

1 Service Level Agreements (SLAs)

The relationship between the cloud provider and the cloud consumer must be described with a Service Level Agreement. Because cloud consumers trust cloud providers to deliver some of their infrastructure services, it is vital to define those services, how they are delivered and how they are used.

An SLA is the foundation of the consumer's trust in the provider. A well-written SLA codifies the provider's reputation.

In addition to the prose that defines the relationship between the consumer and provider, an SLA contains Service Level Objectives (SLOs) that define objectively measurable conditions for the service. The consumer must weigh the terms of the SLA and its SLOs against the goals of their business to select a cloud provider.

1.1 *What is an SLA?*

An SLA defines the interaction between a cloud service provider and a cloud service consumer. An SLA contains several things:

- A set of services the provider will deliver
- A complete, specific definition of each service
- The responsibilities of the provider and the consumer
- A set of metrics to determine whether the provider is delivering the service as promised
- An auditing mechanism to monitor the service
- The remedies available to the consumer and provider if the terms of the SLA are not met
- How the SLA will change over time

The marketplace features two types of SLAs: Off-the-shelf agreements and agreements negotiated between a provider and consumer to meet that consumer's specific needs. It is unlikely that any consumer with critical data and applications will be able to use the first type. Therefore the consumer's first step in approaching an SLA (and the cloud in general) is to determine how crucial their data and applications are.

Most public cloud services offer a non-negotiable SLA. With these providers, a consumer whose requirements aren't met has two remedies:

1. Accept a credit towards next month's bill (after paying *this* month's bill in full), or
2. Stop using the service.

Clearly an SLA with these terms is unacceptable for any mission-critical applications or data. On the other hand, an SLA with these terms will be far less expensive than a cloud service provided under a negotiated SLA.

1.2 Service Level Objectives

An SLO defines a characteristic of a service in precise, measurable terms. Here are some sample SLOs:

- The system should never have more than 10 pending requests.
- Throughput for a request should be less than 3 seconds.
- Data streaming for a read request should begin within 2 seconds.
- At least five instances of a VM should be available 99.99999% of the time, with at least one instance available in each of a provider's three data centers.

Obviously different Service Level Objectives will apply to different use cases, applications and types of data. SLOs can also include an *urgency* to indicate the relative importance of different SLOs. A consumer could use an urgency rating to indicate that availability is more important than response time if the cloud provider cannot deliver both SLOs.

Different roles also affect the SLOs that apply. For example, consider an application written by a cloud consumer, hosted by a cloud provider and accessed by an end user. If the application and its data are hosted by the same cloud provider, chances are the application can access that data without leaving the provider's data center. The cloud consumer will expect very fast response times whenever the application accesses its data. On the other hand, the consumer will have lower expectations of response times whenever an end user accesses the application across the Web.

1.3 Service Level Management

It is impossible to know whether the terms of the SLA are being met without monitoring and measuring the performance of the service. Service Level Management is how that performance information is gathered and handled.

Measurements of the service are based on the Service Level Objectives in the SLA.

A cloud provider uses Service Level Management to make decisions about its infrastructure. For example, a provider might notice that throughput for a particular service is barely meeting the consumer's requirements. The provider could respond to that situation by reallocating bandwidth or bringing more physical hardware online. However, if giving one consumer more resources would make it impossible to meet the terms of another consumer's SLA, the provider might decide to keep one customer happy at the expense of another. The goal of Service Level Management is to help providers make intelligent decisions based on its business objectives and technical realities.

A cloud consumer uses Service Level Management to make decisions about how it uses cloud services. For example, a consumer might notice that the CPU load on a particular VM is above 90%. In response, the consumer might start another VM. However, if the consumer's SLA says that the first 100 VMs are priced at one rate, with more VMs priced at a higher rate, the consumer might decide not to create another VM and incur higher charges. As with the provider, Service Level Management helps consumers make (and possibly automate) decisions about the way they use cloud services.

