. IEEE.

[7] Domanal, S. G. and Reddy, G. R. M. (2013). Load balancing in cloud computingusing modified throttled algorithm In Cloud Computing in Emerging Markets (CCEM), 2013 IEEE International Conference on, pages 1-5. IEEE.

[8] Domanal, S. G. and Reddy, G. R. M. (2014). Optimal load balancing in cloud computing by efficient utilization of virtual machines. In Communication Systems and Networks (COMSNETS), 2014 Sixth International Conference on, pages 1-4. IEEE.

[9] Delavar, A. G. and Aryan, Y. (2014). Hsga: a hybrid heuristic algorithm for work flow scheduling in cloud systems. Cluster computing, 17(1):129-137.