cloud SLA (cloud service-level agreement)

A cloud SLA (cloud service-level agreement) is an agreement between a cloud service provider and a customer that ensures a minimum level of service is maintained.

It guarantees levels of reliability, availability and responsiveness to systems and applications; specifies who governs when there is a service interruption; and describes penalties if service levels are not met.

A cloud infrastructure can span geographies, networks and systems that are both physical and virtual. While the exact metrics of a cloud SLA can vary by service provider, the areas covered are uniform.

- volume and quality of work (including precision and accuracy)
- speed
- responsiveness
- efficiency

The SLA document aims to establish a mutual understanding of the services, prioritized areas, responsibilities, guarantees and warranties provided by the service provider.

It clearly outlines metrics and responsibilities among the parties involved in cloud configurations, such as the specific amount of response time to report or address system failures.

The importance of a cloud SLA

Service-level agreements are fundamental as more organizations rely on external providers for their critical systems, applications and data.

A cloud SLA ensures cloud providers meet certain enterprise-level requirements and provide customers with a clearly defined set of deliverables.

It also describes financial penalties, such as credits for service time, if the provider fails to live up to the guaranteed terms.
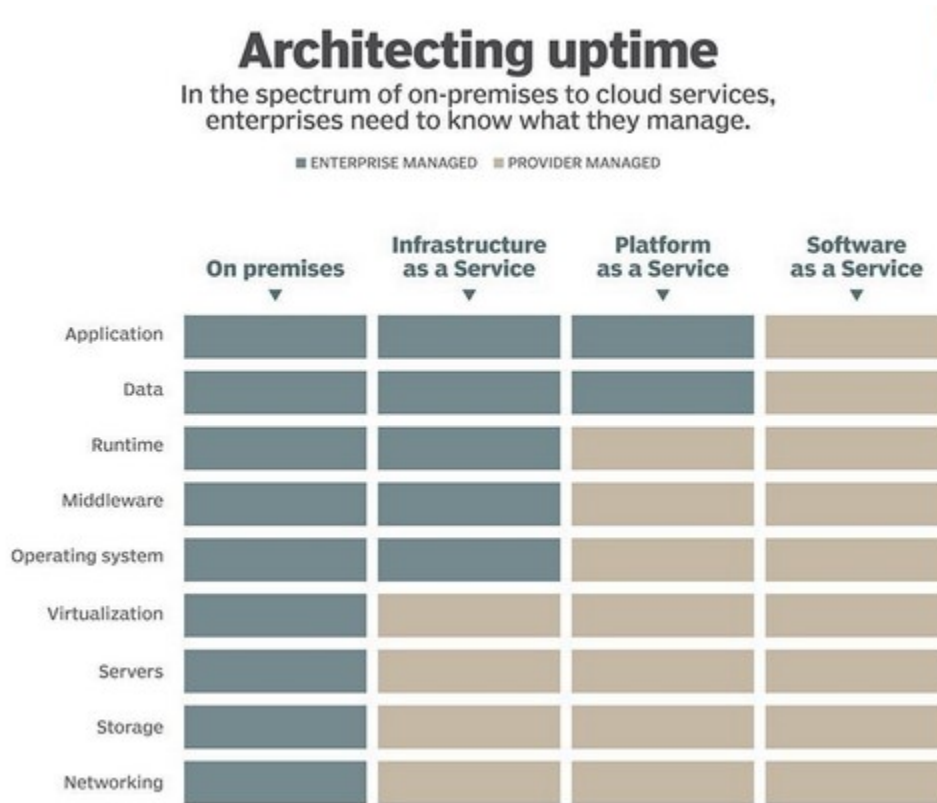
These agreed-upon rules create a trusted foundation upon which a customer commits to use a cloud providers' services. They also reflect the provider's commitments to its quality of service (QoS) and underlying infrastructure.

What to look for in a cloud SLA

The cloud SLA should outline the responsibilities of each party, the acceptable performance parameters, a description of the applications and services covered under the agreement, procedures for monitoring service levels, and a schedule for the remediation of outages.

SLAs commonly use technical definitions to quantify the level of service, such as mean time between failures (MTBF) or mean time to repair (MTTR), which specifies a target or minimum value for service-level performance.

The defined level of services should be specific and measurable, so that they can be benchmarked and, if stipulated by the agreement, trigger rewards or penalties accordingly.

