International Journal of Innovative Research in Information Security (IJIRIS) ISSN: 2349-7017(O)
Issue 2, Volume 4 (April 2015) ISSN: 2349-7009(P)
www.ijiris.com

Availability in Cloud Computing

B.Vani*
Asst.Professor,
Shrimad Andavan Arts and Science College,
Trichy, India.

R.Cynthia Monica Priya
Asst.Professor
Bishop Heber College,
Trichy, India.

Abstract-Cloud computing is a realized wonder. It delights its users by providing applications, platforms and infrastructure without any initial investment. The "pay as you use" strategy comforts the users. The usage can be increased by adding infrastructure, tools or applications to the existing application. The realistic beauty of cloud computing is that there is no need for any sophisticated tool for access, web browser or even smartphone will do. Cloud computing is a windfall for small organizations having less sensitive information. But for large organizations, the risks related to security may be daunting. Necessary steps have to be taken for managing the issues like confidentiality, integrity, privacy, availability and so on. In this paper availability is taken and studied in a multi-dimensional perspective. Availability is taken a key issue and the mechanisms that enable enhancement are analyzed.

Keywords-availability, cloud, deployment, load balancing, replication

I. INTRODUCTION

Cloud computing remains a boon to customers due to its financial payback. The major barriers to adoption are the security and operational risks. Lack of availability is another area that needs attention. It is a ever concerned issue in software concentrated systems. Renowned cloud providers have experienced temporary lack of availability for hours together. Some novel mechanism must be adopted to maintain huge volumes of data over long periods. Availability should be determined using present information, forecasting usage patterns and dynamic resource scaling[1]. The various strategies like incorporating load balancing mechanism, involving efficient replication, proper deployment choices and self-healing mechanism are explored.

II. THE VARIOUS MECHANISMS

A. Load Balancing

This mechanism is very important in cloud computing because the arrival of tasks at the cloud is very erratic and the cloud provides instant scaling up or down of resources to its clients [2]. It is a technique that favors the networks and resources availability by providing a maximum throughput with minimum response time [3].Application of load balancing greatly reduces the number of failures that could instantaneously affect the cloud system. If there is a failure in one part of the system the load balancer is able to switch to the other available resources [4].Dividing the traffic between servers enables data to be sent and received without much delay. Load balancers increase availability and performance as well [2].

Load balancing may be done both statically and dynamically. Static algorithms employ equal distribution of traffic among servers. Due to practical problems, weighted round robin was employed. Based on this approach, the servers with the highest weight received more connections. However, in cases where weights are equal, the servers will then receive balanced traffic. Dynamic load balancing algorithms are used to redistribute available resources among running tasks dynamically. This enables the tasks to use the maximum capacity of each resource in a node. Multi computers with dynamic load balancing feature can allocate and reallocate resources allocation at runtime. This may result in significant improvement in the performance .Load balancing should happen when the scheduler schedules the task to all processors. The following procedure is found to be effective. As new jobs arrive, they are queued to a particular node. The Scheduler may schedule the job to a processor. Rescheduling is to be done if load is not balanced. Allocation and release should be done to the processor. Cloud employs automatic load balancing services that involves in increasing the number of CPU'S or memories to scale with increasing demands [5].

B. Efficient Replication

Cloud computing provides assurance to increase the pace with which applications are deployed, increase novelty and reduce costs without compromising business dexterity. Various companies like Amazon, Google, Microsoft have built huge data centers over the recent years. The data centers are able to provide services at condensed rates which has motivated many institutions to host their services in the cloud[6]. The virtualization concept facilitates one physical node to be projected as several virtual nodes favoring inexhaustible resources. Data replication has been used as a method of increasing data availability [7]. The architecture refers three layers namely application client, storage client and the PC cluster. The application layer offers interfaces for clients to store files, databases and so on. The PC cluster layer provides the hardware and large scale storage devices.The reliability of a system will increase as the number of replicas increase.This would serve to mask more failures. Many replication management schemes are available,out of which one is explored [6].

