Cloud Computing Overview with Load Balancing Techniques

Yatendra Sahu
M.Tech Scholar, Dept. of Computer Science & Engineering, MANIT Bhopal, India

R.K. Pateriya
Associate Professor, Dept. of Computer Science & Engineering, MANIT Bhopal, India

ABSTRACT
Nowadays, cloud computing has become a key technology for online allotment of computing resources and online storage of user's data in a lower cost, where computing resources are available all the time, over the internet with pay per use concept. Cloud computing is business oriented concept where computing resources are outsourced by cloud provider to their client, who demand computing online. There is various advantage of cloud computing including virtual computing environment, on-demand services, maximize resource utilization and easy to use services etc. But there are also some critical issues like security, privacy, load management and fault tolerance etc. In this paper we are providing an overview of cloud technology and its components. We are also focusing on load balancing of cloud computing with some of the existing load balancing techniques, which are responsible to manage the load when some node of the cloud system is overloaded and others are under loaded. In computing, the load may be of various types like memory load, CPU load and network load etc. Load balancing is the process of searching overloaded node and transferring the extra load of the overloaded node to other nodes which are under loaded, for improving resource utilization and decreasing server response time of the jobs.

General Terms
Cloud computing, Advances in the field of Information Technology, Recent technology in computing, On-demand services through internet, issues of cloud computing, Load balancing.

Keywords
Cloud computing, virtualization, resource sharing, load balancing.

1. INTRODUCTION
In the field of information technology, cloud computing is a recent trend that moves computing and data away from desktop and portable computers into large data centers [1]. Cloud computing allows everyone to use software and computing services on-demand at anytime, anywhere and anyplace through the internet. Cloud computing mainly deals with computation, software, data access and storage services that may not require end-user knowledge of the physical location and configuration of the system that is delivering the services [2]. The definition of cloud computing provided by National Institute of Standards and Technology (NIST) says that: "Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, data storage, software applications and other computing services) that can be rapidly provisioned and released with minimal management effort or service provider interaction [3]". By sharing of resource the overall cost reduces.

2. HISTORY
The underlying concept of cloud computing was introduced way back in 1960s by John McCarthy [4]. His opinion was that "someday, computation may be organized as a public utility" [4]. In 1966 Douglas Parkhill investigated the characteristics of cloud computing in his book, "the Challenge of the Computer Utility" for the first time [4].

The history of the term cloud is from the telecommunications world, where telecom companies started offering Virtual Private Network (VPN) services [5] with comparable service quality at a lower cost. Initially before VPN, they provided dedicated point-to-point data circuits which were wastage of bandwidth. But by using VPN services, they can switch traffic to balance the utilization of overall network. Cloud computing now extends this to concept to cover servers and network infrastructure [5].

3. MASSIVE INFRASTRUCTURE
The cloud computing have five characteristics, three delivery models, and four deployment models [3]. The five key characteristics represent the first layer in the cloud environment architecture (see Figure1).

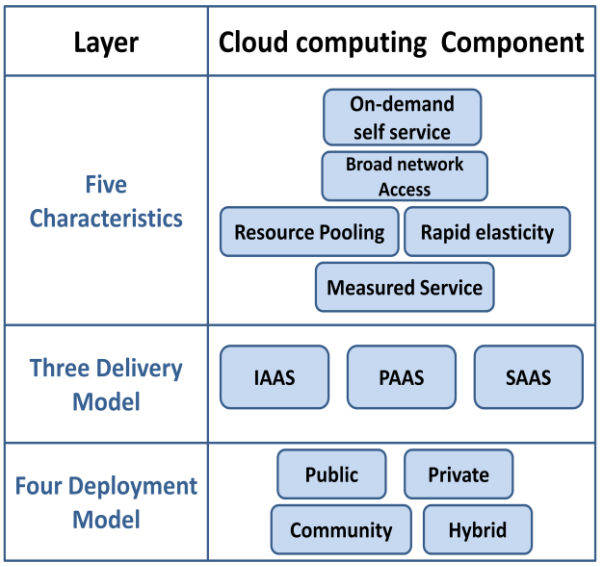

Fig 1: Cloud Environment architecture [6]

3.1 CHARACTERISTICS
Cloud computing is an on-demand, virtualized, cost-effective, elastic, location and device independent, and all time available system. Cost is reduced to a significant level as the infrastructure is provided by a third-party and global utilization of resources which avoids wastage of resources and computing power. The main goal of cloud computing is to make a better use of distributed resources by combine them to achieve higher throughput and be able to solve large scale

computation problems. Some of the most important key characteristics are:

3.1.1 On-demand Self Service

Computing resources are provided online according to the client requirement at specific time without any human interaction. In cloud computing, users access the data, applications or any other services with the help of a browser only, regardless of the other software and hardware.