1.4 Considerations for SLAs

As consumers are deciding what terms they need in an SLA, there are a number of factors they should consider.

1.4.1 Business Level Objectives

Debating the terms of an SLA is meaningless if the organization has not defined its business level objectives. A consumer must select providers and services based on the goals of the organization. Defining exactly what services will be used is worthless unless the organization has defined why it will use the services in the first place.

This is an aspect of cloud computing in which the hardest problems are organizational politics rather than technical issues. Getting all parts of an organization to agree on those goals will involve some groups accepting budget cuts, some groups losing control of their infrastructure and other difficult choices. Despite these difficulties, the organization cannot make the most of cloud computing (or any technology, for that matter) until the business level objectives have been defined.

Consumers should know *why* they are using cloud computing before they decide *how* to use cloud computing.

1.4.2 Business Continuity and Disaster Recovery

Many consumers use the cloud for business continuity. Some consumers store copies of valuable data in multiple clouds for backup. Other consumers use cloudbursting when in-house data centers are unable to handle processing loads. The cloud can be an invaluable resource to keep an organization running when in-house systems are down, but none of that matters if the cloud provider itself does not have adequate continuity and disaster recovery procedures. Consumers should ensure their cloud providers have adequate protection in case of a disaster.

1.4.3 System Redundancy

Many cloud providers deliver their services via massively redundant systems. Those systems are designed so that even if hard drives or network connections or servers fail, consumers will not experience any outages. Consumers moving data and applications that must be constantly available should consider the redundancy of their provider's systems.

1.4.4 Maintenance

Providers handle the maintenance of their infrastructure, freeing consumers from having to do that themselves. However, consumers should understand how and when their providers will do maintenance tasks. Will their services be unavailable during that time? Will their services be available, but with much lower throughput? If there is a chance the maintenance will affect the consumer's applications, will the consumer have a chance to test their applications against the updated service? Note that maintenance can affect any type of cloud offering and that it applies to hardware as well as software.

1.4.5 Location of Data

The physical location of many types of data is restricted. If a cloud provider cannot guarantee that a consumer's data will be stored in certain locations only, the consumer cannot use that provider's services. If a cloud service provider promises to enforce data location regulations, the consumer must be able to audit the provider to prove that regulations are being followed.

1.4.6 Seizure of Data

There have been a few well-publicized instances of law enforcement officials seizing the assets of a hosting company. Even if law enforcement targets the data and applications associated with a particular consumer, the multi-tenant nature of cloud computing makes it likely that other consumers will be affected. Although there are limits to what an SLA can cover, consumers should consider the laws that apply to the provider.

1.4.7 Failure of the Provider

Any cloud provider has the potential to either go out of business or be acquired by another company. Consumers should consider the financial health of their provider and make contingency plans if the provider were to shut its doors. In addition, the provider's policies on access to the consumer's data and applications if the consumer's account is delinquent or in dispute should be made clear.

1.4.8 Jurisdiction

Consumers should understand the laws that apply to any cloud providers they consider. For example, a cloud provider could be based in a country that reserves the right to monitor any data or applications using that cloud provider's services. Given the nature of the consumer's data and applications, this might not be acceptable.

1.4.9 Cloud Brokers and Resellers

If a cloud provider is actually a broker or reseller for another cloud provider, the terms of the SLA should clarify any questions of responsibility or liability if anything goes wrong at the broker, reseller or provider facilities.

1.5 *SLA requirements*

1.5.1 Security

Security as a general requirement is discussed in detail in Chapters 6 and 7 of this paper. The security-related aspects of an SLA should be written with the security controls and federation patterns from Chapter 6 in mind. A cloud consumer must understand their security requirements and what controls and federation patterns are necessary to meet those requirements. In turn, a cloud provider must understand what they must deliver to the consumer to enable the appropriate controls and federation patterns.

1.5.2 Data Encryption

If a consumer is storing vital data in the cloud, it is important that the data be encrypted while it is in motion and while it is at rest. The details of the encryption algorithms and access control policies should be specified in the SLA.

