the Availability Digest

Availability in the Cloud
January 2018

Businesses and their employees demand 100% availability of their applications. Such high availability is an enabler of worker productivity and provides differentiation from the business' competitors.

But some downtime is inevitable, whether from software or hardware upgrades, overloads, software bugs, or system crashes. According to one disaster recovery survey, 77% of companies have a goal of 99.9% availability (less than one hour per month of downtime). Yet 57% of companies had at least one outage in the past three months, each at a cost of $10,000 or higher.

Moving to the Cloud for High Availability

The nature of the cloud can generally guarantee high availability. A typical cloud SLA (Service Level Agreement) will guarantee an availability of 99.99% or higher. This is less than one hour of downtime per year and is definitely considered high availability.

Cloud services are complex. Troubles can arise from the interactions of multiple components and automated systems over distributed networks and data centers. This results in issues that can take a long time to resolve.

The availability of applications can be improved by using multiple cloud providers. Services can be hosted and replicated to various providers at several locations. Even downtime related to maintenance can be reduced by spreading a service over multiple providers.

Do Clouds Really Provide High Availability?

Virtually all cloud providers offer effective data backup and restore solutions. But backing up and restoring data is only part of what is really needed. What is needed is availability, and this is a much broader issue. The level of availability that is provided also affects confidentiality and data integrity.

Confidentiality is keeping sensitive data secret. Data integrity means that only authorized persons can change the data. Availability involves delivering the correct data when it is needed.

Designing a system to enhance any one of these attributes – confidentiality, data integrity, availability – often compromises the other two attributes. Implementing a solution that emphasizes availability provides more resilient access to data and services, but at a cost of less confidentiality and data integrity. Moving toward confidentiality indicates a solution that restricts unauthorized access but at the cost of reduced data integrity and availability.

One of the great advantages of the cloud is the ability to dynamically distribute and manage workloads across servers, data centers, and even cloud providers. Clouds can spin up resources to meet increased demands and can retire these resources when they are no longer needed.

However, applications must be designed to be stateless so that they can be recovered following a server crash. They must be able to scale seamlessly to accommodate peak usage demands. They must plan for significant outages such as power or network failures by providing data replication to different locations. Provisions must be made to manage DNS (Domain Name Services) resources so that traffic can be dynamically routed and balanced across multiple physical locations.

Given these constraints, improved availability and decreased downtime are among the biggest benefits of moving to a cloud-based infrastructure.

Balancing Confidentiality, Integrity, and Availability

To balance confidentiality, data integrity, and availability, several questions must be answered:

- Is it better to lose data permanently or have it fall into the wrong hands?
    This is a balance between availability and confidentiality.

- Is keeping the data tamper-free more important than unplanned data loss?
    The answer helps to determine whether to focus on integrity or availability, or to balance between the two.

- Are all of these decisions unacceptable, and I need absolute confidentiality, integrity, and availability?
    If so, plan on spending time and money. Such comprehensive no-compromise solutions are rarely cost-effective, even in a cloud scenario.

- How long can my company operate without access to cloud data and services?
    This question gets right to the point If the cloud is down, does that result in a minor inconvenience or a profit-shaking catastrophe? Would you gladly risk your data going public in order to get access restored?

Does My Cloud Provider Meet My Expectations?

Once some thought has been given to the questions above, there are two aspects of availability that must be verified with the cloud provider – availability and recovery.

Availability is how much time the service provider guarantees that data and services are available. This is typically presented as a percent of time per year. For instance, 99.999% availability (or five 9s) uptime means that resources will be unavailable for access for no more than about five minutes per year.

An appropriate cloud provider publishes and guarantees availability. The guarantee compensates the company for missing the documented availability metrics. The better the guarantee and availability, the more reliable and expensive the service is.

Recovery is effective for services that are not business-critical. Recovery operations include restore from backup, repair of damaged systems, and resolving viruses and denial-of-service attacks. Most cloud providers provide recovery features such as redundant storage, multi-site mirroring, multiple Internet paths, and rapid patch management.

Four Steps for Achieving High Availability in the Cloud

The following are important steps for achieving cloud availability:

1. Build for server failure:
    Set up auto-scaling.
    Set up master/slave configurations with database mirroring to ensure data integrity and minimum downtime.
    Use dynamic DNS and static IPs so that the application's infrastructure always has the right context.

2. Build for zone failure:
    Zones are distinct locations that are engineered to be insulated from failures on other zones.
    Spread the servers in an application across at least two zones.
    Replicate data across zones.

3. Build for cloud failure:
    Backup or replicate data across multiple cloud providers.
    Maintain sufficient capacity to absorb zone or cloud failures.

4. Automate and test everything:
    Automate backups.
    Set up monitoring and alerts to identify problems as they occur.

Summary

One route businesses are employing to achieve high availability is to take advantage of the cloud. High-availability is the holy grail of the cloud. The cloud embodies the idea of anywhere and anytime access to services, tools and data. It is the enabler of visions of a future with companies with no physical offices and of global companies with completely integrated and unified IT systems.

Acknowledgements

Information for this article was taken from the following sources:

What high availability for cloud services means in the real world, *Tech Republic*; November 20, 2011.
Four Steps to Achieving High Availability in the Cloud, *Right Scale*; May 9, 2012.
Availability Requirements for Cloud Computing, *Petri*; July 18, 2012.
The Benefits of Always-On Cloud Availability, *Nerdio*; September 27, 2016.