3.1.2 Broad Network Access

Cloud users have broad area of cloud services those are accessible through internet. There is no dependency of client platform to access cloud services. Services are always on, anywhere, anytime and anyplace.

3.1.3 Resource Pooling

The cloud computing resources are pooled to serve multiple consumers according to consumer demand.

3.1.4 Measured Service

Cloud users do not need to control and optimize computing resource because all are automatically managed by cloud system. Resource usage can be monitored, controlled, and reported for providing transparency for both the provider and consumer of the utilized service.

3.1.5 Selection of Provider

Selection of cloud service provider is the key to get good service. So according to their choice and their knowledge of cloud provider, users can select the right service provider. One must make sure that the provider is reliable and well-reputed for their customer service and also should have a proven track record in IT- related ventures.

3.2 Cloud Delivery Model

There are three cloud delivery models, Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS) deployed as public, private, community, and hybrid clouds [7].

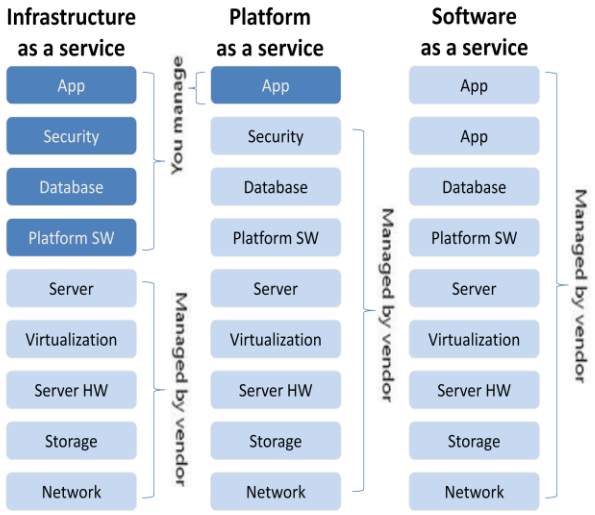

Fig 2: Cloud Delivery Model

In the above diagram we see management permission of cloud user and cloud vender over different layer of cloud system.

3.2.1 Software As A Service

Cloud computing deliver a SAAS service where user do not need to manage installation and configuration of any hardware or software. All the installation and configuration of services are managed by vendor or cloud provider. Examples are Google Online office, Google Docs, Email cloud etc. Main advantages are reduced up-front cost, potential for reduced lifetime cost, Elimination of licensing risk, Elimination of version compatibility and reduced hardware footprint etc. The main disadvantage of SAAS is billing management, synchronization of client and vendor migrations etc.

3.2.2 Platform As A Service

It is a delivery of a computing platform over the web where user can create and install their own application as their need. Configuration of computing platform and server is managed by vendor or cloud provider. Web applications can be created quickly without complexity of buying and managing the storage server, database, and other software/hardware. Example of PAAS is Google App Engine. PASS Enables developers to focus on application code, Instant global platform Elimination of H/W dependencies and capacity concerns Inherent scalability. Disadvantages are Strong governance required to prevent lines of business from building applications without IT involvement.

3.2.3 Infrastructure As A Service

Infrastructure of servers, software, and network equipment is provided as an on-demand service by the cloud provider. It can be used to avoid buying, housing, and managing the basic hardware and software infrastructure components, scales up and down quickly to meet demand. Example: Amazon EC2.

3.3 Cloud Deployment Model

There are four cloud deployment models as public, private, hybrid and community clouds [7].

3.3.1 Public Cloud

Public cloud allows users to access the cloud publicly via interfaces using web browsers. Users need to pay only for the time duration they use the service, i.e., pay-per-use. This can be compared to the electricity system which we receive at our homes. We pay only for the amount of that we use. The same concept applies here. This helps in reducing the operation costs on IT expenditure. However public clouds are less secure compared to other cloud models as all the applications and data on the public cloud are more prone to malicious attacks. The solution to this can be that security checks be implemented.