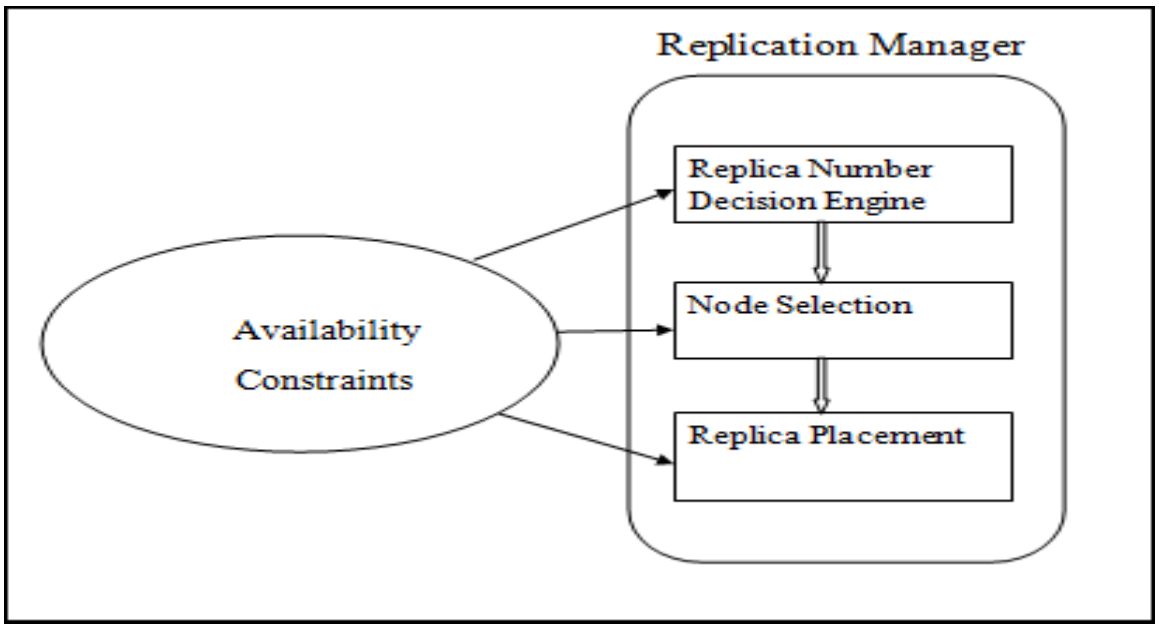

Fig. 1 Three processes of the replication management scheme

The replication management consists of three processes. In order to increase availability, the availability constraints both file availability and storage server availability are included in all the three processes [8]. The replica number decision engine plays a significant role in deciding the number of replicas. Availability needs increase in number of replicas. But once a certain limit has been reached, further increase may not be needed. The adverse effect on further increase may be the management cost. Node selection is the next process. Each node has a calculated weight .Binary weighted tree is used to search the nodes that are highly available in the PC cluster. Replica Placement is the method involving the placement of replicas among data nodes. An efficient data placement algorithm is used for this purpose. To satisfy availability in dynamic environment related to node failure and access frequency, dynamic replication can be done based on access frequency. This method will increase the elasticity of the cloud system.

C. Deployment Choices

Virtualization is the creation of a virtual edition of something, such as a hardware platform, operating system, a storage device or network resources. Virtualization technologies reduces energy and hardware costs .Further, it enhances resource sharing among the applications hosted on different virtual machines [9].In a nutshell, they are used to increase the manageability of software systems and decrease the total ownership cost. This concept allows resources to be owed to different applications but hides the complexity of resource sharing. However, sharing VMs among the application with different resource prerequisite involves uncertainty in availability and performance.

