International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 5 | ISSN : 2456-3307

Review on Service Availability: A Geo-Redundancy Feature of Cloud Based Applications

Riaz Ahmad, Dr. Tamanna Siddiqui
Department of Computer Science, Aligarh Muslim University, Aligarh, Uttar Pradesh, India

ABSTRACT

Cloud computing is a fairly new concept which provides computing services with a variety of advantages (more suitable utilization, reasonably priced and scalability). Cloud computing offers the desired facility/service of an aggregated pool of resources as a utility to user over the internet. If anyone would like to start, taking advantage of cloud services there is a considerable shift of these kinds of resources into the cloud. Essentially the most important areas of cloud computing for users is overall performance and availability. However, critical application have most important concerns about the availability of their services in the Cloud. Conventionally the availability of these has been limited to local installations of hardware and software resources. On the other hand, in cloud computing availability indicate the uptime of a service or system, a network, server and equipment (hardware and software) that are used to provide service. Completely satisfied cloud user' requirements that clarify the quality of service. In this paper, we focus on Geo-redundancy and availability of the service in cloud environment. As well as covering the different concepts of Geo redundancy, availability, issues and the highly effective benefits for the users.

Keywords: Cloud Computing, Testing framework, High Availability, Geo-Redundancy, Cloud Based Application

I. INTRODUCTION

The Cloud has been a pertinent runner of top technology trends in recent times. The Cloud can be perceived as a conceptual layer on the Internet because of that it will make all available hardware options transparent, making them available via a well-defined interface. Essential and common characteristics of the Cloud are the primary factor that has resulted in its popularity and attracts users by reducing initial investments and resource managing costs and alienating the end user with nuances of hardware management. Managing all these service and infrastructure remains a great challenge, considering clients' requirements for zero outage [1] [2].According to Toeroe [6] availability is calculated in terms of percentage, it is used to measure total time duration a service is available for user during promising time period. Trying to achieve high availability (HA) of the service, and provide at least 99.999%, availability ("five nines") services which is an industry benchmark for critical applications.

High availability systems are just like fault tolerant systems without having any single point of failure. In other words, if any system component fails, it does not essentially cause the termination of the service provided by that component.

Cloud provider face lots of challenges, one of the biggest challenges is to deliver a higher level of

availability services. The main aim of this paper is to present a systematic review on service availability and discuss the HA solutions for the Cloud. The authors expect that such solutions could be utilized as a good foundation to addressing most of the current issues in the HA Cloud [4] [11] [13].

This paper is divide into six section. Second and third section, discuss about the geo redundancy and redundancy strategies respectively. In fourth and fifth part of the paper, discuss the availability and the issues related to the cloud service availability respectively. At the end of the paper, give the conclusion.

II. METHODS AND MATERIAL

1. Geo-Redundancy

Geo-Redundancy creates the way to setup computer resource at two or more than two places as a redundant site in case any type of the failure occurs in the primary system. It replicates your data and stores it as a duplicate data in an alternative physical location just in case active site fail or degraded. It contributes greatly to build and maintain the data backup as well as help to improve the application availability. In geographic redundancy setup, clearly state the Geo-master site along with the Geo-redundant service configuration.

To achieve the high availability of the system, duplication play an important component in the redundancy system. Active-Active, Active –Standby and N+K load sharing, are the redundancy model use to maintain the availability [10] [13].

1.1. **Internal Redundancy**- This type of redundancy takes into account both the redundancy (such as hardware and software) which is interior and maintained by one particular system instance. Because of that, it should really minimize a large amount of single hardware and software failure situations.

1.2. **External Redundancy**- This type of redundancy takes into account a set of non- redundant systems or internally redundant systems, which can be organized like a pool to provide more accessibility and/or power to clients.

It can be further divide into two different types:

- **Co-Located External Redundancy -** It is actually creating either at the exact same data center or near with another. For example, one pool of rack-installed servers is mounted in the same equipment rack.
- **Geographically Distributed External Redundancy, or "Geo-redundancy**," - These kinds of options are actually done separately and distribute systems geographically apart to reduce the hazard that an accident, such as an earthquake would probably destroy the option to make service not available for a generally very long time-period. A few examples of distributed pool of web servers promoting the exact same website.