3.3.2 Private Cloud

A private clouds operation is within an organization's internal enterprise data center. The main advantage here is that it is easier to manage security, maintenance and upgrades and also provides more control over the deployment and use. Private cloud can be compared to intranet. Compared to public cloud where all the resources and applications were managed by the service provider, in private cloud these services are pooled together and made available for the users at the organizational level. The resources and applications are managed by the organization itself. Security is enhanced here as only the organizations' users' have access to the private cloud.

3.3.3 Hybrid Cloud

It is a combination of public cloud and private cloud. In this model a private cloud is linked to one or more external cloud services. It is more secure way to control data and applications and allows the party to access information over the internet. It enables the organization to serve its needs in the private cloud and if some occasional need occurs it asks the public cloud for intensive computing resources.

3.3.4 Community Cloud

When many organizations jointly construct and share a cloud infrastructure, their requirements and policies then such a cloud model is called as a community cloud. The cloud infrastructure could be hosted by a third-party provider or within one of the organizations in the community.

4. CLOUD ARCHITECTURE

Cloud computing system can be divided into two sections as front end and back end [9].

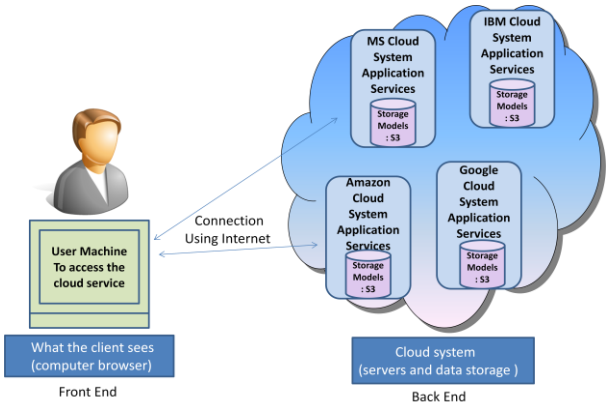

Fig 3: Cloud computing system architecture

They both are connected with each other through a network, usually the internet. Front end is what the client (user) sees whereas the back end is the cloud system. Front end has the client's computer and the application required to access the cloud (Browser) and the back has the cloud computing services like on-demand computing and data storage from various servers.

The difference between traditional system and cloud system is represented in the next diagram. Using hypervisor [10], also called virtual machine manager (VMM), is one of many hardware virtualization techniques allowing multiple operating systems, termed guests, to run concurrently on a host computer. It is so named because it is conceptually one level higher than a supervisory program.

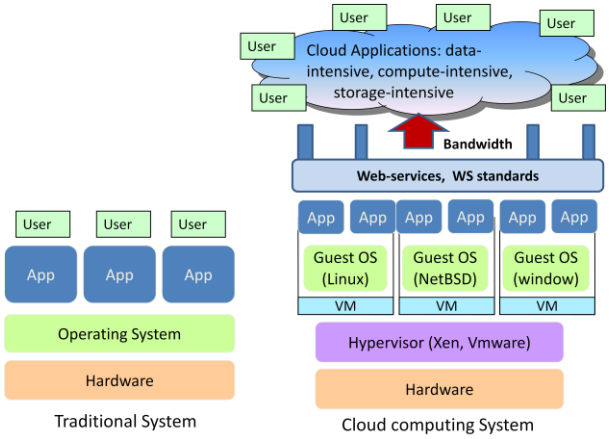

Fig 5: Compare b/w traditional system and cloud system [9]

The hypervisor [11] presents to the guest operating systems as a virtual operating platform and manages the execution of the guest operating systems. Multiple instances of a variety of operating systems may share the virtualized hardware resources. Hypervisors are very commonly installed on server hardware, with the function of running guest operating systems, that themselves act as servers.

5. CHALLENGES IN CLOUD [12]

- Security
- Efficient load balancing
- Performance Monitoring
- Consistent & Robust Service abstractions
- Resource Scheduling
- Scale and QoS management
- Interoperability & Portability
- Requires a constant & speedy Internet connection.

