A Study of Effective Load Balancing Approaches in Cloud Computing

R.R. Kotkondawar
PG Scholar, Department of Computer Engineering, DBATU, Lonere-Raigad, India

P.A. Khaire
Department of Computer Science and Engineering, DBACER, Nagpur, India

M.C. Akewar
Department of Computer Technology, YCCE, Nagpur, India

Y.N. Patil
Department of Computer Engineering, DBATU, Lonere-Raigad, India

ABSTRACT
Cloud computing is the most recent technology in today's world of computing and it overcomes deficiencies of traditional ways of computing. Cloud computing is a new way of providing the essential services to cloud users on "Pay As You Go" basis. Cloud computing provides different features like on demand access, flexibility, instant response, pay per use etc. to customers. In order to provide all these features to cloud users, cloud computing systems must be structured and managed efficiently to provide the Quality of Services (QOS) to users. Various technological concepts such as abstraction and virtualization are used that hides the implementation details from an average cloud user. Cloud load balancing plays a very important role in providing all the cloud features to users which is the main topic of interest in our research. Different archictures apply altogether different load balancing algorithms. This paper includes the Study of different approaches of effective management of cloud systems. The study includes load balancing approaches in different system architectures like Centralized, Distributed and Cluster based architecture. Finally various algorithms have been compared based on the different parameters like response time, efficiency and throughput etc.

General Terms
Cloud Computing, Algorithms, Clustering

Keywords
Load balancing, QOS, Response time

1. INTRODUCTION
Cloud computing in the broad sense means the computing performed on the web or internet. Each of the following Section gives the details of various aspects of cloud computing. Cloud computing provides numerous major benefits on adopting it. Some most important of them are given below [3]

• Cloud Computing enables economies of scale. On the provider side, this leads to a higher productivity in provisioning infrastructure, software and platform as a service and a higher survivability when difficulties occur. On the user side, these economies of scale decrease investment and running costs [3].

• It allows organizations to focus on their core competencies in a sustainable manner. Non-IT-service providers can sustainably out-source the IT services they need for their business activities. In contrast to conventional outsourcing, it makes sense for Cloud Computing service providers to continually reinvest in the modernization of the services they provide [3].

•Cloud computing breaks the barrier of being a computer literate acquainted with all the technologies related with computers. This facility allows an average cloud user to enjoy cloud services without being an expert in all these things. This is again an advantage of cloud computing that users don't need to be specialized or educated about how cloud works or implemented.

•Cloud computing services can be accessed from any device. It can be our own desktop, laptops, and tablets or can be simply a smartphone like blackberry or a windows compatible phone. Means cloud computing services are device independent.

•Cloud computing services can be accessed from any place (location independent) and any time whenever a cloud user wants to access these services from cloud.

1.1 Definition of Cloud Computing
Many definitions are proposed for cloud computing. Here some of them are given here,

1) Cloud is a parallel and distributed computing system consisting of collection of the interconnected and virtual computers that are dynamically provisioned and presented as one or more unified computing resources based on the "service-level-agreements" (SLA) that are established through the negotiation between the service provider and customer [1].

2) NIST definition 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 inter-action [20].

3) 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), so we use that term. The datacenter hardware and software is what we will call a cloud [19].

1.2 Features of Cloud Computing
Cloud computing shows various special characteristics which proved superior when compared with other computing platforms. Here some of them have been listed in relation to the load balancing aspect of cloud computing.

• *Resource pooling*- Cloud computing being the biggest platform to provide the on demand services, the cloud service providers use certain technologies such as virtualization and multi tenancy to make all the necessary resources readily available to the numerous consumers or the cloud users who are paying for their needs.

• *Rapid elasticity*- The flexibility and rapid elasticity allows the system to scale up and scale down quickly according to the needs of a cloud user and also the facility to release the resources when no longer needed. Thus an Elastic cloud computing system should always employ a very efficient load balancing strategy for scaling the loads up and down inside the system.

• *Scalability*- Cloud computing provides resources and services for users on demand. The resources are scalable over several data centers. In order to achieve a highly scalable system, balancing of the loads when the load increases at a large extent and a cloud user demands more resources online rapidly is very important.