- Simple configuration
- Redundancy
- Single point of failure

Table 1: Type of System Redundancy

Single System configuration		*Multiple system configuration*	
Redundancy Free	*Internally Redundancy agreement*	*Co-located external redundancy*	*Geographically distributed external redundancy*

…	ACTIVE – ACTIVE ACTIVE – STANDBY N-K load share and complex redundancy arrangements	A number of homogeneous systems at the similar actual site. For example data center	A number of homogeneous systems at multiple than two geographically separated or divorced sites. For example, data centers
…	…	For external redundancy, equivalent redundancy choices are available in the same way with internally redundancy preparations	
Simple Redundancy	Internal Redundancy	External Redundancy	

2. Redundancy Strategies

The following external redundancy schemes are commonly used:

2.1. Active – Standby:

In this model, two units are required for deployment, one is installed as the "active" unit, which offers services for all of the traffic, and the other one installed as the "standby" unit. In case active module fails to provide the services, then the standby unit takes over service for users. The standby unit must be configured with sufficient capacity to support the fully engineered traffic load from the active unit. Standby units have changeable degrees of readiness to restore service; these degrees of readiness are:

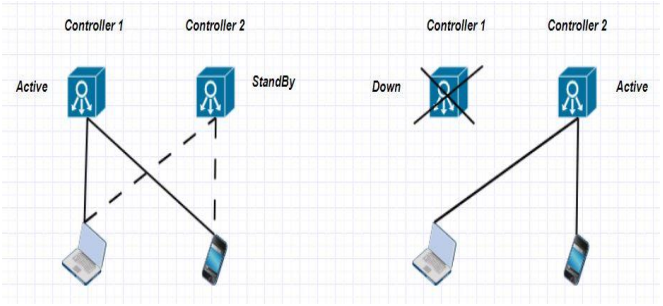

Figure-1: Active/Standby

Hot Standby (Hot Spare) -In this redundant method one system runs at the same time with an identical primary system. At the time of failure of the primary system, the hot standby system activate and immediately takes over, replacing the primary system. Thus, both systems have identical data. A hot standby network element can be either automatically triggered into service or manually activated. Periodically sent a sort of heartbeat message between the active and standby units (either the active sends a heartbeat to the standby or the standby sends a heartbeat to the active and expects to receive an acknowledgment). If the standby unit does not receive any acknowledgment from the active that means it stop processing heartbeats, then it makes itself active and traffic is redirected to it.

- **Warm Standby.** In this redundancy method, secondary system runs in the background of the primary system. At regular intervals, secondary server copies data from the primary server, which means that there are times when both servers do not contain the exact same data. It is used for replication and mirroring.
- **Cold Standby**. In this redundancy method having one system as a backup for another identical primary system. It is called at the time when, only on failure of the primary system. Cold standby unit can restore service from a failed active unit, the standby must be powered on, the operating system and platform software must boot up, persistent data must be loaded from a backup copy and the application must be completely started up.

2.2. Active–Active.

In this model, two units are required for deployment. Both unit are installed as the "active" unit, each capable of serving the entire engineered service traffic and both the units are fully operational and carrying traffic during normal operation. In general, half of the traffic is catered by each unit. When any of the active units fails, traffic served by the failed unit, can be carried by the other unit. Active–Active can provide a few added advantages over active–standby planning.

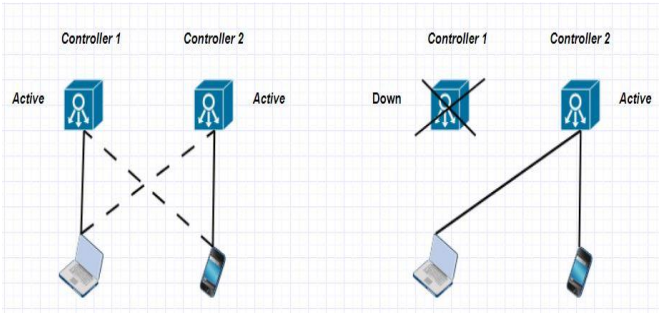

Figure-2: Active/Active

- ***Better service during normal operations*** both active unit should be running 50% of fully engineered capacity not more than that. Thus, service latency may be lesser than on standby systems on which the active unit routinely serves more traffic.
- ***Lower risk of uncovered or silent failure***-easy to find the failures of units as compare to standby units (essentially idle) because it is actively delivering service.