Architecting uptime

In the spectrum of on-premises to cloud services, enterprises need to know what they manage.

■ ENTERPRISE MANAGED   ■ PROVIDER MANAGED

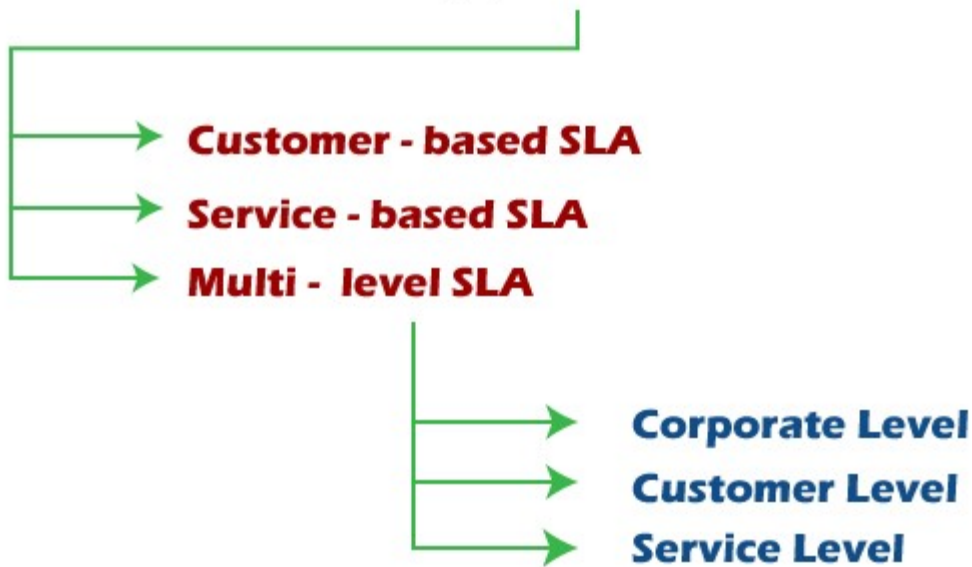|	On premises ▼	Infrastructure as a Service ▼	Platform as a Service ▼	Software as a Service ▼
Application				
Data				
Runtime				
Middleware				
Operating system				
Virtualization				
Servers				
Storage				
Networking				

A **Service Level Agreement (SLA)** is the bond for performance negotiated between the cloud services provider and the client.

Types of SLA

- Present market place features two types of SLAs :
  - Off-the-shelf SLA or non-negotiable SLA or Direct SLA
    - Non-conducive for mission-critical data or applications
    - Provider creates the SLA template and define all criteria viz. contract period, billing, response time, availability, etc.
    - *Followed by the present day state-of-the-art clouds.*
  - Negotiable SLA
    - Negotiation via external agent
    - Negotiation via multiple external agents

Most Service Level Agreements are standardized until a client becomes a large consumer of cloud services. Service level agreements are also defined at **different levels** which are mentioned below:

Types of SLA

- **Customer - based SLA**
- **Service - based SLA**
- **Multi - level SLA**
  - **Corporate Level**
  - **Customer Level**
  - **Service Level**

While some are targeted at individual customer groups, others discuss issues relevant to entire companies. This is because the needs of one user differ from another. Here are some types of SLAs used by businesses today and how each one is utilized for specific situations:

1. Customer-based SLA

This type of agreement is used for individual customers and comprises all relevant services that a client may need while leveraging only one contract. It contains details regarding the type and quality of service that has been agreed upon.

For example, a telecommunication service includes voice calls, messaging, and internet services, but all exist under a single contract.

2. Service-based SLA

This SLA is a contract that includes one identical type of service for all of its customers. Because the service is limited to one unchanging standard, it is more straightforward and convenient for vendors.

For example, using a service-based agreement regarding an IT helpdesk would mean that the same service is valid for all end-users that sign the service-based SLA.

3. Multi-level SLA

This agreement is customized according to the needs of the end-user company. It allows the user to integrate several conditions into the same system to create a more convenient service. This type of SLA can be divided into the following subcategories:

- **Corporate level:** This SLA does not require frequent updates since its issues are typically unchanging. It includes a comprehensive discussion of all the relevant aspects of the agreement and applies to all customers in the end-user organization.
- **Customer level:** This contract discusses all service issuesthat are associated with a specific group of customers. However, it does not take into consideration the type of user services.
  For example, when an organization requests that the security level in one of its departments is strengthened. In this situation, the entire company is secured by one security agency but requires that one of its customers is more secure for specific reasons.
- **Service level:** In this agreement, all aspects attributed to a particular service regarding a customer group are included.

