Comparing Public Cloud Service Level Agreements
Questions to ask when evaluating public cloud SLAs

Introduction

Service level agreements (SLAs) are one of the major considerations for every buyer of cloud computing services. The question often asked is how many nines of availability a given provider guarantees.

Most buyers with performance-sensitive applications typically require a minimum of three nines (99.9%), with many seeking a provider who can guarantee four (99.99%) or more nines.

Oftentimes, SLAs cannot be simplified into the number of nines. Finding the SLA truth, and whether a service provider will meet your requirements, depends on the details of how they define their SLA measures and penalties. This white paper explores the questions to ask when comparing public cloud SLAs to determine whether a cloud providers SLA will meet your business needs.
As a reference, below is the amount of unscheduled monthly downtime acceptable under various SLA levels:

<table>
<thead>
<tr>
<th>Availability % (SLA)</th>
<th>Downtime per year</th>
<th>Downtime per month*</th>
<th>Downtime per week</th>
</tr>
</thead>
<tbody>
<tr>
<td>90% ('one nine')</td>
<td>36.5 days</td>
<td>72 hours</td>
<td>16.8 hours</td>
</tr>
<tr>
<td>99% ('two nines')</td>
<td>3.65 days</td>
<td>7.20 hours</td>
<td>1.68 hours</td>
</tr>
<tr>
<td>99.5%</td>
<td>1.83 days</td>
<td>3.60 hours</td>
<td>50.4 minutes</td>
</tr>
<tr>
<td>99.9% ('three nines')</td>
<td>8.76 hours</td>
<td>43.8 minutes</td>
<td>10.1 minutes</td>
</tr>
<tr>
<td>99.95%</td>
<td>4.38 hours</td>
<td>21.56 minutes</td>
<td>5.04 minutes</td>
</tr>
<tr>
<td>99.99% ('four nines')</td>
<td>52.56 minutes</td>
<td>4.32 minutes</td>
<td>1.01 minutes</td>
</tr>
<tr>
<td>99.999% ('five nines')</td>
<td>5.26 minutes</td>
<td>25.9 seconds</td>
<td>6.05 seconds</td>
</tr>
<tr>
<td>99.9999% ('six nines')</td>
<td>31.5 seconds</td>
<td>2.59 seconds</td>
<td>0.605 seconds</td>
</tr>
<tr>
<td>99.99999% ('seven nines')</td>
<td>3.15 seconds</td>
<td>0.259 seconds</td>
<td>0.0605 seconds</td>
</tr>
</tbody>
</table>


Service level agreements (SLAs) are one of the major considerations for every buyer of cloud computing services.

Areas covered by SLAs

Cloud hosting and SLAs may provide protection at several different levels across infrastructure operating systems, and applications. Below are several of the important coverage levels that may be covered by a cloud SLA.

- **Facilities-level SLA** –
  In a traditional colocation contract structure, the colocation provider will typically offer an SLA covering the data centre facilities required to support the client-owned hardware. These include items such as power, on-site generators, cooling, etc. This type of coverage is generally considered the bare minimum in the hosting market. A cloud service provider (CSP) will typically choose a colocation facility that maintains a facilities-level SLA.

- **Platform-level SLA** –
  The next level of protection in a public cloud hosting agreement typically covers physical server, virtualisation platform and network hardware owned by the cloud service provider and used by the cloud hosting client. Typically, the physical server and virtualisation software are covered by a Platform SLA. The cloud network (network between Cloud Servers) may be covered by a separate availability SLA.

- **Operating system SLA** –
  The operating system (OS) is the next potential area of coverage for a cloud hosting SLA. Cloud Service Providers offering an OS-level SLA typically provide some degree of managed services to a client. This additional service allows the vendor to guarantee that the operating system is properly maintained so that it is consistently available, and typically has some caveats (further details below).

- **Application-level SLA** –
  This type of SLA provides protection against application level failures up to and including the custom application running on the provider’s hardware. This type of SLA is extremely rare in the cloud hosting market. Under this model, the cloud hosting provider is guaranteeing the availability and performance of their client’s software, which is a difficult commitment to meet.
The SLA truth is in the details – questions to ask

Not surprisingly, the committed uptime percentage and areas of coverage should only be the starting points. The SLA truth is very much in the details. There are a number of questions to ask and details to confirm with any provider's SLA. For example:

**How are uptime and availability calculated?**

**Best practices:** SLAs are typically calculated monthly, and only consider the period during which you were a client in the calculation of uptime versus downtime.

**Cautions:** As an example of SLAs not meeting this best practice, Amazon Web Services calculates its availability period starting 12 months prior to a client joining the service, with the assumption that the previous 12 months were delivered at 100% uptime (whether you were a client then or not)

This means that if a client experiences a 60-minute outage on the first day of service, the AWS SLA would be calculated at 99.99% uptime because the previous 12-months are always included in their analysis.

Amazon, Microsoft, and HP are three well-known cloud providers who require multiple availability zones to be in use for an outage to be considered an SLA violation.

**What should be in place for me to be covered?**

**Best practices – physical hardware:** If hosting with physical hardware, providers will commonly mandate that clients deploy redundant and/or clustered systems in order to protect against failure of isolated hardware components, which no one can totally control (hardware inevitably will fail). This means that clients need to deploy web server farms, clustered database solutions, etc. in order to be covered by a server hardware SLA.

**Best practices – virtual machines (VMs):** In the case of virtual machines, the hardware layer has been abstracted and typically the hypervisor allows VMs to recover from hardware failures without human interaction. In this case, a client should not expect exceptions in the SLA for hardware failures as current technology allows infrastructure providers to architect around those limitations.