2.3. N + K Load Sharing-

In this model, "N" identical units are required to manage the engineered load, as well as "*K*" extra units can be deployed because of that system can easily manage and withstand up to "*K*" simultaneous element failures without any loss of capacity. It is most helpful to share the engineered load of a system across a pool of an identical unit. Service load is equally distribute across all the working elements in the pool of N + K identical units. This model offer several benefits. Generally, all network elements are active, in case any failure of a network element found either by a client or by a load-sharing device, messages can be linked to any one of the remaining network elements.

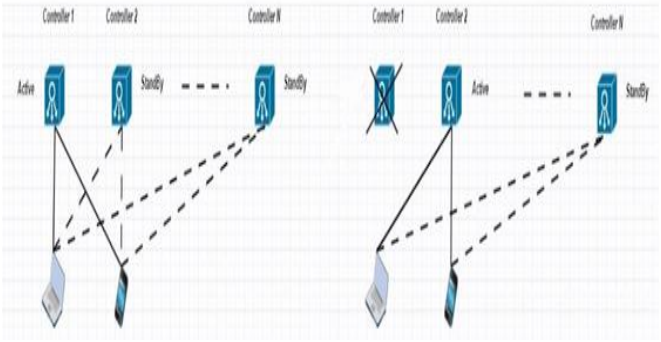

Figure-3: N+K Load Sharing

2.4. Comparison of Redundancy strategies

Table 2: Comparison of Redundancy strategies

Redundancy Type	Pro	Con
Active- Active	• The failure domain and outage duration are very smaller, because fewer APs must fail that take less time to recover. • Every controller is working at all times that has a reduced load.	• Expensive because of licensing requirements. • Load distribution is less than the hardware capacity, so generally hardware goes un-utilized.

Active-Standby	• If APs fail to the backup controller, as data synchronization is performed at regular intervals.	• Similar to Active-Active from cost perspective. • Outage duration is comparatively larger.
N+K Load Sharing	• In N+K load sharing strategies, required controller with licensed, and scaled in such a way to handle the maximum number of failed controllers. • In general, deploy only one redundant controller.	• Preemption enablement is a requirement, which could result into an outage, albeit a planned one.

2.5. Availability

The objective is to make sure that application is always available, even if a cluster member is down. It also describes the abilities of a user to easy access important information or resources in a specified location as well as in the appropriate format [10].

According to the IBM Community

Availability = High Availability + Continuous Operations

High Availability (HA) - (Minimized Unplanned Downtime) all the attribute of a system to deliver service during specific time-period, at most appropriate or approved levels and masks unplanned outages from end-users.

Unplanned = Automated failures detection, network outages, Recovery, Testing, data centre failures, etc.

Continuous Operations (CO) - (Minimize Planned Downtime) all those attribute of a system to provide non-disruptive service to the end user, 24/7 service.

Planned = Patching, application upgrades, code releases, data centre maintenance etc.

III. DISCUSSION AND ANALYSIS

3. Service Availability

If the failure event continues for more than a few seconds, it is likely to impact isolated user service requests and the user initiated retries of those failed requests. Brief service impact events may cause individual transactions or sessions to fail, thus prompting the user to retry the transaction or session, if the first retried attempt fails because service is still impacted, then the event will often be considered a service outage, and thus it impacts the service availability metrics. When service is impacted so long that retried user operations fail thereby causing users to abandon their efforts to access the service, the service is generally deemed unavailable [3][5]. MTBF (Mean Time between Failures) and MTTR (Mean time to repair) are helpful parameters to calculate Availability of a system or service.

