Availability of Services in the Era of Cloud Computing

Sanjay P. Ahuja[1] & Sindhu Mani[1]

[1] School of Computing, University of North Florida, Jacksonville, America

Correspondence: Sanjay P. Ahuja, School of Computing, University of North Florida, Jacksonville, FL 32224, America. E-mail: sahuja@unf.edu

Received: February 24, 2012 Accepted: March 15, 2012 Online Published: June 1, 2012

doi:10.5539/nct.v1n1p2 URL: http://dx.doi.org/10.5539/nct.v1n1p2

Abstract

One of the most important areas for consumers is security, performance and availability when it comes to cloud computing. Availability refers to the uptime of a system, a network of systems, hardware and software that collectively provide a service during its usage. Traditionally the availability of these has been limited to local installations of hardware and software resources which businesses and consumers deployed and maintained. With the advent of cloud services there is a considerable shift of these resources into the cloud. While cloud computing presents some cost effective benefits for the consumers and businesses, it is also extremely important for the cloud service providers to offer environments that are highly scalable and high in availability. This will in many ways dictate the credibility of these cloud services. Regardless of the size of an organization prolonged downtime of the service might be disastrous to its business, customer loyalty and brand value. This paper discusses the state of availability of services in the cloud.

Keywords: availability, scalability, EC2

1. Introduction

Cloud service providers in recent years claim to provide both the cost and efficiency benefits to the businesses. They are offering cost saving models to the customers that can take advantage of the infrastructure they provide and minimum setup is required for these models to get started. However, for several businesses some of the areas in the cloud are still of concern such as security, performance and availability. The cloud service providers are expected to provide mission critical services so the businesses are run in a highly efficient, scalable, safe and are consistent yet flexible environment. Companies like Amazon, Microsoft, Google, salesforce.com have been spending hundreds of thousands of dollars in the research and development of cloud services. It will be very important to ensure these organizations retain their brand loyalty and customer base.

IDC, a premier global market intelligence firm in a recent report predicted that spending on cloud and related services would grow by six times by the year 2013. It also predicted that 25 percent of total IT spending would be towards the cloud services in the next 2 to 5 years. If this is going to be true then it will be a major shift for the companies, consumers from how businesses will be performed and architected. The ownership of the hardware will be in a major transition and the corporate sector will see a reduction of hardware assets (Silva, 2010).

The services providers are expected to offer much more robust cloud infrastructures to ensure that their consumers are presented with an environment that is highly scalable and has high availability and service delivery capabilities. This is true for the public, private and hybrid models of the cloud.

1.1 Services in the Cloud

The cloud services providers have come up with different service types that benefit the businesses and consumers according to their needs. These cloud services are categorized into three categories namely Software as a service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS) (Sun, 2009). Organizations provide SaaS on demand. Essentially SaaS is based on the concept of renting the cloud-based application instead of buying it. Couple of good examples that fall in this category are Google Apps that manage pictures, email service, calendar etc. Second example is salesforce.com. Salesforce.com provides software solutions for sales and marketing on the cloud. It implies that applications in this category can be used effectively and get benefited by both the individual consumers and corporate organizations.

The next category of the cloud services is Platform as a Service (PaaS). In this category, the cloud service

providers provide a platform for developing and executing the applications. Services provided include database management, security, workflow management, application serving, and so on. In addition PaaS also provides tools to maintain these applications. Google App Engines is an example of PaaS that provides an infrastructure and environment application developers.

The third category is Infrastructure as a Service (IaaS). While the cloud service provider provides the required infrastructure, the software that runs on it is essentially the consumer's. Infrastructure includes data centers, servers, networks and so on.