Few Service Level Agreements are enforceable as contracts, but mostly are agreements or contracts which are more along the lines of an Operating Level Agreement (OLA) and may not have the restriction of law.

Service Level Agreements usually specify **some parameters** which are mentioned below:

1. Availability of the Service (uptime)
2. Latency or the response time
3. Service components reliability
4. Each party accountability
5. Warranties

In any case, if a cloud service provider fails to meet the stated targets of minimums then the provider has to pay the penalty to the cloud service consumer as per the agreement. So, Service Level Agreements are like insurance policies in which the corporation has to pay as per the agreements if any casualty occurs.

The Service Level Agreement includes:

- Detailed service overview
- Speed of service delivery
- Plan for performance monitoring
- Description of the reporting procedure
- List of penalties that will be applied in case of agreement violations
- Constraints

Components of SLA

An SLA highlights what the client and the service provider want to achieve with their cooperation and outlines the obligations of the participants, the expected performance level, and the results of cooperation.

An SLA usually has a defined duration time that is provided in the document. The services that the provider agrees to deliver are often described in detail to avoid misunderstanding, including procedures of performance monitoring, assessment, and troubleshooting. Here are the following components necessary for a good agreement:

- **Document overview:** This first section sets forth the basics of the agreement, including the parties involved, the start date, and a general introduction of the services provided.

- **Strategic goals:** Description of the agreed purpose and objectives.
- **Description of services:** The SLA needs detailed descriptions of every service offered under all possible circumstances, including the turnaround times. Service definitions should include how the services are delivered, whether maintenance service is offered, what the hours of operation are, where dependencies exist, an outline of the processes, and a list of all technology and applications used.
- **Exclusions:**Specific services that are not offered should also be clearly defined to avoid confusion and eliminate room for assumptions from other parties.
- **Service performance:** Performance measurement metrics and performance levels are defined. The client and service provider should agree on a list of all the metrics they will use to measure the provider's service levels.
- **Redressing:** Compensation or payment should be defined if a provider cannot properly fulfill their SLA.
- **Stakeholders:**Clearly defines the parties involved in the agreement and establishes their responsibilities.
- **Security:** All security measures that the service provider will take are defined. Typically, this includes the drafting and consensus on antipoaching, IT security, and nondisclosure agreements.
- **Risk management and disaster recovery:**Risk management processes and a disaster recovery plan are established and communicated.
- **Service tracking and reporting:** This section defines the reporting structure, tracking intervals, and stakeholdersinvolved in the agreement.
- **Periodic review and change processes.** The SLA and all established key performance indicators (KPIs) should be regularly reviewed. This process is defined as well as the appropriate process for making changes.
- **Termination process.** The SLA should define the circumstances under which the agreement can be terminated or will expire. The notice period from either side should also be established.
- Finally, all stakeholders and authorized participants from both parties must sign the document to approve every detail and process.

Common Metrics of SLA

Service-level agreements can contain numerous service-performance metrics with corresponding service-level objectives. A common case in IT-service management is a call center or service desk. Metrics commonly agreed to in these cases include:

- **Abandonment Rate**: Percentage of calls abandoned while waiting to be answered.
- **ASA**(Average Speed to Answer): Average time (usually in seconds) it takes for a call to be answered by the service desk.
- **Resolution time:**The time it takes for an issue to be resolved once logged by the service provider.
- **Error rate:**The percentage of errors in a service, such as coding errors and missed deadlines.
- **TSF**(Time Service Factor): Percentage of calls answered within a definite timeframe, e.g., 80% in 20 seconds.
- **FCR**(First-Call Resolution): A metric that measures a contact center's ability for its agents to resolve a customer's inquiry or problem on the first call or contact.
- **TAT**(Turn-Around-Time): Time is taken to complete a particular task.
- **TRT**(Total Resolution Time): Total time is taken to complete a particular task.
- **MTTR**(Mean Time To Recover): Time is taken to recover after an outage of service.
- **Security:**The number of undisclosed vulnerabilities, for example. If an incident occurs, service providers should demonstrate that they've taken preventive measures.