The method of deploying application components into virtual machines and the placement of these on physical machines helps to improve availability [10].The selection of the best deployment strategy for the available software components, the number of replicas of each component ,the components that should be placed on the same machine should be analyzed. Deployment choices are significant in determining the availability of cloud applications. However, sharing VMs among applications that need different resources introduces doubts about performance and availability.

The term “executor” refers to either a component or a service that runs on a VM. Many approaches may be followed. The first approach involves consolidating all the task executors into one node. The second approach involves one node per task and the third involves task executor groupings. The placement problem deals with the selection of executors for their corresponding VMs. The task executors are hosted on virtual machines, having a collection of resources[9].

D. Self-Healing and Consistency Mechanism

The middleware architecture in multi-master pattern has four components namely User/ Client, Cloud Manager (CM), Cluster Controller (CC) - Node Controller (NC) and Backup Manager (BM).

User/ Client: The cloud user needs to interact with the cloud manager before interacting with the interface.

CM: They are the master nodes. They are interconnected in a full-mesh topology.

They serve as an interface for the user to request a list of the available hardware resources as well as to set the connection to one of the hardware components.

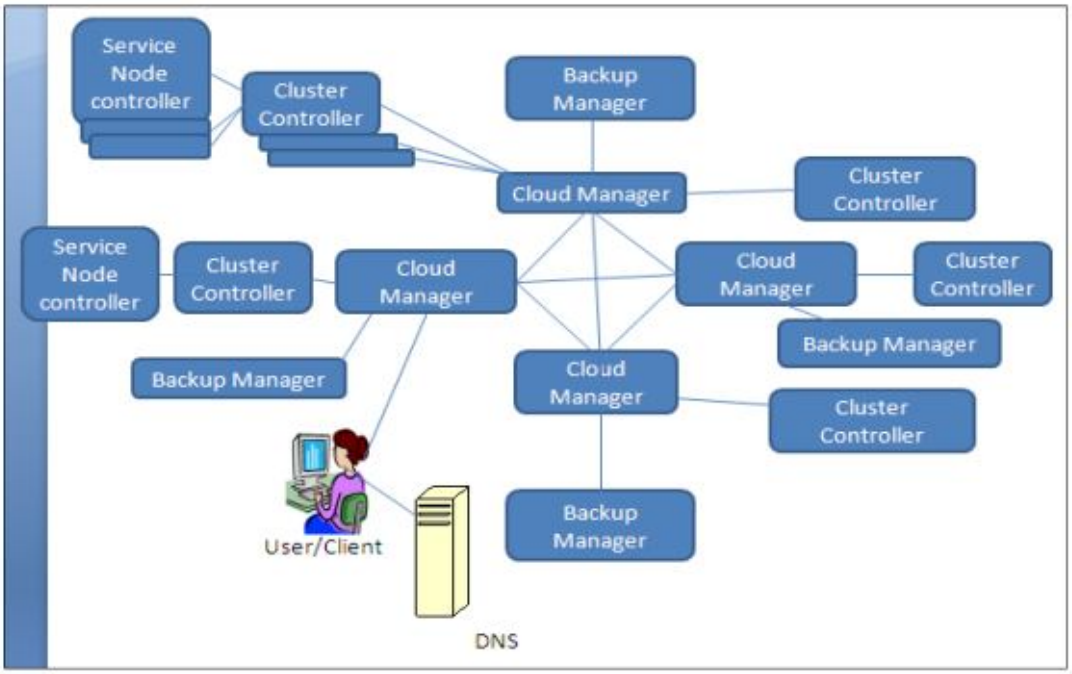

Fig.2. Middleware architecture in Multi-master pattern

CC-NC: They are the individual worker nodes which are connected to exactly one of the master nodes. They announce and update a list of all known CMs in the system.

BM : When some of CMs fail and the Cluster Controllers lose the connection to the cloud interface an automatic self-healing is provided.

