Cloud IaaS: Service-Level Agreements

Published: 7 March 2011

Analyst(s): Lydia Leong

Cloud infrastructure-as-a-service (IaaS) providers typically offer SLAs that cover the various elements of the service. IT managers need to understand how SLAs are derived, how SLA penalties work and what to negotiate for.

Key Findings

■ Cloud IaaS SLAs are similar to SLAs for network services, hosting and data center outsourcing.
■ SLA penalties are usually applied as a service credit and are typically capped at the equivalent of a month’s service fees.

Recommendations

■ Match your application availability and performance needs to the appropriate IaaS offering and SLAs.
■ Don’t waste your time trying to get a service provider to accept uncapped legal liability for service failures; most service providers do not agree to uncapped damages.

Table of Contents

Analysis..................................................................................................................................................2
SLAs in the Cloud..................................................................................................................................2
  Penalties for SLA Violation..................................................................................................................2
  Compute SLAs................................................................................................................................3
  Network SLAs..................................................................................................................................4
  Storage SLAs...................................................................................................................................4
  Other SLAs.......................................................................................................................................4
Limitation of Liability.........................................................................................................................5
Sourcing Considerations.....................................................................................................................5
Recommended Reading.........................................................................................................................6
Analysis

As described in "Evaluating Cloud Infrastructure as a Service," all cloud IaaS offerings are not created equal, despite superficial similarities in the way the offerings are described. There is considerable variance in service provider design goals, the quality of technical implementations, and the cost-effectiveness and value for money of those implementations.

This is part of a series of reports detailing the market options, including technical and business characteristics, for IaaS offerings. This document is focused on SLAs.

SLAs in the Cloud

When considering the usefulness of an SLA for the availability and performance of cloud IaaS, it is important to consider three different aspects:

- The engineered level (mathematically calculated).
- The SLA that the customer receives.
- The provider’s track record of delivering on its SLA.

There is a significant difference between engineering and hope. Many internal IT organizations, for instance, set service-level objectives (SLOs) that are based on what they hope to achieve, rather than the level that the solution is engineered to achieve and can be mathematically expected to deliver, based on calculated mean time between failures (MTBF) of each component of the service. Many organizations are lucky enough to achieve service levels that are higher than the engineered reliability of their infrastructure. IaaS providers, however, are likely to base their SLAs on their engineered reliability, not on hope.

The SLA that is offered to customers may be higher or lower than the provider’s internal SLO and the solution’s engineered level. This is because there are other factors involved in determining what the SLA should be, such as the SLA’s marketing value and the business risks that the provider is willing to take.

Penalties for SLA Violation

In general, SLAs for cloud IaaS are structured quite similarly to SLA agreements for network services, hosting and data center outsourcing. SLA metrics are typically calculated on a monthly basis. If an SLA is violated, you are entitled to a credit for some part of your service fee. The amount of the credit is usually based on the length of time for which the SLA was violated, multiplied by the cost of the affected service. Longer single violations, longer cumulative violations or repeated violations may carry disproportionately higher penalties.

The penalty for an SLA violation is typically capped at 100% of one month’s service fees. In the case of traditional data center outsourcers offering IaaS, the cap might be much lower, matching the common outsourcing cap of 20%. This cap may be based on everything you typically pay the service provider during a month, or it may be capped at just the portion of your bill that is directly
related to the specific type of violation; for instance, if you experience a compute outage, you might only be allowed to get credits equal to your payment for compute, not counting anything that you pay for storage or network bandwidth. As cloud IaaS fees are often variable on a monthly basis, the credit amount might be based on the previous month’s bill, your average bill over a period of time (such as the previous year), or your baseline minimum contracted obligation.

Like other outsourcers, providers generally do not agree to SLA penalties that are based on metrics other than the service fee. For instance, even if the customer is an e-commerce provider and downtime represents lost revenue, any SLA penalty will normally be based on the service fee, not on the customer’s lost revenue.

Many providers have standard contracts that do not contain an option to terminate for nonperformance. We strongly recommend that customers negotiate contracts that allow them to terminate for cause, both for repeated SLA violations and for single incidents of a scope that would reasonably cause them to doubt the provider’s continued ability to deliver successfully on the contract.