6. LOAD BALANCING IN CLOUD

Load balancing is the process of reassigning the total load to the individual nodes of the collective system to make effective resource utilization and to improve the response time of the jobs, simultaneously removing a condition in which some of the nodes are over loaded while some others are under loaded. Depending on system state, load balancing algorithms divided into two types as static and dynamic. A load balancing algorithm which is dynamic in nature, does not consider previous state or behavior of the system, that is, it depends on the present behavior of the system. Depending on who initiated the process, load balancing algorithms can be divided into three types as sender Initiated, receiver Initiated and symmetric.

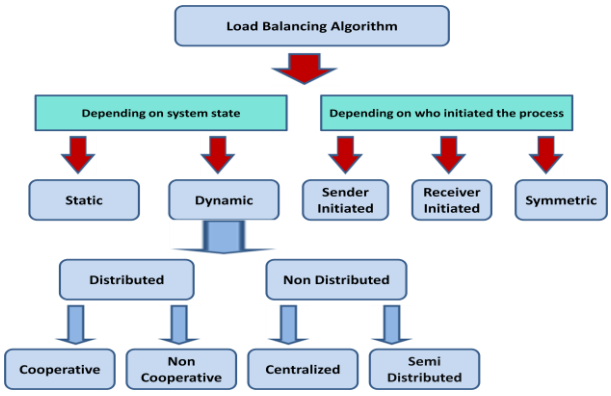

Fig 5: Classification of load balancing algorithms [13]

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. This load considered can be in terms of CPU load, amount of memory used, delay or Network load.

7. DYNAMIC LOAD BALANCING

Dynamic load balancing can be categorized in two types as distributed and non-distributed. In distributed algorithms all nodes present in the system execute the algorithm and the task of load balancing is shared among them [13]. Nodes interaction can take two forms: cooperative and non-cooperative [13]. In the cooperative, the nodes work side-by-side to achieve a common objective to improve the overall response time, etc. In the non-cooperative, each node works independently toward a goal to improve the response time of a local task.

In non-distributed type, either one node or a group of nodes do the task of load balancing. Non-distributed dynamic load balancing algorithms can take two forms: centralized and semi-distributed. In the first form, the load balancing algorithm is executed only by a single node in the whole system: the central node. In semi-distributed form, nodes of the system are partitioned into clusters, where the load balancing in each cluster is of centralized form. A central node is elected in each cluster by appropriate election technique which takes care of load balancing within that cluster [13].

7.1 HONEYBEE FORAGING

This algorithm [14] is derived from the behavior of honey bees. The honey bees use a procedure for finding and reaping food that is also applicable on distributed system to balance the load. There is a class of bees that known as forager bees, forage for food sources and after finding one they come back to give the idea of the quality or quantity of food with its distance from the beehive. The class of bees known as Scout bees then follows the foragers to reap it. After return they gives an idea of how much food is remain and hence results in more exploitation or abandonment of the food source. It is a decentralized loads balancing technique based on natural procedure. System performance is enhanced with increased system dissimilarities but throughput is not increased with an increase in system size [15]. The best performance of this algorithm is shown on the conditions where the multiform types of services are required.

7.2 BIASED RANDOM SAMPLING

This algorithm [15] is based on the construction of the virtual graph having connectivity between the all nodes of the system where each node of the graph is corresponding to the node computer of the cloud system. Edges b/w nodes are two types as Incoming edge and outgoing edge that is used to consider the load of particular system and also allotment the resources of the node. It is scalable technique to balance the load of the cloud system. It is also reliable and effective load balancing approach that is mainly developed to balance the load of distributed system [15]. Load balancing is achieved without the need to monitor the nodes for their resources availability.

7.3 ACTIVE CLUSTERING

Active Clustering [15] works on the concept where same type nodes of the system are grouped together and working on these groups. It works like as self-aggregation load balancing technique where network is rewired to balance the load of the system. Systems optimize using similar job assignments by connecting similar services [15]. The performance of the system is enhanced with high resources thereby increasing the throughput by using these resources effectively.

7.4 COMPARE AND BALANCE

This algorithm [16] uses the concept of compare and balance to reach an equilibrium condition and manage unbalanced system's load. On the basis of probability (no. of virtual machine running on the current host and whole cloud system), current host randomly select a host and compare their load. If load of current host is more than the selected host, it transfers extra load to that node. Each host of the system performs the same procedure. This load balancing model is also designed and implemented to reduce virtual machines' migration time by shared storage memory.