The CM forwards the requests for the available hardware components to all the connected Cluster Controller. The CC propagates the request to the Node Controllers(NC) which executes the request and returns its response to the CM, which is passed back to the user. The middleware architecture in multi-master pattern prevents an overall cloud failure in case of a failed master node [11].

III. **Exploring the various mechanisms**

A. Mechanism 1

Static algorithms used in load balancing employ equal distribution of traffic among servers. Due to practical problems, weighted round robin was employed. Based on this approach, the servers with the highest weight received more connections. However, in cases where weights are equal, the servers will then receive balanced traffic [4].Dynamic load balancing algorithms are used to redistribute available resources among running tasks dynamically. This enables the tasks to use the maximum capacity of each resource in a node. Multi computers with dynamic load balancing feature can allocate and reallocate resources allocation at runtime. This may result in significant improvement in the performance. Load balancing should happen when the scheduler schedules the task to all processors. The following procedure is found to be effective. As new jobs arrive, they are queued to a particular node. The Scheduler may schedule the job to a processor. Rescheduling is to be done if load is not balanced. Allocation and release should be done to the processor. Dynamic algorithms designated proper weights on servers and by searching in whole network a lightest server preferred to balance the traffic. However, selecting an appropriate server needed real time communication with the networks, which will lead to extra traffic added on system.

Having a closer view on the three kinds of load balancing helps us to understand the pros and cons of each. The static method uses algorithms that use the principle of equal distribution of traffic among servers. The next approach was to impose weighted round robin. Servers are assigned weights. This method is acceptable for unequal weights, whereas remains an overhead for equal weights imposed on servers. Dynamic load balancing is the second method considered. This method employs algorithms that designate proper weights to servers. It involves searching the whole network for the lightest server that would balance the traffic. Even though, this method sounds good, it suffers from a serious problem.

Selecting the appropriate server needs real time communication with the network, which in turn increases the network traffic [12]. The third method is to employ automatic load balancing. It enables entities to increase the resources according to the raise in demands. This method seems to be the best among the three. The cost involved may remain a constraint to its effectiveness. Any of the three methods are employed according to the situation. The pros overshadow the cons, as load balancing meets the needs of availability and performance.

B. Mechanism 2

Data availability is one of the key requirements for the cloud system. Data replication has been used as means of increasing the availability in traditional distributed databases, peer-to-peer (P2P) systems, and grid systems [13]. Cloud systems differ from the previous frameworks in that they are intended to support large numbers of customer-oriented applications, each with different quality of service (QoS), requirements and resource utilization. In the explored replication management system an optimum replica number is provided. The explored system is found to improve data availability depending on the expected availability and failure probability of each node in the cluster [7].

C. Mechanism 3

The ways and means of improving availability of deployed software applications in cloud environments have been explored. The placement of executors on VMs improves availability. The factors that affect the placement are as follows:

• Failure of Infrastructure Resources

The errors on the individual nodes and failure of cloud infrastructure may be the two reasons that cause the failure of infrastructure resources. Resource overloading, crashing of the operating system, hardware errors, problems in networks, operation errors may be some of the examples for the causes for failure.

• Task Executor's Failure : The availability of the task executor is found to be a function of the mean time to recovery (MTTR) and the mean time to failure (MTTF) [9].

• The Multi-tenancy of the Task Executors: Multiple task executors may share resources on the same node; failure of one executor may have impact on the availability of other executors on the node. Let d1 and d2 be two task executors placed on the same VM. If suppose for some request classes d1 alone is executed. Failure of t2 may impact the availability of d1 for the requests that use d1 and not d2. For example, if d2 fails in a busy state, sufficient CPU cycles may not be provided to d1 to enable it to respond to the requests.

The approach that is involved is referred to as ACR(Availability, Correlation Factor, Resource Consumption) model. The placement decision maker takes into account the request availability, cost of resources and interference between the task executors on the same node.The results produced by this model are as follows:

• Availability-aware placements are very effective in enhancing the end-to-end availability.

• Increasing the application availability does not always leads to extra costs on resource consumption.