**Cautions:** Be wary of providers mandating that numerous ‘availability zones’ or ‘regions’ need to fail for them to consider the failure an SLA violation. While architecting an application across numerous cloud locations is a good idea to minimise risk of downtime, the cloud provider should offer some SLA against failures of their cloud product in a single location as well.

Amazon, Microsoft, and HP are three well-known cloud providers who require multiple availability zones to be in use for an outage to be considered an SLA violation. In HP's cloud, the client is also required to attempt to build another server in the failed region and have that fail as well before they will consider the event an SLA violation.

**What about performance degradation as opposed to hard downtime?**

**Best practices:** Cloud SLAs should cover both unavailability (hard downtime), as well as performance degradation. Many providers offer clear SLA language explaining what happens if their infrastructure goes completely offline, but fail to mention whether performance degradation is also considered an SLA violation. Be certain to read the SLA carefully and ask that performance degradation and unavailability both be covered.

**Cautions:** Cloud architectures are extremely complex. A simple server-only uptime SLA fails to address one of the most important components of cloud architecture, namely the network. Consider the implication to your business should network performance vary widely. Is this something you can easily architect around? You should have a clear understanding as to whether the provider guarantees network performance as well as uptime performance.
What are the penalties for SLA violations?

**Best practices:** After a contractual threshold of downtime has been reached, most cloud providers will commit to a percentage refund of the fees paid for a given period of time (generally monthly). Be certain to carefully compare the time durations and associated percentage of fees the provider agrees to refund in the event of an outage.

**Cautions:** Providers’ definitions of an SLA violation vary widely, and clients need to review them carefully to ensure that penalties paid under various SLAs are comparable. To compare accurately, be certain you understand the following:

- The amount of downtime allowed under the SLA and associated credit amounts
- At what amount SLA refunds are capped (i.e. how much of the fees paid for a given period is the provider obligated to refund)
- What portion of the fee is eligible for refund, for example:
  - Will the provider only refund my compute-related fees and exclude network, storage, CDN fees, etc., or is the credit against my entire invoice for the month?
  - If a server (or servers) fails with the result that my entire application fails, am I credited only for the server(s) that failed, for my entire invoice, or for the compute portion of my invoice?

The answers to the questions above can greatly impact the usability and value of the SLA to your business, so ensure that you understand the level of commitment the provider is offering under its SLA.

As current industry examples, Terremark pays USD 96 per 24-hour period of downtime, up to 50% of monthly usage fees\(^i\); Amazon’s SLA caps credits at 10% of the entire monthly bill\(^ii\); HP caps credits at 30% of the costs specifically associated with the resources that failed (i.e. compute only, or storage only, etc.\(^iii\); and Rackspace Cloud pays up to 100% of the invoiced amount, but provides credits only against the costs of individual servers that failed\(^iv\); even if the impact to the affected server rendered dozens of other servers useless in delivering the affected application.

What do I have to do to request a credit?

**Best practices:** In a perfect world, the provider would automatically credit a client’s invoice for any downtime, but this is not the case in the cloud market today. Most providers require their clients to submit a written request for the credit within a given period of time from the date of the outage. A 30-day notification requirement is relatively standard.

**Cautions:** A period of less than 30 days for a valid credit request would be considered somewhat restrictive given that for some amount of time following an outage, clients are likely addressing the business issues caused by the unavailability of their systems.

Also, review what amount of information is required to justify to the provider that downtime occurred. If the burden of proof is so high that the work required to justify the outage costs more in terms of time than the value of the credit, the value of the SLA is considerably reduced.

Summary

As is demonstrated by the examples above, there is much more to a provider’s cloud SLA than simply an uptime percentage guarantee, and savvy buyers need to investigate their options carefully to ensure that they are comfortable with the guarantee offered by their provider.

Ultimately, if and when downtime occurs (as it will everywhere at some point), a provider’s SLA and related monetary performance penalties will rarely offset the damage done to your business by the outage (and it’s not realistic to assume that you can ever negotiate an agreement that will fully mitigate the impact).

If that’s the case, do SLAs really matter? According to a NetworkWorld interview with Gartner’s lead cloud analyst, many cloud SLAs are “practically useless.”\(^v\) In reality, the penalties paid under an SLA are arguably the least meaningful part of the agreement given all of the restrictions and caveats described above. Much more important is the message communicated by the detailed terms and requirements within each provider’s SLA.

For example, does the provider consider downtime the same way as you do in your business? If a server fails and your business goes offline, do you have to carefully evaluate the SLA to determine whether the provider considers this a downtime event, or is it clear that their service has failed and the SLA has been violated?

SLA exclusions and caveats are often an indication of how and how often the provider expects its service to fail, and in which way. If a provider requires their clients to build in 2+ availability zones, and both zones have to fail before the SLA is violated, this indicates that the provider expects single zones to fail. If a provider requires clients to attempt to build a new server when one server fails before considering it an SLA violation, the message is that single servers are expected to fail regularly. If a provider offers a server SLA, but not a network performance SLA, it’s likely because it cannot guarantee performance of its network.

Despite the public cloud limitations discussed in this report, carefully reviewing your cloud provider’s SLA will give you an idea of how much downtime the provider views as acceptable, and to what extent it is willing to commit its own financial resources to compensate for unexpected outages.

If you have additional questions about your options in the public cloud arena and would like to speak to one of our experts, please contact us at cloud-sales@dimensiondata.com or visit our website at www.dimensiondata.com/cloud www.dimensiondata.com/cloud/sla

---

\(^i\) http://aws.amazon.com/ec2-sla/, as of 12/11/2012
\(^iv\) http://www.rackspace.com/cloud/legal/sla, as of 12/11/12