2. Achieving High Availability

Could computing is, by all means, a third party service and consumers heavily rely on the service providers for their computing needs. These computing needs range from research to businesses to high performance computing. Researchers are heavily involved in finding new technologies that can make cloud computing more reliable from a security, performance and availability's perspective. Traditionally the resources required for businesses have been locally installed, setup and maintained by the organizations. The organizations interact with each other in a very controlled and secured environment. They often sign the service level agreements (SLAs) that hold each party engaged with certain accountabilities. These accountabilities must be very concrete and measurable (Nimsoft, 2009) when they interact with each other in the cloud. A lot of groundwork needs to be done to ensure reliability. In some situations a downtime of few hours can lead to a loss of hundreds of thousands of dollars. Establishing robust monitoring tools and practices will bring long terms benefits in terms of achieving high availability in the cloud.

Technically there are several levels where high availability can be achieved. These levels include application level, data center level, infrastructure level and geographic location level (Rackspace, 2010). One of the very basic goals of high availability is to avoid single point of failures as much as possible to achieve operational continuity, redundancy and fail-over capability (Rackspace, 2010). At the infrastructure level the basic configuration might look like the one in the Figure 1. This configuration has two or more load balancers, two or more web servers and two or more database servers. The consumer accesses the cloud via the Internet. At each level both active and passive nodes are provisioned to provide high availability. If one node goes bad, the second node supports the load and hence reducing the downtime. This is replicated at each level of the configuration as shown in Figure 1.

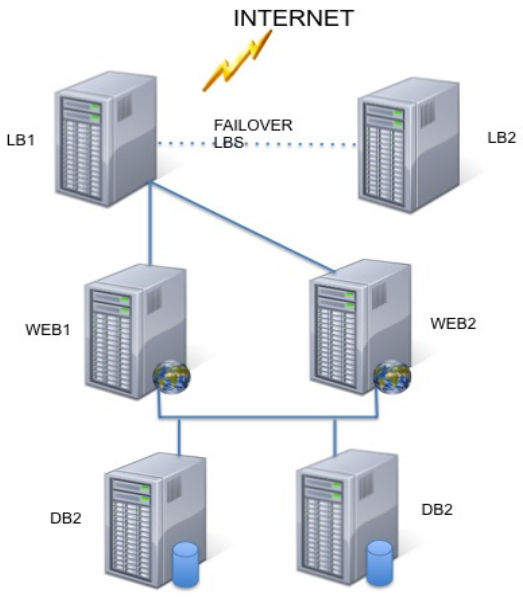

Figure 1. Infrastructure level high availability configuration. Courtesy (Rackspace, 2011)

Dynamic scalability of the services is one of the very important features of the cloud. This goes a long way in achieving high availability. Amazon's EC2 scales up the services by provisioning additional servers very easily and in a short amount of time. It provides dynamic scalability capabilities which help in load balancing and effective handling of the sudden and unexpected increase in the network traffic. This dynamic scalability can be

programmatically controlled via cloud servers API. Programmatically controlled environments provide near real time scalability capabilities. With a single API call several virtual machine instances can be added to a cluster (Kupferman, Silverman, Jara, & Browne, 2009). And since the resources are fixed at the beginning of the computation the applications can be scaled up or scaled down as the requirement to adjust the workload arises. These adjustments can be in the form of requesting more machines or terminating the ones that are not needed. Amazon's EC2 provides capabilities to control and manage the resources per user needs. This in turn helps the web services to provide high availability.

Figure 2 illustrates a typical Windows Azure and roles of some of the core components that are responsible for scalability and availability. In this environment the physical hardware resources are abstracted away and are exposed as compute resources for the cloud applications to use them. A Windows Azure Fabric is controlled by a Fabric Controller. This windows fabric is responsible for exposing the storage and computing resources by abstracting the hardware resources. In addition the instances of the applications are monitored for availability and scalability. This is done automatically in the environment. If one instance of the application goes down for some reason the Fabric Controller is notified and the application is instantiated in another virtual machine. This process ensures that the application availability is achieved with minimal impacts to the downtime in a consistent manner. At this time Windows Azure environment heavily supports. NET applications, ASP.NET applications, and WCF-based Web services, using tools supported in Visual Studio (Joseph, 2009).