Calculating availability by using MTBF and MTTR [1].

Availability = (MTBF/ (MTBF + MTTR))

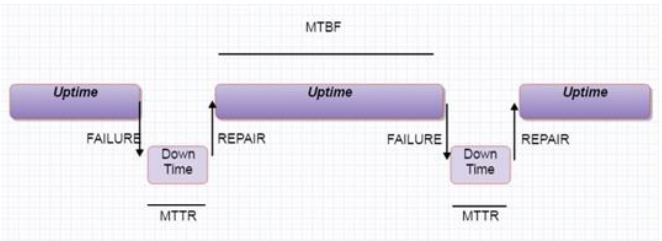

Figure-4: Estimate the availability from MTBF and MTTR

Figure give the better understanding of the given equation. MTTR is the time duration, to return a system to service (downtime) and MTBF is the time duration in which the system should come up (Uptime).

CRITICAL	(99.999% service availability)	Loss of this capability would raise to an unacceptable level, the risk associated with providing safe and efficient operations.
ESSENTIAL	(99.9% service availability)	Loss of this capability would significantly raise, the risk associated with providing safe and efficient operations.
ROUTINE	(99% service availability)	Loss of this capability would have a minor impact on the risk, associated with providing safe and efficient operations.

According to the U.S. Federal Aviation Administration, the criticality definitions in cloud environment, cloud providers offer different availability guarantees. Clod providers may offer you 99.999%, 99.99%, 99.95% or 99.9% uptime. In case there is a failure they may specify a time to resolve for that failure, which could be form half hours to a few hours, many provider does not mention any type TTR (Time to Resolve).

Service availability is necessary because the provider gives you credits only when the uptime is less than the availability in the SLA. If the service availability is 99.54% they are reliable to get 40 hours of non-schedules downtime per years that amounts to around 200 minutes every month. In that condition credit are allow only when outage is greater than the 200 minutes for every billing month.

3.1. Penalty for Non-Availability Service

On the behalf outages, (Outages are dependent upon the amount of service downtime) it permits you to estimate the credits. Credits are commonly a percentage or a part of the bill amounts. Apart from the calculation process the total credits allowed should also be consider. Different provider have their own credits rule, some providers offer at most 20% to 30% of the bill amount as credit and some other may offer up to 100% of the bill amount[4][12]. However, no one offers more than the 100%. For example, a cloud provider's SLA state that 99.9% availability. In any case, if the uptime is 96.4% then offer up to 85% refund for the service charge.

Table 3: Example credit for an SLA uptime of 99.9%

Uptime % per Month	Reduction (credit) in Monthly Service Fee
99.89% - 99.5%	10%
99.49% - 99.0%	25%
98.99% - 98.0%	40%
97.99% - 97.5%	55%
97.49% - 97.0%	70%
96.99% - 96.5%	85%
96.5% or less	100%

If the downtime for the month is 200 minutes (that is .4630%) the uptime is 99.54% and hence the credit back to you is 10% of the monthly fee. Different provider offer different Service availability (Uptime %) on the behalf provide credits.

Table 4: Service credit offer by the different Service provider

Provider	*Uptime% per Month*	*Service Credit Percentage*
Amazon S3	*99.9%-99.0%*	*10%*
	99.0% or less	*25%*

HP Cloud	*Uptime% per Month*	*Service Credit Percentage*	
	100%-99.95%	*No*	
	99.95%-99.9%	*5%*	
	99.9% - 99.5%	*10%*	
	99.5%-99%	*20%*	
	99.0 or less	*30%*	
Google Apps	*Uptime% per Month*	*Days of Service added to end of the service term*	
	99.9%-99%	*3*	
	99%-95.0%	*7*	
	99.5% or less	*15*	
The Rackspace Cloud	*Uptime% per Month*	*Service Credit Percentage*	
	Network	*30min (100% - 99.93%)*	*5%*
	Data Center	*30min (100% - 99.93%)*	*5%*
	Cloud Server Hosts	*30min (100% - 99.93%)*	*5%*
	Migration	*30min (100% - 99.93%)*	*5%*

3.2. Issues on Cloud Service Availability

In cloud service, service availability is the one of the features that is impacted by a weak hypervisor [8] [9].

