Service Level Agreements

By Aaron Johns

Service-level agreement (SLA)

* A service-level agreement (SLA) is a contract between a service provider and its customers that documents what services the provider will furnish and defines the service standards the provider is obligated to meet.
* A service-level commitment (SLC) is a broader and more generalized form of an SLA.
* The two differ because an SLA is bidirectional and involves two teams. In contrast, an SLC is a single-directional obligation that establishes what a team can guarantee its customers at any given time.
* Service providers need SLAs to help them manage customer expectations and define the severity levels and circumstances under which they are not liable for outages or performance issues. Customers can also benefit from SLAs because the contract describes the performance characteristics of the service -- which can be compared with other vendors' SLAs -- and sets forth the means for redressing service issues.


SLA checklist
Statement of objectives
Scope of services to be covered
Service provider responsibilities
Customer responsibilities
Performance metrics (response time, resolution time, etc.)
Penalties for contract breach/exclusions

Key components of an SLA

* **Agreement 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.

* **Description of services.**

  The SLA needs detailed descriptions of every service offered, under all possible circumstances, with the turnaround times included. 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 service levels of the provider.

* **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 will be taken by the service provider 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 clearly communicated.

- **Service tracking and reporting.**

  This section defines the reporting structure, tracking intervals and stakeholders involved 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.

- **Signatures.**

  Finally, all stakeholders and authorized participants from both parties must sign the document to show their approval of every detail and process.

Three types of SLAs

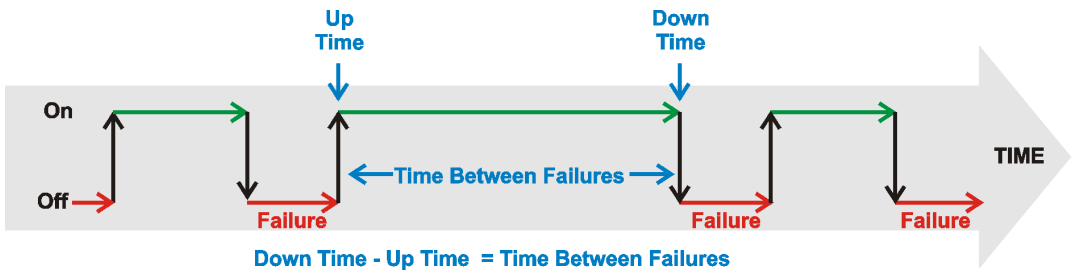
* There are three basic types of SLAs: customer, internal and multilevel service-level agreements.
    * A **customer service-level agreement** is between a service provider and its external customers. It is sometimes called an external service agreement.
    * An **internal SLA** is between an organization and its internal customer -- this could be another organization, department or site.
    * A **multilevel SLA** will divide the agreement into various levels that are specific to a series of customers using the service. For example, a software as a service (SaaS) provider might offer basic services and support to all customers using a product, but they could also offer different price ranges when buying the product that dictates different service levels. These different levels of service will be layered into the multilevel SLA.

Some Important Terminologies

Mean time between failures (MTBF)

* This prediction uses previous observations and data to determine the average time between failures. MTBF predictions are often used to designate overall failure rates, for both repairable and replaceable/non-repairable products.
* Here is the simplest equation for mean time between failure:

MTBF=total operational uptime between failures / number of failures



Example

Assume a manufacturer has recorded the following data points between product failures for one of its network switch models

FAILURE NUMBER	RECORDED OPERATIONAL UPTIME PRIOR TO FAILURE (IN HOURS)
1	10,000
2	9,500
3	11,000
4	9,000
Total	39,500

Substituting the values in the formula

MTBF=(total operational uptime between failures / number of failures)

we get

MTBF = ((10,000 + 9,500 + 11,000 + 9,000)/4) = 9,875

That means the manufacturer has approximately 9,875 hours of uptime on this network switch before it experiences a failure.

MTTR (mean time to repair)

* MTTR (mean time to repair) is the average time required to fix a failed component or device and return it to production status.
* Mean time to repair includes the time it takes to find out about the failure, diagnose the problem and repair it. MTTR is a basic measure of how maintainable an organization's equipment is and, ultimately, is a reflection of how efficiently an organization can fix a problem.
* Data storage professionals can use MTTR as a performance metric to evaluate how efficiently they are using their data storage resources. Once the mean time to repair is known, it can be used to modify and improve an organization's processes to reduce that figure and decrease the amount of lost productivity.
* If the mean time to repair is already low for a device, it indicates that a component can be repaired quickly and efficiently.