1.5.3 Privacy

Basic privacy concerns are addressed by requirements such as data encryption, retention and deletion. In addition, an SLA should make it clear how the cloud provider isolates data and applications in a multi-tenant environment.

1.5.4 Data Retention and Deletion

Many organizations have legal requirements that data must be kept for a certain period of time. Some organizations also require that data be deleted after a certain period of time. Cloud providers must be able to prove they are compliant with these policies.

1.5.5 Hardware Erasure and Destruction

A common source of data leaks is the improper disposal of hardware. If a cloud provider's hard drive fails, the platters of that disk should be zeroed out before the drive is disposed or recycled. On a similar note, many cloud providers offer the added protection of zeroing out memory space after a consumer powers off a VM.

1.5.6 Regulatory Compliance

Many types of data and applications are subject to regulations. Some of those are laws (HIPAA for medical records in the United States), while others are industry-specific (PCI DSS for retailers who accept credit cards). If regulations must be enforced, the cloud provider must be able to prove their compliance.

1.5.7 Transparency

Under the SLAs of some cloud providers, the consumer bears the burden of proving a that the provider failed to live up to the terms of the SLA. A provider's service might be down for hours, but consumers who are unable to prove that downtime are not eligible for any sort of compensation.

For critical data and applications, providers must be proactive in notifying consumers when the terms of the SLA are breached. This includes infrastructure issues such as outages and performance problems as well as security incidents.

1.5.8 Certification

There are many different certifications that apply to certain types of data and applications. For example, consumer might have the requirement that their cloud provider be ISO 27001 certified. The provider would be responsible for proving their certification and keeping it up-to-date.

1.5.9 Terminology for key performance indicators

The term uptime can be defined in many ways. Often that definition is specific to a provider's architecture. If a provider has a data center on six continents, does uptime refer to a particular data center or *any* data center? If the only available data center is on another continent, that uptime is unlikely to be acceptable. To

make matters worse, other cloud providers will use definitions specific to *their* architectures. This makes it difficult to compare cloud services.

A set of industry-defined terms for different key performance indicators would make it much easier to compare SLAs in particular (and cloud services in general).

1.5.10 Monitoring

If a failure to meet the terms of an SLA has financial or legal consequences, the question of who should monitor the performance of the provider (and whether the consumer meets its responsibilities as well) becomes crucial. It is in the provider's interest to define uptime in the broadest possible terms, while consumers could be tempted to blame the provider for any system problems that occur. The best solution to this problem is a neutral third-party organization that monitors the performance of the provider. This eliminates the conflicts of interest that might occur if providers report outages at their sole discretion or if consumers are responsible for proving that an outage occurred.

1.5.11 Auditability

Many consumer requirements include adherence to legal regulations or industry standards. Because the consumer is liable for any breaches that occur, it is vital that the consumer be able to audit the provider's systems and procedures. An SLA should make it clear how and when those audits take place. Because audits are disruptive and expensive, the provider will most likely place limits and charges on them.

1.5.12 Metrics

Monitoring and auditing require something tangible that can be monitored as it happens and audited after the fact. The metrics of an SLA must be objectively and unambiguously defined. Cloud consumers will have an endless variety of metrics depending on the nature of their applications and data. Although listing all metrics it is impossible, some of the most common are:

- Throughput – How quickly the service responds
- Reliability – How often the service is available
- Load balancing – When elasticity kicks in (new VMs are booted or terminated, for example)
- *Need your inputs here. I'm guessing there's a basic set of six or eight.*

1.5.13 Machine-Readable SLAs

A machine-readable language for SLAs would enable an automated cloud broker that could select a cloud provider dynamically. One of the basic characteristics of cloud computing is on-demand self-service; an automated cloud broker would extend this characteristic by selecting the cloud provider on demand as well. The broker could select a cloud provider based on business criteria defined by the consumer. For example, the consumer's policy might state that the broker should use the cheapest possible provider for some tasks, but the most secure provider for others. Although substantial marketplace demand for this requirement will take some time to develop, any work on standardizing SLAs should be done with this in mind.