- In cloud environment, cloud services run on virtual machines and a hypervisor that allows more than one operating system to share a single hardware host. Weak hypervisor always results in negative impact of service availability for cloud applications.
- No standard technique or framework available for testing non-functional requirement.
- Determine the single point of failure, which can be present at application, data center, infrastructure or geographic level and provide security for confidentiality and integrity of data.

IV. CONCLUSION

In these days cloud computing offer services to the user in safe and secure manner. Many different service providers offer different services and make a SLA (Service Level Agreement) between the users and providers. SLA contracts clearly mention service description and condition. In this paper, we have discussed the Geo-Redundancy and availability of cloud-based services. We also compared the cloud service vendors in terms of availability and the level of service reliability offered by them. After analysing, we observed that a standard testing technique or framework for testing Geo-Redundancy feature is required.

V. REFERENCES

1. Eric Bauer and Randee Adams, “Reliability and Availability of Cloud Computing”,-Wiley-IEEE Press (2012).
2. Tamanna Siddiqui, Riaz Ahmad; Cloud Testing: A Systematic Review; International Research Journal of Engineering and Technology (IRJET); Volume: 02 Issue: 03 June-2015; e-ISSN: 2395 - 0056 p-ISSN: 2395-0072.
3. T. Parveen, and S. Tilley, “When to Migrate Software Testing to the Cloud?,” In proc. 2nd International Workshop on Software Testing in the cloud (STITC), 3rd IEEE International Conference on Software Testing, Verification and Validation (ICST), April 2010, pp. 424-427.
4. Jayaswal, Kailash. Cloud Computing Black Book. John Wiley & Sons, 2014.
5. Tamanna Siddiqui, Riaz Ahmad, "A review on software testing approaches for cloud applications", Perspectives in Science (2016) (Elsevier) 8, 689—691.
6. Toeroe M, Tam F (2012) Service Availability: Principles and Practice. John Wiley & Sons.

http://www.wiley.com/WileyCDA/WileyTitle/productCd-1119954088.html.

7. S. Suhel, Mohd. Salim, “Cloud Based Application Testing”, IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 3 Issue 7, July 2016.
8. Najeeb Ahmad Khan, Tamanna Siddiqui and Riaz Ahmad, "REVIEW ON CLOUD SECURITY ISSUES: CHALLENGES AND SOLUTION", International Journal of Recent Scientific Research Vol. 7, Issue, 5, pp. 10846-10853, May, 2016 ISSN: 0976-3031.
9. Ahuja, Sanjay P., and Sindhu Mani. "Availability of services in the era of cloud computing." Network and Communication Technologies 1.1 (2012): 2.
10. Eric Bauer, Randee Adams, Daniel Eustace-Beyond Redundancy_ How Geographic Redundancy Can Improve Service Availability and Reliability of Computer-Based Systems-Wiley-IEEE Press (2011).
11. Mrs. Priyanka Hariom Singh, "Introduction Minimizing Planned Downtime of SAP Systems with the Virtualization echnologies & systematic review and research challenges in cloud computing", International Journal of Advance Research in Science and Engineering (IJARSE),Volume No.06, Special Issue No.01, December 2017.
12. Dr. Tamanna Siddiqui and Riaz Ahmad, "Testing Framework for Geo-Redundant Cloud Based Applications" Journal of Advanced Research in Dynamical and Control Systems (JARDCS), Vol. 10, 04-Special Issue, 2018.
13. T. Narula and E. G. Sharma, "Framework for Analyzing and Testing Cloud based Applications," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 4, no. 6, pp. 592-596, 2014.
14. S. Hosseini, R. Nasiri, G. L. Shabgahi, “Framework for Testing Cloud Base Applications”, International Journal of Scientific Engineering and Applied Science (IJSEAS) - Volume-1, Issue-3, May 2015.