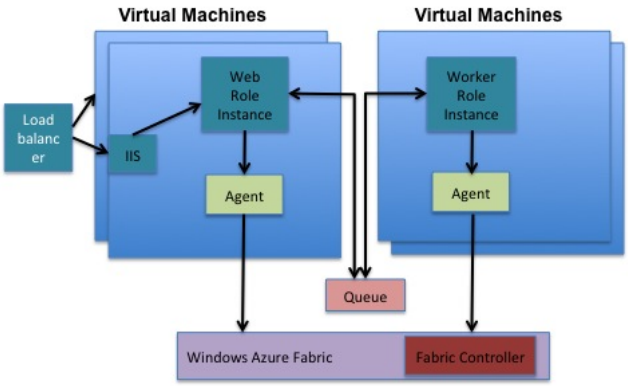

Figure 2. Window Azure and roles. Courtesy (Joseph, 2009)

Open source private cloud vendor Eucalyptus has targeted high-availability in Eucalyptus 3.0 (Kerner, 2011). In Eucalyptus 3.0 there are now multiple controllers for high-availability. The controllers are web services that help to orchestrate the real time operation of the cloud. In terms of deployment, this can be made across two or more racks with separate controllers on each. The high-availability feature will detect networking, compute, memory and hardware failures and then fail-over to a working stable node.

3. Management and Monitoring of the Cloud

As mentioned earlier, establishing robust monitoring and management tools and practices for the cloud architectures will create long term benefits in terms of achieving high availability. Traditional infrastructures have several dedicated servers for a service to be available in a consistent and continuous manner. However, managing these servers can get extremely complex with more and more additions to the servers and the applications they support. There are several third party vendors that are on the rise that provide monitoring and cloud management tools such as Cloudkick (Rackspace, 2011), LogicMonitor (http://www.logicmonitor.com/downloads/Architecture.pdf?84cd58), Pandora FMS (Pandora, 2011) etc. More importantly several vendors provide tools that only provide monitoring of the server or the service up-time (Barry, 2011). Service availability, however, is much more than just ensuring the server uptime. It is also to ensure that the communication infrastructure between the cloud and the consumer is assured since it is directly outside of the consumer's control.

In order for the consumer and businesses to adopt cloud services it becomes essential that they are testing adequately before using them.

3.1 Testing the Cloud

As with any testing, the need for testing the cloud is to achieve a high level of customer satisfaction in addition to mitigating the risks associated with the processes. Some of the potential risks involved can be loss of confidential consumer's data, loss of data integrity, security breaches, downtime etc. Using testing and monitoring tools can go a long way in ensuring that service providers are providing the cloud services with high availability.

There are a number of layers where the testing can be implemented (Barry, 2011). These layers are Wide Area Network (WAN), which provides communication between the customer and the data communication services. The second layer is between the data communication services and data centers in a LAN. Third layers is monitoring of the data center's performance and availability. And finally testing can be performed on the individual servers for their performances.

Since the WAN layer is not owned by the cloud service provider it may sometimes be a bottleneck. However, it is possible to monitor the performance of the data communication layer between the data centers and service provider using other third-party monitoring tools. The acceptance criteria should be satisfying the SLAs already agreed upon.

Functional and non-functional testing techniques can be established to certify these layers. These range from documenting clear, concise and complete business requirements. Static testing can be performed to ensure that this is accomplished. Other standard testing techniques include system testing to ensure individual components are functioning as expected. Integration testing to ensure individual components talk to each other as expected in a broader network. Load, stress and performance testing techniques can be utilized for certifying performance of individual servers. Corporates such as HP and Yahoo are also working on Test Beds require for testing the cloud services. These test beds are expected to work at Internet scale (Applabs, 2009).

4. Open Issues with Cloud Availability

Security is one of the concerns for a cloud service and can be impacted by a flawed hypervisor (Myerson, 2011). All cloud services run on virtual machines and a hypervisor allows multiple operating systems to share a single hardware host. If a hypervisor is flawed, it will negatively impact all the instance resources. This can significantly impact the availability of the cloud service.

Protecting the confidentiality and integrity of data is another issue. It is very important that appropriate cryptographic measures like algorithms are in place to ensure the data is protected.