1.6 A note about reliability

In discussions of reliability, a common metric bandied about is the number of “nines” a provider delivers. As an example, five nines reliability means the service is available 99.99999% of the time, which translates to total system outages of roughly 5 minutes out of every 12 months. One problem with this metric is that it quickly loses meaning without a clear definition of what an outage is. (It loses even more meaning if the cloud provider gets to decide whether an incident constitutes an outage.)

Beyond the nebulous nature of nines, it is important to consider that many cloud offerings are built on top of other cloud offerings. The ability to combine multiple infrastructures provides a great deal of flexibility and power, but each additional provider makes the system less reliable. If a cloud provider offers a service built on a second cloud provider's storage service and a third cloud provider's database service, and all of those providers deliver five nines reliability, the reliability of the entire system is less than five nines. The service is unavailable when the first cloud provider's systems go down; the service is equally unavailable when the second or third providers' systems have problems. The more cloud providers involved, the more downtime the overall system will have.

Finally, as the number of cloud providers increases, the number of outside factors increases as well. If a VM and a cloud database are in the same data center, communication between the VM and the database don't require network access. On the other hand, if the cloud database is delivered by another provider, the available bandwidth between the VM and the database affects the performance and reliability of the overall system. Both cloud providers could be up and running with healthy systems, but if the network connection between them fails, the overall system is down.

To sum up, any consumer who needs to evaluate the reliability of a cloud service should know as much as possible about the cloud providers that deliver that service, whether directly or indirectly.

1.7 Cross-reference: SLA Requirements and Use Case Scenarios

At its best, an SLA protects the interests of both the cloud consumer and cloud provider. Just as a given SLA doesn't meet the needs of all consumers, every requirement discussed here doesn't apply to all customer scenarios. The following table cross-references the seven use case scenarios from Chapter 3 with the SLA requirements listed here.

Requirement	End User to Cloud	Enterprise to Cloud to End User	Enterprise to Cloud	Enterprise to Cloud to Enterprise	Private Cloud	Changing Cloud Vendors	Hybrid Cloud
Data Encryption			✓				
Privacy	✓	✓	✓	✓	✓	✓	✓
Data Retention and Deletion			✓	✓			✓
Hardware Erasure and Destruction			✓	✓			✓
Regulatory Compliance	✓	✓	✓	✓	✓	✓	✓
Transparency	✓	✓	✓	✓	✓	✓	✓
Certification	✓	✓	✓	✓	✓		✓
Terminology for Key Performance Indicators			✓	✓	✓	✓	✓
Metrics	✓	✓	✓	✓	✓		✓
Auditability	✓						

Requirement	End User to Cloud	Enterprise to Cloud to End User	Enterprise to Cloud	Enterprise to Cloud to Enterprise	Private Cloud	Changing Cloud Vendors	Hybrid Cloud
Monitoring	✓	✓	✓	✓	✓		✓
Machine-Readable SLAs				✓			

1.8 Cross-reference: SLA Requirements and Cloud Delivery Models

The following table cross-references the three NIST delivery models from Chapter 2 with the SLA requirements listed here.

Requirement	Platform as a Service	Infrastructure as a Service	Software as a Service
Data Encryption	✓	✓	
Privacy	✓	✓	✓
Data Retention and Deletion		✓	✓
Hardware Erasure and Destruction		✓	✓
Regulatory Compliance	✓	✓	✓
Transparency	✓	✓	✓
Certification	✓	✓	✓
Terminology for Key Performance Indicators		✓	✓

Requirement	Platform as a Service	Infrastructure as a Service	Software as a Service
Metrics	✓	✓	✓
Auditability			
Monitoring	✓	✓	✓
Machine-Readable SLAs		✓	

2 SLA Use Case Scenarios

This section describes specific use cases that illustrate how SLAs are used in the real world.