• *Efficiency*-An efficient cloud computing system should work for all the possible configurations of cloud where users are requesting the resources on any extent unknown to the cloud service providers providing the important features like rapid elasticity and high scalability with the much needed fault tolerance. A proper distribution of tasks among the processors can achieve these features for the cloud systems.

• *Dynamic Resource Allocation*-Cloud computing systems are allocating the loads across the system either statically or dynamically. A dynamic resource allocation policy proves to be better than the static one to sustain the dynamic requirements of a cloud user. An effective load balancing technique applied on the cloud system will solve the purpose to provide the QOS to cloud users.

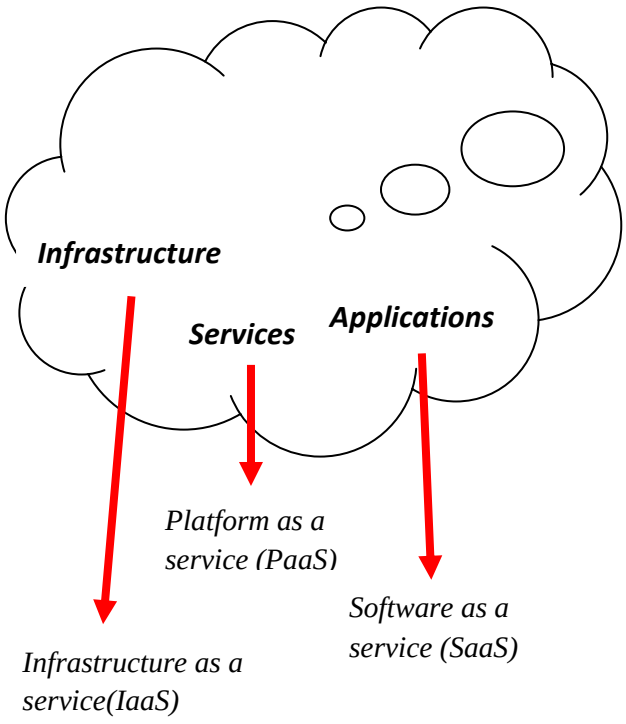

Fig 1: Cloud Layered Model

1.3 Services Provided By Cloud

Service in the broad sense means the way in which one is adopting the cloud facilities. Everything on cloud is given as a service. There are three types of services provided over the cloud which allows to define three cloud service models as follows.

1) Infrastructure as a service (IaaS)- IaaS is the lower most layer. It refers to the operating system and virtualization. The cloud users will be allowed to use a dedicated CPU and the storage resources such as memory virtually depending upon their accountability [12].

2) Platform as a service (PaaS)- PaaS is the middle layer referring to the platform used for implementing a particular service like programming models and environment, method of execution employed, database and the web sources etc. [12].

3) *Software as a service (SaaS)*- SaaS is the uppermost layer of the cloud architecture. It is the most important layer from user's perspective. SaaS offers a whole application as a service on demand. The cloud provides the application software to the customers that can be easily installed and used without knowledge of the method employed for software development on the cloud [12].

As described in the three layered architecture, cloud provides all three types of services to the users that is IaaS, PaaS and SaaS on demand basis. Customers are just paying for any of these services on the cloud. Different techniques such as virtualization and multi tenancy are employed for this purpose to provide these resources and QOS to users or customers. The most important service being the SaaS as it provides a whole application as a service.

As mentioned in [12] the cloud users can access these software's from cloud clients and are prevented from directly accessing infrastructures of the cloud and the platform on which the application is running that is the users would only be accessing the software online and storing the data back eradicating the need of locally installing the applications on users machine, thus minimizing the burden of software maintenance and support for various levels of users accountability.

2. DIFFERENT ARCHITECTURES EMPLOYING CLOUD COMPUTING-

Features of cloud computing can be easily employed on different computing platforms or architectures like centralized. client-server, distributed, peer to peer etc. These different architectures employ altogether different load balancing algorithms. Hence selecting a platform first is necessary before developing a load balancing algorithm. Researchers are still doing their work in this area. In this paper mainly three architectures which are Distributed architecture, Clustered architecture and the Centralized architecture are considered. Further the different load balancing techniques employed by each of these architectures have been discussed and compared.