Identifying the single of point of failure is another issue. These could be present at any one of these levels like application level, data center level, infrastructure level and geographic location level. Failover strategies must be in place to ensure the failed nodes are identified and replaced immediately upon failure.

5. Conclusion

Availability of the cloud will be critical for its long-term success. As mentioned above it is not limited to availability of just the applications being used. They also need to be delivered to the consumer in a consistent way without delays. Some of the basic properties such as scalability and flexibility are still important factors. Analyzing and proactively mitigating the risks and identifying single points of failure involved in the cloud can go a long way in achieving high availability of the cloud. This will protect the cloud assets also resulting in consumer buy into taking advantage of cloud services.

Deploying the appropriate monitoring and performance logging tools can be instrumental in showing the areas of improvements and optimization. Service level agreements (SLAs) between the consumer and cloud service providers can be help improvement the cloud availability.

Disaster recovery for the cloud-based services is another important area to explore. Lot of companies offering services often invest in Disaster Recovery setup and infrastructure to ensure their services are not disrupted in the event of a disaster such as natural calamity. It will be interesting to see how a similar setup can be established and experimented.

Acknowledgement

This research has been supported by the Fidelity National Financial Distinguished Professorship in Computer and Information Sciences.

References

Applabs. (2009). *Testing the Cloud, White Paper by AppLabs.* Retrieved from

http://www.qaguild.com/upload/app_whitepaper_testing_the_cloud_1v00.pdf

Barry, J. (2011). *Testing the Cloud: Assuring Availability, Napatech.* Retrieved from http://www.hpcinthecloud.com/hpccloud/2011-08-16/testing_the_cloud:_assuring_availability.html

Joseph, J. (2009). *Patterns For High Availability, Scalability, and Computing Power With Windows Azure.* Retrieved from http://msdn.microsoft.com/en-us/magazine/dd727504.aspx

Kerner, S. M. (2011). *Eucalyptus 3.0 Advances Private Cloud Availability.* Retrieved from http://www.serverwatch.com/server-news/eucalpytus-3.html

Kupferman, J., Silverman, J., Jara, P., & Browne, J. (2009). *Scaling Into the Cloud.* Retrieved from http://cs.ucsb.edu/~jkupferman/docs/ScalingIntoTheClouds.pdf

LogicMonitor, Architecture White Paper. Retrieved from http://www.logicmonitor.com/downloads/Architecture.pdf?84cd58

Myerson, J. M. (2011). *Cloud services: Mitigate risks, maintain availability; Maintain high availability in a cloud environment using cloud service security policy.* Retrieved from http://www.ibm.com/developerworks/cloud/library/cl-cloudservicerisks/index.html?ca=drs-

Nimsoft. (2009). Ensuring High Service Levels in Cloud Computing, Keys to Effective Service Management. Retrieved from http://www.techrepublic.com/whitepapers/ensuring-high-service-levels-in-cloud-computing-keys-to-effective-service-management/1188477

Pandora. (2011). Virtualization and cloud computing monitoring; Artica Soluciones Tecnológicas (1st ed.). Retrieved from http://pandorafms.com/downloads/PandoraFMS_Virtual_Enviroment_Monitoring.pdf

Rackspace, H. (2011). *High Availability Cloud Environments, White Paper by Rackspace Hosting.* Retrieved from http://www.codeproject.com/Articles/157992/High-Availability-Cloud-Environments

Rackspace. (2010). Architecting High Availability Linux Environments within the Rackspace Cloud, A detailed exploration into the technical requirements and business implications of Cloud High Availability, White Paper by The Rackspace Cloud Engineering Team. Retrieved from http://c0179631.cdn.cloudfiles.rackspacecloud.com/High_Availibilty_Cloud_Feb_16.pdf

Silva, P. (2010). Availability and the Cloud. Retrieved from http://ismny.org/uploads/F5_Availability_and_Cloud.pdf

Sun, M. (2009). Introduction to Cloud Computing Architecture White Paper (1st ed.). Retrieved from http://eresearch.wiki.otago.ac.nz/images/7/75/Cloudcomputing.pdf