### Compute SLAs

Providers typically offer an SLA for compute infrastructure availability. This SLA is normally calculated over the course of a calendar month. There are generally two aspects to this SLA:

- **Cumulative availability.** This is the availability over the course of the entire month. This SLA is usually 99.5% or higher, and many cloud IaaS providers offer 99.99% or higher. We consider reasonable cumulative availability SLAs to be:
  - 99.5% for single-data-center, non-high-availability, non-production infrastructure.
  - 99.75% for single-data-center, high-availability, production infrastructure.
  - 99.95% for multiple-data-center, high-availability, production infrastructure.

- **Outage length.** This SLA provides compensation for each outage exceeding a certain length. The typical minimum outage length for compensation is 10 minutes or more.

In both cases, there is usually a sliding scale for SLA penalties; the more by which the SLA is missed, the greater the penalty relative to the amount of downtime. Planned and/or emergency maintenance windows may be excluded from the SLA, but best-in-class SLAs exclude only customer-initiated maintenance. The burden of the claim process varies with the provider. For instance, some require you to have filed a trouble ticket at the time of the outage, or to be able to independently prove that the outage occurred via system logs, monitoring data or the like. However, in practice, most providers with claim rules do not enforce them; if the provider’s own monitoring data indicates an outage, SLA claims will generally be respected.

It is relatively common for cloud providers to offer 100%-availability SLAs. Such SLAs should be read carefully. Usually, these are outage-length SLAs, but they may also have a cumulative availability aspect; for instance, such an SLA may offer credits for any outage greater than 15 minutes, or total downtime exceeding 15 minutes (equivalent to 99.999% cumulative availability).
Because compute capacity is not usually oversubscribed, there are normally no SLAs for compute performance; customers are simply guaranteed to have the compute for which they paid, with technical enforcement at the hypervisor level.

Some providers may offer provisioning-related SLAs, such as guaranteeing the ability to provision a certain amount of compute resources, or provisioning a virtual machine within a certain time.

Network SLAs

Providers usually offer a network SLA, which covers availability, packet loss and jitter for Internet bandwidth. Internet SLAs are usually aggressive, with 100% availability and no more than 0.1% cumulative packet loss over the course of a month. Denial-of-service (DoS) attacks may be explicitly excluded from the SLA, even though cloud IaaS providers typically protect all customers with DoS mitigation measures. Like most Internet bandwidth SLAs, this is not end to end; it generally governs the cloud provider’s data center connectivity to the Internet as a whole, not to any specific provider or customer.

The market is evolving toward SLAs that cover Internet performance to specific points of presence. Such SLAs are likely to use third-party measurement services, such as Keynote Systems, Gomez (Compuware) and Webmetrics (Neustar).

Some providers may also provide an internal network SLA that covers availability, latency, packet loss and jitter between devices within their data center and, if they have multiple data centers, between those data centers.

Customers that obtain private network connectivity, instead of Internet connectivity, will receive whatever SLA their carrier is willing to offer them.

Storage SLAs

Providers are increasingly offering storage SLAs covering availability and performance. Storage performance is usually measured in throughput, or in input/output operations per second (IOPS). However, sometimes a storage performance SLA will actually be an internal network SLA, covering latency between compute and storage.

Service providers do not usually offer SLAs concerning backup and recovery, but when they do, they generally guarantee a restore time, rather than a commitment that a restore will work. We expect that backup services will eventually have more stringent SLAs.

We also expect that, over time, the market will evolve to offer SLAs that cover data loss and data integrity.

Other SLAs

Providers may also offer more granular SLAs, including application SLAs or even SLAs for individual transactions. Such SLAs are typically associated with managed services or more complex forms of outsourcing, such as Infrastructure Utility for SAP (IU4SAP). The more services that are provided above the "bare metal" compute infrastructure layer, the more likely it is that SLAs will be focused
on relevancy to the business use of the infrastructure, rather than the infrastructure components themselves.

Providers may offer SLAs related to customer service, such as the amount of time it takes for a particular type of service request to be responded to and carried out, or failure on the part of the service provider to meet certain escalation procedures.