2.1 Distributed Architecture

In case of distributed architecture of cloud computing, distributed computing systems are used to solve computational problems. As mentioned in [8], a heterogeneous collection of computers that work together forms a distributed computing system. Here several autonomous entities are present each of which has its own local memory and communication among these entities is done through the message passing system. Every entity is acting as a server having same copies of data and resources that may be requested by the users. Resource sharing is possible in this architecture. Load balancing here is the function of dispatching the several tasks to each of the nodes to share load of complete system. Here, the two load

balancing strategies have been discussed as mentioned in [13] and [14] further with respect to this architecture.

2.2 Clustered Architecture

In case of clustered architecture of cloud computing, clustered computing systems are used to solve computational problem. Here a group of linked computers that are tightly coupled with the high speed networking and work closely together. In many respects they form a single computer, provide a single image illusion and operate mostly in a shared memory mode. Computers that form a cluster are homogeneous in their operating system and hardware specifications and contained in a single location in contrast to the distributed architecture as mentioned in [8]. Load balancing here is again the task of dispatching the tasks to the individual nodes inside the system. In this paper two load balancing strategies have been discussed as mentioned in [15] and [18] further with respect to this architecture.

2.3 Centralized Architecture

The centralized architecture of cloud computing uses the centralized form of computing. Centralized computing follows a master-slave relationship inside a system. Here a central server is Master and the remaining can be thought of as the slave processors. The cloud users requests services through these servers that are accepted by the central server or node. Unlike the clustered or distributed architecture, resources are granted through the central server only. In this system load balancing is done on the central server only and is assigning the tasks to individual slave processors equally to maintain balance inside the system.

In this paper the load balancing strategies as mentioned in [12], [16] and [17] have been discussed with respect to this architecture.

3. LOAD BALANCING

Load balancing is the main issue in cloud computing. In cloud computing, load balancing basically means adjusting the loads across the nodes forming the cloud which may be the CPUs, network links or other resources. It is a mechanism 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 as mentioned in [12].As mentioned in [4], the computing power of any distributed system can be realized by allowing its "Computational Elements" (CEs) or nodes to work co-operatively so that large loads are distributed or allocated among them in fair and the effective manner."Any strategy for load distribution among the CEs is known as Load Balancing."

3.1 Virtualization and Load Balancing

The term virtualization broadly describes the separation of a resource or request for a service from the underlying physical delivery of that service. With virtual memory, for example, computer software gains access to more memory than is physically installed, via the background swapping of data to disk storage. Similarly, virtualization techniques can be applied to other IT infrastructure layers including networks, storage, laptop or server hardware, operating systems and applications. The cloud is the virtualization of resources that maintains and manages itself. It is done by using the virtual machines or VMs [21].

The advantages of using the virtual machines as mentioned in [22] are

1. Instant provisioning-fast scalability
2. Live migration is possible
3. Load balancing in datacenters is possible
4. Low downtime for maintenance
5. Security and fault isolation

Both the load balancing and virtualization are closely related. Some of the features of cloud computing systems like fast response time, elasticity are achieved by load balancing techniques through virtualization. Hence both load balancing and the virtualization are equally important.

3.2 Load Balancing in Different Architectures

The pros and cons of each of the three architectures explained above have been discussed here. Centralized architecture is useful when the number of the slave processors or clients is less and is well managed by the central server. But as the number of the clients increase main server may find it difficult to balance the system loads accordingly. Also there is a big issue of security and fault tolerance when the central server fails due to some reasons and as a result the slave processors also fails and the system is no more a reliable system. In contrast to this distributed systems are open and scalable. The computers in the system can come and go and the distributed system server must assign the work and collect the output from the systems as they arrive and leave. The system is more reliable than the centralized architecture in the sense that failures of one or two servers do not lead to the system failure at all. Same is the scenario with clustered architecture only the difference being that the system is located in the same geographical location whereas in other two architectures the locations of the nodes in the cloud may be different. A distributed system can be considered more reliable and hence load balancing proves to be more prominent in this system.

4. LITERATURE STUDY

In this section of paper the approaches used for load balancing in all of the three architectures discussed so far are compared.