8. CONCLUSION

This paper discuss a new wave of online services in the field of information technology: cloud computing with its challenges. In cloud computing, there are infinite computing capabilities with attractive pay-per-use scheme. Cloud computing provide everything as a service to their users, like as: storage of data as a service, application software as a service, computing platform as a service and computing infrastructure as a service etc. However this wave still needs to resolve some of its existing issues with urgency. One of the major issues of cloud computing is system load balancing, because overloading of a particular node makes it slow down resulting poor system efficiency. So there is always a requirement of efficient load balancing algorithms for improving the utilization of computing resource.

9. ACKNOWLEDGMENTS

The Success of this research work would have been uncertain without the help and guidance of a dedicated group of people in our institute MANIT Bhopal. We would like to express our true and sincere acknowledgements as the appreciation for their contributions, encouragement and support. The researchers also wish to express gratitude and warmest appreciation to people, who, in any way have contributed and inspired the researchers.

10. REFERENCES

[1] Marios D. Dikaiakos, George Pall is, Dimitrios Katsaros, Pankaj Mehra, Athena Vakali, 2009 "Cloud computing : Distributed Internet Computing for IT and Scientific Research" IEEE Internet Computing, Published by the IEEE Computer Society.

[2] Panagiotis Kalagiakos, Panagiotis Karampelas, 2011 "Cloud Computing Learning" IEEE, Hellenic American University Manchester, N.H. - U.S.A, pp: 7/11.

[3] Peter Mell,Timothy Grance, The NIST Definition of "Cloud Computing" National Institute of Standards and Technology - Computer Security Resource Center-www.csrc.nist.gov.

[4] Cloud Computing, accessed (23/01/2013), from http://en.wikipedia.org/wiki/Cloud_computing.

[5] John Harauz, Lorti M. Kaufinan. Bruce Potter, 2009 "Data Security in the World of Cloud Computing" IEEE Security & Privacy, Copublished by the IEEE Computer and Reliability Societies.

[6] Mohammed A. AlZain, Eric Pardede, Ben Soh, 2012 "Cloud Computing Security: From Single to Multi-Clouds" 45th Hawaii International Conference on System Sciences, IEEE, pp: 7/12.

[7] Yashpalsinh Jadeja, Kirit Modi, 2012 "Cloud Computing-Concepts, Architecture and Challenges" International Conference on Computing, Electronics and Electrical Technologies, IEEE, pp: 4/12.

[8] Bhaskar Prasad Rimal, Eunmi Choi, 2009 "A taxonomy and survey of cloud computing systems" Fifth International Joint Conference on INC, IMS and IDC, published by IEEE Computer Society.

[9] Cloud Computing Architecture, accessed (27/01/2013), from http://communication.howstuffworks.com/cloud-computingl.htm

[10] VMware, Cloud computing Virtual infrastructure and virtual machine manager, accessed (25/01/2013), from

http://www.vmware.com/virtualization/virtual-infrastructure.html.

[11] Xen ,virtual machine manager in Cloud computing, accessed 11/01/2013, from http://www.xen.org.

[12] Ramgovind S, Eloff MM, Smith E, 2010 "The management of security in cloud computing" IEEE.

[13] Ali M. Alakeel, 2010 "A Guide to Dynamic Load Balancing in Distributed Computer Systems" IJCSNS International Journal of Computer Science and Network Security, VOL.10 No.6.

[14] Nakrani and C. Tovey, 2004 "On Honey Bees and Dynamic Server Allocation in Internet Hosting Centers" Adaptive Behavior 12, pp: 223-240.

[15] Martin Randles, David Lamb, A. Taleb-Bendiab, 2010 "A Comparative Study into Distributed Load balancing Algorithms for Cloud Computing" IEEE 24th International Conference on Advanced Information Networking and Applications Workshops, pp: 1/10.

[16] Yi Zhao, Wenlong Huang, 2009 "Adaptive Distributed Load Balancing Algorithm based on Live Migration of Virtual Machines in Cloud" Fifth International Joint Conference on INC, IMS and IDC, pp: 6/09.

[17] Balazs Gerofi, Hajime Fujita and Yutaka Ishikawa, 2010 "An Efficient Process Live Migration Mechanism for Load Balanced Distributed Virtual Environments", IEEE International Conference on Cluster Computing, pp: 1/10.