Uptime is also a common metric used for data services such as shared hosting, virtual private servers, and dedicated servers. Standard agreements include the percentage of network uptime, power uptime, number of scheduled maintenance windows, etc. Many SLAs track to the ITIL specifications when applied to IT services.

Metrics for Monitoring and Auditing

- **Throughput** – How quickly the service responds
- **Availability** – Represented as a percentage of uptime for a service in a given observation period.
- **Reliability** – How often the service is available
- **Load balancing** – When elasticity kicks in (new VMs are booted or terminated, for example)
- **Durability** – How likely the data is to be lost
- **Elasticity** – The ability for a given resource to grow infinitely, with limits (the maximum amount of storage or bandwidth, for example) clearly stated
- **Linearity** – How a system performs as the load increases

Metrics for Monitoring and Auditing *(Contd...)*

- **Agility** – How quickly the provider responds as the consumer's resource load scales up and down
- **Automation** – What percentage of requests to the provider are handled without any human interaction
- **Customer service response times** – How quickly the provider responds to a service request. This refers to the human interactions required when something goes wrong with the on-demand, self-service aspects of the cloud.
- **Service-level violation rate** – Expressed as the mean rate of SLA violation due to infringements of the agreed warranty levels.
- **Transaction time** – Time that has elapsed from when a service is invoked till the completion of the transaction, including the delays.
- **Resolution time** – Time period between detection of a service problem and its resolution

Types of SLA Penalties

A natural reply to any violation is a penalty. An SLA penalty depends on the industry and business. Here are the two most common SLA penalty types.

Types of SLA Penalties
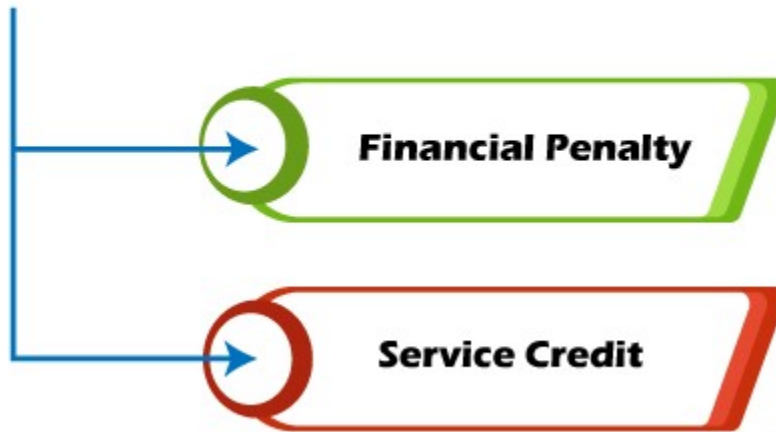
1. Financial penalty

This kind of penalty requires a vendor to pay the customer compensation of damages equal to the one written in the agreement. The amount will depend on the extent of a violation and damage and may not fully reimburse what a customer paid for the eCommerce service or eCommerce support.

- **License extension or support:**It requires the vendor to extend the license term or offer additional customer support without charge. This could include development and maintenance.

2. Service credit

In this case, a service provider will have to provide a customer with complimentary services for a specific time. To avoid any confusion or misunderstanding between the two parties in SLA violation, such penalties must be clearly articulated in the agreement. Otherwise, they won't be legitimate.

- **Service availability:**It includes factors such as network uptime, data center resources, and database availability. Penalties should be added as deterrents against service downtime, which could negatively affect the business.
- **Service quality:**It involves performance guarantee, the number of errors allowed in a product or service, process gaps, and other issues that relate to quality.

These penalties must be specified in the language of the SLA, or they won't be enforceable. In addition, some customers may not think the service credit or license extension penalties are adequate compensation. They may question the value of continuing to receive a vendor's services that cannot meet its quality levels.

Consequently, it may be worth considering a combination of penalties and including an incentive, such as a monetary bonus, for more than satisfactory work.

Revising and Changing an SLA

Since business requirements are subject to change, it's important to revise an SLA regularly. It will help to always keep the agreement in line with the business's service level objectives. The SLA should be revised when changes of the following occur:

- A company's requirements
- Workload volume
- Customer's needs
- Processes and tools

The contract should have a detailed plan for its modification, including change frequency, change procedures, and change log.

