Load Balancing Techniques in Cloud Computing: A Study

Deepak B S
NIE, Mysore

Shashikala S V
Dept. of CSE,BGSIT

Radhika K R
Dept of CSE, MIT

ABSTRACT

Cloud Computing is an emerging computing paradigm. It aims to share data, calculations, and service transparently over a scalable network of nodes. Cloud Computing is nothing but a collection of computing resources and services pooled together over internet and is provided to the users on pay-as-needed basis. In cloud computing, load balancing is required to distribute the dynamic local workload evenly across all the nodes. It helps to achieve a high user satisfaction and resource utilization ratio by ensuring an efficient and fair allocation of every computing resource. Proper load balancing aids in minimizing resource consumption, implementing fail-over, enabling scalability, avoiding bottlenecks and over-provisioning etc. It helps in optimal utilization of resources and hence in enhancing the performance of the system. A few existing scheduling algorithms can maintain load balancing and provide better strategies through efficient job scheduling and resource allocation techniques as well. In order to gain maximum profits with optimized load balancing algorithms, it is necessary to utilize resources efficiently. This paper discusses some of the existing load balancing algorithms in cloud computing and also their challenges.

Keywords: Load balancing, cloud computing

1. INTRODUCTION

Cloud computing is emerging as a new paradigm of large scale distributed computing. It has moved computing and data away from desktop to portable PCs into large data centers [1]. It has the capability to harness the power of Internet and wide area network to resources that are available remotely, thereby, providing cost effective solution to the most of the real life requirement. It provides the scalable IT resources such as applications and service, as well as infrastructure on which they operate, over the Internet, as pay-per- use basis to adjust the capacity quickly and easily. It helps to accommodate changes in demand. Thus cloud computing is a framework for enabling a suitable, and resource utility of the system. It also ensures for the fair distribution of work and resources.

Load balancing in computer networks is a technique used to spread workload across multiple network links of computers [2]. It facilitates networks and resources by providing a maximum throughput with minimum time, thus it helps to improve performance by optimally using available resources and helps in minimizing latency and response time. Load balancing is achieved by using multiple resources that is, multiple servers that are able to fulfill a request or by having multiple paths to a resource. Load balancing helps to achieve a high user satisfaction and resource utilization. When one or more components of any service fail, load balancing facilitates continuation of the service by implementing fair-over, that is, it helps in provisioning and de-provisioning of instances of applications without fail. It also ensures that every computing resource is distributed efficiently and fairly [3].

Consumption of resources and conservation of energy is not always a prime focus of discussion in cloud computing. However, resource consumption can be kept to minimum with proper load balancing which not only helps in reducing costs but making enterprise greener. , One of the very important features of cloud computing, is also enabled by load balancing. Hence, improving resource utility and the performance of a distributed system in such a way will reduce the energy consumption and carbon footprints to achieve Green computing [1].

The real world example of load balancing can be a website which has thousands of users at the same time. If not balanced then the users have to face the problem of timeouts, response delays and long processing time. The solutions involve making use of duplicate servers to make the website available by balancing the network traffic.

2. CHARECTERISTICS

According to the National Institute of Standards and Technology (NIST) [4], the characteristics of the cloud computing are discussed as follows:

On-demand Self-service- The customer can avail all computing capabilities, such as server time and network storage, when he needed automatically without requiring human interaction with each service provider.

Broad Network Access- The customers can avail all the services provided by broad band network over the cloud system and standard techniques are used access it, that promote use by different types of users like thin or thick client platforms (e.g., mobile phones, tablets, laptops, and workstations).

Resource Pooling- The multiple customers are served using provider's computing resources that are pooled together using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to customer demand.

There is a sense of location independence in that the customer generally has no control or knowledge over the exact location of the provided resources but may be able to specify location at a higher level of abstraction (e.g., country, state, or datacenter). Examples of resources include storage, processing, memory, and network bandwidth.

Rapid Elasticity- The cloud system can be used to provide and release, resources and platforms automatically in some cases, to scale rapidly outward and inward to accommodate with customer demand. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be appropriated in any quantity at any time for all different types of users.

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

3. LOAD BALANCING

Load balancing is relatively new technique that facilitates networks and resources by providing a maximum throughput with minimum response time[5]. Dividing the traffic between servers, data can be sent and received without major delay. Different kinds of algorithms are available that helps traffic loaded between available servers. A basic example of load balancing in our daily life can be related to websites. If there is no load balancing, users could experience delays, timeouts and possible long system responses. Load balancing solutions usually apply redundant servers which help a better distribution of the communication traffic so that the website availability is conclusively settled [5].

There are many different types of load balancing algorithms available, which can be categorized into two groups. The following section will discuss these two main categories of load balancing algorithms, those categories are as follows.