• Error interference affects the availability of task executors running in the same node. Considering effective interference correlations helps in choosing good placement decisions that may improve the availability of application.

D. Mechanism 4

A communication model for middleware architecture in multi-master pattern has been put forth. It has been explored that it reduces failure and enhances availability. The self healing mechanism helps in the automatic reconnection of worker nodes to another master of the cloud. Furthermore the reconnection also guarantees the network consistency [11].

IV. **Conclusion**

Cloud Computing is the talk of the hour. It provides services to its clients, "on-demand". The core principle of cloud computing is its ability to serve its clients on demand. Any technology that needs celebration also needs attention to the issues. One such is availability. This paper has explored means of enhancing availability from different perspectives. The available mechanisms may help at certain situations. As more and more situations rise, more will the need of enhancing availability. This paves the way future research.

References

[1] M. Armbrust, Armando Fox, Rean Griffith, Anthony D. Joseph, Randy Katz, Andy Konwinski, Gunho Lee, David Patterson, Ariel Rabkin, Ion Stoica, Matei Zaharia , "A view of cloud computing", Communications of the ACM, vol. 53 , 2010 , pp. 50-58, http://cacm.acm.org/magazines/2010/4/81493-a-view-of-cloud-computing/fulltext.

[2] Suneel K S, Ravichandra A J , H S Guruprasad, "Enhanced Load Balancing Algorithm in Three-Tier Cloud Computing" , 2014, pp. 296-301.

[3] D. Durkee, Why Cloud Computing Will Never Be Free , Volume 8 ,Issue 4, April 2010 , pp. 20-29.

[4] Borko Furht , Armando Escalante Handbook of Cloud Computing, ISBN 978-1-4419-6523-3, Springer, 2010.

[5] Ruixia Tong ,Xiongfeng Zhu, A Load Balancing Strategy Based on the Combination of Static and Dynamic, Database Technology and Applications (DBTA)2010, pp. 1-4.

[6]Julia Myint , Thinn Thu Naing Improving Data Availability in Cloud Storage with EfficientReplication,ICCA2012,http://www.academia.edu/5171739/Improving_Data_Availability_in_Cloud_Storage_with_Efficient_Replication.
[7] Qingsong Wei , Veeravalli B., Bozhao Gong ,Lingfang Zeng , Dan Feng CDRM: A Cost-effective Dynamic Replication Management Scheme for Cloud Storage Cluster, IEEE International Conference on Cluster Computing, 2010.
[8] Suruchee V.Nandgaonkar, A. B. Raut, "A Comprehensive Study on Cloud Computing"; , IJCSMC, Vol. 3, Issue. 4, April 2014, pp.733 – 738.
[9] Jim (Zhanwen) Li,Qinghua Lu, Liming Zhu,Len Bass, Xiwei Xu,Sherif Sakr ,Paul L. Bannerman,Anna Liu, Improving Availability of Cloud-Based Applications through Deployment Choices, IEEE Sixth International Conference on Cloud Computing (CLOUD) , 2013,pp. 43-50.
[10] Jeffrey Dean, Luiz André Barroso, "The Tail at Scale", Communications of the ACM, Vol. 56 No. 2, Feb. 2013, pp.74-80.
[11] Megha Balnath Rode et al., High Availability of Cloud through Different Self-Healing and Consistency Mechanism , Volume 2, Issue 11, November 2014, pp. 200-205.
[12] A Khiyaita, M Zbakh, H El Bakkali, D El Kettani, "Load balancing cloud computing: state of art," 2nd National Days of Network Security and Systems [JNS2], 20-21 April 2012, pp 106-109.
[13] Gueyoung Jung, Kaustubh R. Joshi, Matti A. Hiltunen, Richard D. Schlichting, Calton Pu, "Performance and availability aware regeneration for cloud based multitier applications", Dependable Systems and Networks , 2010, pp. 497-506.