Limitation of Liability

Like other outsourcers, cloud IaaS providers normally explicitly limit their liability in their contracts, although the precise amount is typically negotiable. However, liability limits might not be enforceable; for instance, if you were to successfully sue your service provider for negligence, it is possible that a judge or jury might award damages greater than the liability limits in the contract. In practice, however, most of these type of lawsuits are settled in arbitration. Do not use a provider’s willingness to assume greater liability as a proxy for assessing the risks of outsourcing to this provider, and do not waste time trying to get a provider to assume uncapped liability.

It is more useful to negotiate stronger SLAs and steeper SLA penalties than it is to negotiate greater amounts of liability. (See "How to Structure Incentives and Penalties in Outsourcing Contracts" for more on contract negotiation.) SLA penalties are routinely successfully claimed by customers without any more trouble than going through the provider’s paperwork process for claims; recovering damages requires a lawsuit. Ensure, however, that your requests are realistic, given your architecture; if you want higher levels of availability and greater security, you must be prepared to implement, and pay for, the technology to achieve them.

If your risks in the event of, for example, downtime or a data breach significantly exceed the limits of liability provided in the contract, you may wish to consider some other way of financially mitigating your risk. For instance, property and casualty (P&C) insurance may include business interruption insurance, though you will need to ensure that it includes coverage for third-party service providers. You could also consider purchasing cyber-risk insurance (see "Managing Availability and Performance Risks in the Cloud: Expect the Unexpected").

Sourcing Considerations

When sourcing an IaaS solution, for each application that you plan to run on it, consider:

- How critical is this application?
- What is this application’s acceptable availability, recovery time objective and recovery point objective?
- What are the performance characteristics and needs of this application?

Compare these needs with the service levels available from the providers to determine which environments can meet your requirements.
Recommended Reading

"Cloud IaaS: How Compute Resources Are Delivered"

"Cloud IaaS: Adding Storage to Compute"

"Cloud IaaS: Networking Options"

"Cloud IaaS: Security Considerations"

"Cloud IaaS: Service and Support Models"
### Regional Headquarters

**Corporate Headquarters**  
56 Top Gallant Road  
Stamford, CT 06902-7700  
USA  
+1 203 964 0096

**Japan Headquarters**  
Gartner Japan Ltd.  
Aobadai Hills, 6F  
7-7, Aobadai, 4-chome  
Meguro-ku, Tokyo 153-0042  
JAPAN  
+81 3 3481 3670

**European Headquarters**  
Tamesis  
The Glanty  
Egham  
Surrey, TW20 9AW  
UNITED KINGDOM  
+44 1784 431611

**Latin America Headquarters**  
Gartner do Brazil  
Av. das Nações Unidas, 12551  
9° andar—World Trade Center  
04578-903—São Paulo SP  
BRAZIL  
+55 11 3443 1509

**Asia/Pacific Headquarters**  
Gartner Australasia Pty. Ltd.  
Level 9, 141 Walker Street  
North Sydney  
New South Wales 2060  
AUSTRALIA  
+61 2 9459 4600

---

© 2011 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner is a registered trademark of Gartner, Inc. or its affiliates. This publication may not be reproduced or distributed in any form without Gartner’s prior written permission. The information contained in this publication has been obtained from sources believed to be reliable. Gartner disclaims all warranties as to the accuracy, completeness or adequacy of such information and shall have no liability for errors, omissions or inadequacies in such information. This publication consists of the opinions of Gartner’s research organization and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice. Although Gartner research may include a discussion of related legal issues, Gartner does not provide legal advice or services and its research should not be construed or used as such. Gartner is a public company, and its shareholders may include firms and funds that have financial interests in entities covered in Gartner research. Gartner’s Board of Directors may include senior managers of these firms or funds. Gartner research is produced independently by its research organization without input or influence from these firms, funds or their managers. For further information on the independence and integrity of Gartner research, see “Guiding Principles on Independence and Objectivity” on its website, [http://www.gartner.com/technology/about/ombudsman/omb_guide2.jsp](http://www.gartner.com/technology/about/ombudsman/omb_guide2.jsp).