The load balancing algorithms can of three categories depending up on who initiated the process as given in [6]:

Sender Initiated: If load balancing is initiated by the sender then the algorithm is called sender initiated.

Receiver Initiated If load balancing is initiated by the receiver then the algorithm is called sender initiated.

Symmetric: It is the combination of both sender initiated and receiver initiated.

The load balancing algorithms can be divided into two categories depending on the current state of the system, as given in [6]:

Static: The static Load balancing algorithms doesn't depend on the current state of the system. Prior knowledge of the system is needed.

Dynamic: In the dynamic load balancing algorithms decisions on load balancing will be based on current state of the system. No prior knowledge is needed. So it is better than static approach.

4. LOAD BALANCING TECHNIQUES

4.1 Carton

R. Stanojevic[7] proposed a mechanism CARTON for cloud control the unifies the use of load balancing(LB) and distributed rate limiting(DRL). LB is used to equally distribute the jobs to different Servers so that the associated costs can be minimized and DRL is used to make sure that the resources are distributed in a way to keep a fair resource allocation. With very low computation and communication overhead, this algorithm is simple and easy to implement.

4.2 Compare and Balance

Zhao[8] addressed the problem of intra-cloud load balancing amongst physical hosts by adaptive live migration of virtual machines. A load balancing model is designed and implemented to reduce virtual machines' migration time by shared storage to balance load amongst servers according to their processor or IO usage and to keep virtual machines zero-downtime in the process. A distributed load balancing algorithm COMPARE and BALANCE is also proposed that is based on sampling and reaches equilibrium very fast. This algorithm assures that the migration of VMs is always from high-cost physical hosts to low-cost host but assumes that each physical host has enough memory which is a weak assumption.

4.3 Events-Driven

V. Nae[9] presented an event-driven load balancing algorithm for real time massively multiplayer online games (MMOG). This algorithm after receiving capacity events as input, analyzes its components in context of the resources and the global state of the game session, thereby generating the game session load balancing actions. It is capable of scaling up and down a game session on multiple resources according to the variable user load but has occasional QoS breaches.

4.4 Scheduling Strategy on LB of VM Resource

A scheduling strategy on load balancing of VM resources that uses historical data and current state of the system. This strategy achieves the best load balancing and reduced dynamic migration by using a genetic algorithm. It helps in resolving the issue of load imbalance and high cost of migration thus achieving better resource utilization [11].

4.5 Honeybee Foraging Behavior

M.Randles investigated a decentralized honey-bee based load balancing technique that is a nature- inspired algorithm for self -organization. It achieves global load balancing through local server actions. Performance of the system is enhanced with increased system diversity but throughput is not increased with an increase in system size. It is best suited for the conditions where the diverse population of service types is required [10].

4.6 Biased Random Sampling

M. Randles investigated a distributed and scalable load balancing approach that uses random sampling of the system domain to achieve self organization thus balancing the load across all nodes of the system. The performance of the system is improved with high and similar population of resources thus resulting in an increased throughput by effectively utilizing the increased system resources [10].

4.7Message Oriented Model

Zenon Chaczko presented a model that uses XMPP for load balancing. This technology is open for real time communication between various parties. XMPP clients send presence information to XMPP presence servers and XML streams containing details of presence information of clients produced by these servers. Using a load balancer on the top of an XMPP server allowed incoming requests to be prioritized and handled by a generic service [2].

4.8 Open Flow Model

Hardeep Uppal presented a model in which open flowswitch is used. Open flow switches are like a standard switch with a flow table performing packet lookup and forwarding. The difference lies in how flow rules are inserted and updated inside the switch's flow table [12].

4.9 Self-organized Load Balancing Algorithm

Giuseppe has presented a new method for load balancing in which a node with highest capacity serves as super peers. At first level, algorithm find out the capacity of every peer, i.e., the amount of service requests that peer is able to fulfill in a client time unit. This, in turn, is reflected in the Myconet overlay as the target Bs of peers maintained by a super peer. This way, super peers are well positioned to effectively balance their neighbors' request queues [13].

4.10 Ant Colony Optimization

Ratan Mishra has proposed a model in which Individual ants are behaviorally much unsophisticated insects. They have a very limited memory and exhibit individual behavior that appears to have a large random component. Acting as a collective however, ants manage to perform a variety of complicated tasks with great reliability and consistency [14].

5. METRICS FOR LOAD BALANCING

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 for efficient load balancing.
- **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 inter process communication. This should be minimized so that a load balancing technique can work efficiently.

6. COMPARISON OF EXISTING LOAD BALANCING TECHNIQUES BASED ON VARIOUS METRICS

The various load balancing metrics are discussed in the previous section. In this section we will compare all load balancing techniques according to the metrics discussed earlier. this comparison is shown table 1.