* MTTR can be calculated by dividing the total time required for maintenance -- downtime -- by the total number of repairs within a specific time frame.

* If the total time required to fix the issue is 120 minutes, and four breakdowns caused that downtime, an organization can conclude that a breakdown will take approximately 30 minutes to repair.

* The goal of any organization should be to decrease MTTR and increase the MTBF of a system. Generally, mean time to repair indicates efficiency in correcting processes, and mean time between failures indicates the reliability of a system.

Mean time to failure

* Mean time to failure (MTTF) is a maintenance metric that measures the average amount of time a non-repairable asset operates before it fails. Because MTTF is relevant only for assets and equipment that cannot or should not be repaired, MTTF can also be thought of as the average lifespan of an asset.

* To calculate MTTF, divide the total number of hours of operation by the total number of assets in use.

MTTF = Total hours of operation ÷ Total assets in use

Let’s assume we tested three identical water pumps in a datacenter until all of them failed. The first pump system failed after eight hours, the second one failed at ten hours, and the third failed at twelve hours.

Your MTTF calculation would look like this:

MTTF = Total hours of operation / Total assets in use
MTTF = (8 + 10 + 12) / 3
MTTF = 10 hours

This would lead us to the conclusion that this particular type and model of the pump will need to be replaced, on average, every 10 hours.

High Availability

* High availability (HA) is a characteristic of a system which aims to ensure an agreed level of operational performance, usually uptime, for a higher than normal period.
* Modernization has resulted in an increased reliance on these systems. For example, hospitals and data centers require high availability of their systems to perform routine daily activities.
* Availability refers to the ability of the user community to obtain a service or good, access the system, whether to submit new work, update or alter existing work, or collect the results of previous work. If a user cannot access the system, it is – from the user's point of view – unavailable.
* Generally, the term downtime is used to refer to periods when a system is unavailable.

Availability %	Downtime per year	Downtime per quarter	Downtime per month	Downtime per week	Downtime per day
90% ("one nine")	**36.53 days**	**9.13 days**	**73.05 hours**	**16.80 hours**	**2.40 hours**
95% ("one and a half nines")	18.26 days	4.56 days	36.53 hours	8.40 hours	1.20 hours
97%	10.96 days	2.74 days	21.92 hours	5.04 hours	43.20 minutes
98%	7.31 days	43.86 hours	14.61 hours	3.36 hours	28.80 minutes
99% ("two nines")	**3.65 days**	**21.9 hours**	**7.31 hours**	**1.68 hours**	**14.40 minutes**
99.5% ("two and a half nines")	1.83 days	10.98 hours	3.65 hours	50.40 minutes	7.20 minutes
99.8%	17.53 hours	4.38 hours	87.66 minutes	20.16 minutes	2.88 minutes
99.9% ("three nines")	**8.77 hours**	**2.19 hours**	**43.83 minutes**	**10.08 minutes**	**1.44 minutes**
99.95% ("three and a half nines")	4.38 hours	65.7 minutes	21.92 minutes	5.04 minutes	43.20 seconds
99.99% ("four nines")	**52.60 minutes**	**13.15 minutes**	**4.38 minutes**	**1.01 minutes**	**8.64 seconds**
99.995% ("four and a half nines")	26.30 minutes	6.57 minutes	2.19 minutes	30.24 seconds	4.32 seconds
99.999% ("five nines")	**5.26 minutes**	**1.31 minutes**	**26.30 seconds**	**6.05 seconds**	**864.00 milliseconds**
99.9999% ("six nines")	**31.56 seconds**	**7.89 seconds**	**2.63 seconds**	**604.80 milliseconds**	**86.40 milliseconds**
99.99999% ("seven nines")	**3.16 seconds**	**0.79 seconds**	**262.98 milliseconds**	**60.48 milliseconds**	**8.64 milliseconds**
99.999999% ("eight nines")	**315.58 milliseconds**	**78.89 milliseconds**	**26.30 milliseconds**	**6.05 milliseconds**	**864.00 microseconds**
99.9999999% ("nine nines")	**31.56 milliseconds**	**7.89 milliseconds**	**2.63 milliseconds**	**604.80 microseconds**	**86.40 microseconds**

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; and
  - efficiency.