4.1 Distributed Systems

In Distributed environment two algorithms are considered as mentioned in [13] and [14] for the distributed environment. A decentralized content aware load balancing algorithm as mentioned in [13] uses the content information to narrow down the search. It uses the unique and special property of nodes to help the scheduler to decide the best node for processing the request. This algorithm improves the searching performance hence the overall performance and reduces the idle time of the nodes.

The load balancing algorithm for distributed services as mentioned in [14] uses a protocol to limit the redirection rates to avoid the remote server overloading. It improves the response time. The mean response time is 29% smaller than Round robin (RR) algorithm. Only disadvantage being that a middleware is always needed to support the protocol used.

4.2 Clustered Systems

In Clustered environment two algorithms are considered as mentioned in [15] and [18]. Active clustering as in [15] is designed for large scale cloud systems and optimizes the job assignment by connecting similar services by local re-wiring. It performs better with high resources and utilizes the increased system resources to increase throughput. Its disadvantage being system degrades as system diversity increases. A resource aware scheduling algorithm (RASA) as

mentioned in [18] for the clustered or grid based environment achieves scheduling in batch mode and used to reduce the make span.

4.3 Centralized Systems

In Centralized environment three algorithms are considered as mentioned in [12], [16] and [17]. As mentioned in [12], Prof. A.A Jaiswal and *et al.* proposed two algorithms for the design of virtual servers that are better in terms of their efficiency and dynamic response time. The algorithm2 has a comparatively low CPU overhead as it is based on centralized non-distributed scheme and employs minimum messages being exchanged among nodes and has a higher throughput. The two other algorithms are central queue algorithm as in [16] and central manager algorithm as in [17]. Both these algorithms are fault tolerant. The central queue algorithm provides overhead rejection while the central manager algorithm does not provide it. The forecasting accuracy of central manager algorithm is more as compared to central queue algorithm. Comparison is given in Table 1.

Table 1. Comparison of Approaches

Author , Year And Publication	**Title**	**Architecture**	**Advantages/Disadvantages**
W.Leinberger and *et al.* 2000 (IEEE)	Central Queue Algorithm	Centralized Systems	1.Better in terms of fault tolerance 2.Provides the facility for overload rejection 3.Forecasting accuracy is small
P. L. McEntire and *et al.* 1984 (IEEE)	Central Manager Algorithm	Centralized Systems	1.Better in terms of fault tolerance 2.Forecasting accuracy is large 3.Do not provide the overhead rejection
H. Mehta and *et al.* 2011 (ICWET)	Decentralized content aware	Distributed Computing	1. Improves the searching performance hence increasing overall performance 2. Reduces idle time of the nodes
A. M. Nakai and *et al.* 2011 (IEEE)	LB for Internet distributed services	Distributed Web Servers	1. Reduces service response times 2. Mean response time is 29% smaller than RR(Round Robin) and 31% smaller than SL(Smallest Latency)
M. Randles and *et al.* 2010 (IEEE)	Active Clustering	Large Scale Cloud Systems	1. Performs better with high resources 2.Improves throughput by utilizing the increased system resources 3. Degrades as system diversity increases
Pinal Salot 2013 (IJRET)	Resource Aware scheduling Algorithm(RASA)	Grid/Clustered Systems	1.Used to reduce makespan
Prof. A.A.Jaiswal 2012 (IJAIEM)	Design of optimized virtual server for efficient management of cloud load in multiple cloud environment	Centralized Systems	1.Both algorithms are better in terms of efficiency and dynamic response time 2.Algorithm 2 has comparatively low CPU overhead

5. CONCLUSION AND FUTURE WORK

In this paper, numerous proposed load balancing algorithms have been examined with respect to certain parameters like response time or efficiency etc. Three architectures are considered clustered, distributed and centralized for our research. In cloud computing the most important process is load balancing which leads to faster response to requests of cloud users. An efficient load balancing is thus needs to be employed to cloud. Researches have been done in this area. But still some work needs to be done to improve the response time. Some researchers have found the ways to improve it and their research proved very useful. In the next stage of our work we will try to develop a load balancing algorithm for the centralized clustered systems in cloud that will be utilized to improve the system response time. The algorithm will be developed by using certain newer approaches that can be used to get a measurable improvement in the system response time as compared to the traditional approaches.