Table 1: Comparison of different load balancing techniques

metric	carton	compare& balance	event driven	vm resources	honey bee	biased random sampling	message oriented	open flow model	self organized	ant colony
Throughput	No	No	No	No	Yes	Yes	No	Yes	No	Yes
Overhead	Yes	Yes	No	Yes	No	No	No	No	No	No
Response Time	No	Yes	No	No	No	No	Yes	No	No	No
Resource utilization	Yes	No	Yes	Yes	No	No	No	Yes	Yes	Yes
performane	No	No	No	No	Yes	Yes	Yes	No	Yes	Yes

7. LOAD BALANCING CHALLENGES IN THE CLOUD COMPUTING

Although cloud computing has been widely adopted. Research in cloud computing is still in its early stages, and some scientific challenges remain unsolved by the scientific community, particularly load balancing challenges [15].

- **Automated service provisioning:** The resources can be allocated or de-allocated automatically this is a key feature of cloud computing called elasticity. How then can we use or release the resources of the cloud, by keeping the same performance as traditional systems and using optimal resources?
- **Migration of Virtual Machines:** With virtualization, an entire machine can be seen as a file or set of files, to unload a physical machine heavily loaded, it is possible to move a virtual machine between physical machines. The main objective is to distribute the load in a datacenter or set of datacenters. How then can we dynamically distribute the load when moving the virtual machine to avoid bottlenecks in Cloud computing systems?

- **Management of Energy:** The benefits that advocate the adoption of the cloud is the economy of scale. Energy saving is a key point that allows a global economy where a set of global resources will be supported by reduced providers rather that each one has its own resources. How then can we use a part of datacenter while keeping acceptable performance?

- **Stored data management:** In the last decade data stored across the network has an exponential increase even for companies by outsourcing their data storage or for individuals, the management of data storage or for individuals, the management of data storage becomes a major challenge for cloud computing. How can we distribute the data to the cloud for optimum storage of data while maintaining fast access?

- Emergence of small data centers for cloud computing: Small datacenters can be more beneficial, cheaper and less energy consumer than large datacenter. Small providers can deliver cloud computing services leading to geo-diversity computing. Load balancing will become a problem on a global scale to ensure an adequate response time with an optimal distribution of resources.

8. CONCLUSION

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 all 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 throughput, resource utilization, performance, response time and overhead associated.

8. REFERENCES

[1] Nidhi Jain Kansal, Inderveer Chana, Cloud Load balancing techniques:A step towards green computing, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 1, January 2012.

[2] Zenon Chaczko, Venkatesh Mahadevan, Shahrzad Aslanzadeh and Christopher Mcdermid, Availabilty and load balancing in cloud computing, 2011 International Conference on Computer and Software Modeling, IPCSIT vol.14 (2011) ACSIT Press, Singapore.

[3] Tanveer Ahmed, Yogendra Singh, Analytic study of load balancing techniques using tool cloud analyst

[4] Peter Mell, Timothy Grance, “The NIST Definition of Cloud Computing”, NIST Special Publication 800-145, September 2011.

[5] R. Shinmonski. Windows 2000 & Windows Server 2003 Clustering and Load balancing. Emeryville. McGraw-Hill professional publishing, CA, USA (2003), p2, 2003.

[6] David Escalante and Andrew J. Korty, Cloud Services: Policy and Assessment, EDUCASE Review, Vol. 46, No.4(2011).

[7] Stanojevic R and Shorten R(2009) IEEE ICC,1-6.

[8] Zhao Y. and Huang W. (2009) 5th International Joint Conference on INC, IMS and IDC, 170-175.

[9] Nae V., Prodan R. and Fahringer T.(2010) 11th IEEE/ACM International Conference on Grid Computing (Grid), 9-17.

[10] Randles M., Lamb D. and Taleb-Bendiab A. (2010) 24th International Conference on Advanced Information Networking and Applications orkshops, 551-556.

[11] Meenakshi Sharma, Pankaj Shrama, Dr. Sandeep Sharma, Efficient Load [6] Hardeep Uppal, dane Brandon, OpenFlow based load balancing.

[12] Hardeep Uppal, dane Brandon, OpenFlow based load balancing.

[13] Giuseppe Valetto, Paul Snyder, Daniel J. Dubois, Elisabetta Di Nitto and Nicolo M. Calcavecchia, A self-organized load balancing algorithm for overlay based decentralized service networks.

[14] Ratan Mishra, Anant jaiswal, Ant colony optimization: A Solution of load balancing in cloud, International Journal of Web & Semantic Technology (IJWesT) Vol.3, No.2, April 2012.

[15] A. Khiyaita, M. Zbakh, H. El Bakkali and Dafir El Kettani, “Load Balancing Cloud Computing: State of Art” , 9778-1-4673-1053-6/12/$31.00, 2012 IEEE.