1. SLA Calculation

SLA assessment and calculation determine a level of compliance with the agreement. There are many tools for SLA calculation available on the internet.

2. SLA uptime

Uptime is the amount of time the service is available. Depending on the type of service, a vendor should provide minimum uptime relevant to the average customer's demand. Usually, a high uptime is critical for websites, online services, or web-based providers as their business relies on its accessibility.

3. Incident and SLA violations

This calculation helps determine the extent of an SLA breach and the penalty level foreseen by the contract. The tools usually calculate a downtime period during which service wasn't available, compare it to SLA terms and identify the extent of the violation.

4. SLA credit

If a service provider fails to meet the customer's expectations outlined in the SLA, a service credit or other type of penalty must be given as a form of compensation. A percentage of credit depends directly on the downtime period, which exceeded its norm indicated in a contract.

Service Level Management

Service level management is the process of managing SLAs that helps companies to define, document, monitor, measure, report, and review the performance of the provided services. The professional SLA management services should include:

- Setting realistic conditions that a service provider can ensure.

- Meeting the needs and requirements of the clients.
- Establishing the right metrics for evaluating the performance of the services.
- Ensuring compliance with the terms and conditions agreed with the clients.
- Avoiding any violations ofSLA terms and conditions.

An SLA is a preventive means to establish a transparent relationship between both parties involved and build relationships in the cooperation. Such a document is fundamental to a successful collaboration between a client and a service provider.

Example!

Example Cloud SLAs

Cloud Provider	Service	Type of Delivery Model	Service Level Agreement Guarantees
Amazon	EC2	IaaS	Availability (99.95%) with the following definitions : Service Year : 365 days of the year, Annual Percentage Uptime, Region Unavailability : no external connectivity during a five minute period, Eligible Credit Period, Service Credit
	S3	Storage-as-a-Service	Availability (99.9%) with the following definitions: Error Rate, Monthly Uptime Percentage, Service Credit
	SimpleDB	Database-as-a-Service	No specific SLA is defined and the agreement does not guarantee availability
Salesforce	CRM	PaaS	No SLA guarantees for the service provided
Google	Google App Engine	PaaS	Availability (99.9%) with the following definitions : Error Rate, Error Request, Monthly Uptime Percentage, Scheduled Maintenance, Service Credits, and SLA exclusions

Example Cloud SLAs (contd...)

Cloud Provider	Service	Type of Delivery Model	Service Level Agreement Guarantees
Microsoft	Microsoft Azure Compute	IaaS/PaaS	Availability (99.95%) with the following definitions : Monthly Connectivity Uptime Service Level, Monthly Role Instance Uptime Service Level, Service Credits, and SLA exclusions
	Microsoft Azure Storage	Storage-as-a-Service	Availability (99.9%) with the following definitions: Error Rate, Monthly Uptime Percentage, Total Storage Transactions, Failed Storage Transactions, Service Credit, and SLA exclusions
Zoho suite	Zoho mail, Zoho CRM, Zoho books	SaaS	Allows the user to customize the service level agreement guarantees based on : Resolution Time, Business Hours & Support Plans, and Escalation

Example Cloud SLAs *(contd...)*

Cloud Provider	Service	Type of Cloud Delivery Model	Service Level Agreement Guarantees
Rackspace	Cloud Server	IaaS	Availability regarding the following: Internal Network (100%), Data Center Infrastructure (100%), Load balancers (99.9%) Performance related to service degradation: Server migration, notified 24 hours in advance, and is completed in 3 hours (maximum) Recovery Time: In case of failure, guarantee of restoration/recovery in 1 hour after the problem is identified.
Terremark	vCloud Express	IaaS	Monthly Uptime Percentage (100%) with the following definitions: Service Credit, Credit Request and Payment Procedure, and SLA exclusions

Example Cloud SLAs *(contd...)*

Cloud Provider	Service	Type of Cloud Delivery Model	Service Level Agreement Guarantees
Nirvanix	Public, Private, Hybrid Cloud Storage	Storage-as-a-Service	Monthly Availability Percentage (99.9%) with the following definitions: Service Availability, Service Credits, Data Replication Policy, Credit Request Procedure, and SLA Exclusions