6. REFERENCES

[1] Rajkumar Buyyaa and *et.al.* Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility ,ELSEVIER.

[2] S. K. Garg, C. S. Yeob, A. Anandasivamc, and R. Buyya, "Environment-conscious scheduling of HPC applications on distributed Cloud-oriented data centers", Journal of Parallel and Distributed Computing, Elsevier, Vol. 70, No. 6, May 2010, pages 1-18.

[3] SATW, White Paper Cloud Computing

[4] Sagar Dhakal, Majeed M. Hayat, Jorge E. Pezoa, Cundong Yang, and David A. Bader, 2007 Dynamic

[5] Load Balancing in Distributed Systems in presence of Delays- A Regeneration theory approach

[6] Roedig, U., Ackermann, R., Steinmetz, R.: Evaluating and Improving Firewalls for IP Telephony Environments. In: IP-Telephony Workshop (IPTel) (April 2000)

[7] Abbas Karimi , Faraneh Zarafshan , Adznan b. Jantan,A.R. Raml, M. Iqbal b.Saripan, A New Fuzzy Approach for Dynamic Load Balancing Algorithm 2009, IJCSI

[8] Abirami M S, Niranjana G, "Dynamic Load Balancing in Distributed Systems", International Conference on Computing and Control Engineering (ICCCE 2012), ISBN 978-1-4675-2248-9, 2012.

[9] Grace Rammurthy 3/4/2011-Distributed Systems And cloud computing A Comparative Study

[10] S. Sharma, S. Singh, and M. Sharma, "Performance Analysis of Load Balancing Algorithms," World Academy of Science, Engineering and Technology, vol. 38, 2008.

[11] G. R. Andrews, D. P. Dobkin, and P. J. Downey, "Distributed allocation with pools of servers," in Proceedings of the first ACM SIGACT-SIGOPS symposium on Principles of distributed computing. Ottawa, Canada: ACM, 1982, pp. 73-83.

[12] Nidhi Jain Kansal, Inderveer Chana, Cloud load balancing techniques-A Step Towards Green Computing,IJCSI,vol 9,issue 1,Nov 2012

[13] Ajay A. Jaiswal , Dr. S. K. Shriwastava Design of an Optimized Virtual Server for Efficient Management of Cloud Load in Multiple Cloud Environments. International Journal of Application or Innovation in Engineering & Management (IJAIEM) Volume 1, Issue 3, November 2012

[14] H. Mehta, P. Kanungo, and M. Chandwani, "Decentralized content aware load balancing algorithm for distributed computing environments", Proceedings of the International Conference Workshop on Emerging Trends in Technology (ICWET), February 2011, pages 370-375.

[15] A. M. Nakai, E. Madeira, and L. E. Buzato, "Load Balancing for Internet Distributed Services Using Limited Redirection Rates", 5th IEEE Latin-American Symposium

[16] M. Randles, D. Lamb, and A. Taleb-Bendiab, "A Comparative Study into Distributed Load Balancing Algorithms for Cloud Computing", Proceedings of 24th IEEE International Conference on Advanced Information Networking and Applications Workshops, Perth, Australia, April 2010, pages 551-556.

[17] William Leinberger, George Karypis, Vipin Kumar, "Load Balancing Across Near-Homogeneous Multi-Resource Servers", 0-7695-0556-2/00, 2000 IEEE.

[18] P. L. McEntire, J. G. O'Reilly, and R. E. Larson, Distributed Computing: Concepts and Implementations. New York: IEEE Press, 1984. Pinal Salot A Survey of Various Scheduling Algorithm In Cloud Computing Environment

[19] Fox et al: Above the Clouds: A Berkeley View of Cloud computing feb 2009

[20] NIST: Nist definition of cloud computing

[21] VMware white paper: Virtualization overview

[22] Intro_to_Virtualization.pdf-Introduction to Cloud Computing and Virtualization By Mayank Mishra Sujesha SudevalayamPhD Students CSE, IIT Bombay