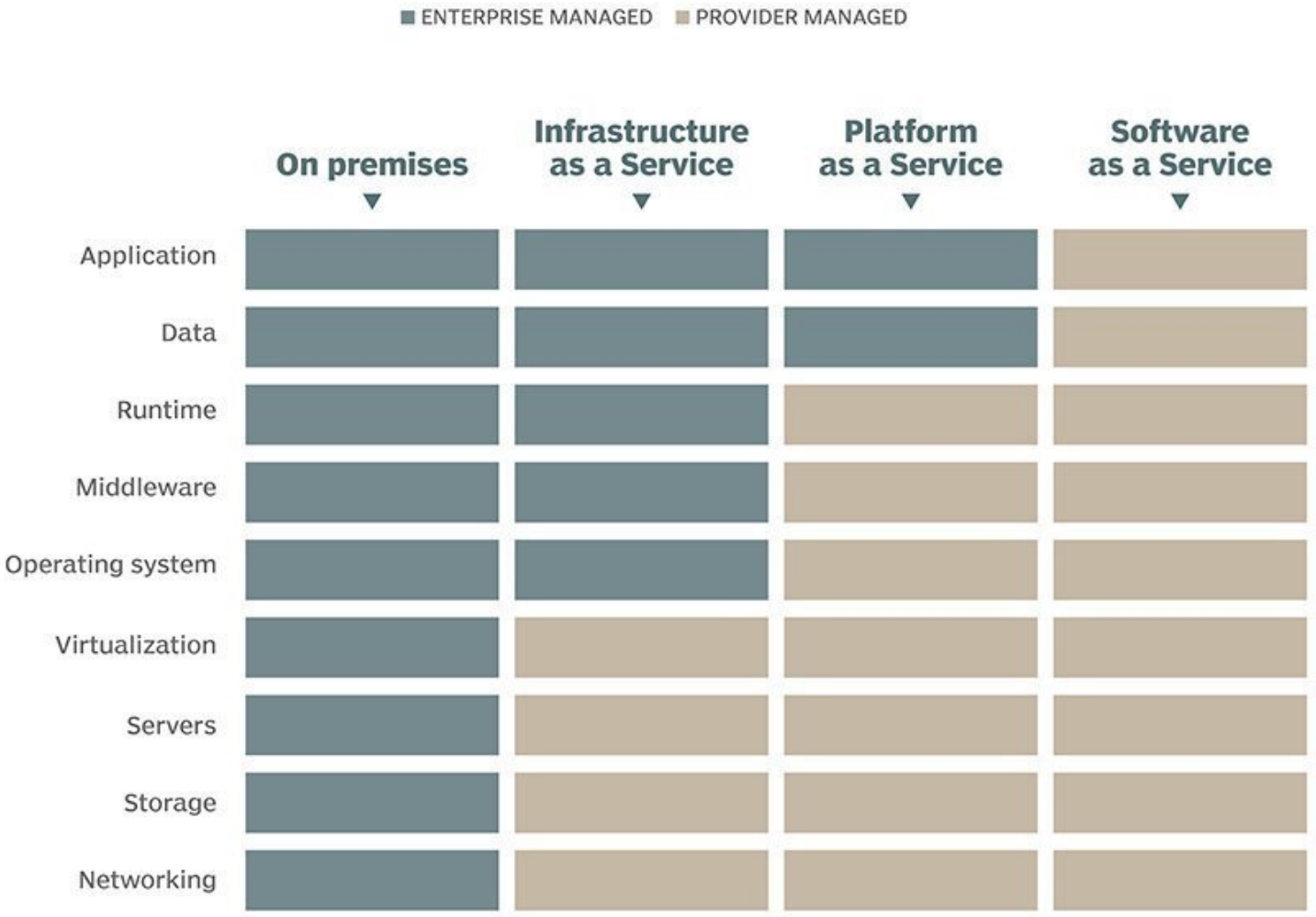
- There are three principles of systems design in reliability engineering which can help achieve high availability.
  - **Elimination of single points of failure.** This means adding or building redundancy into the system so that failure of a component does not mean failure of the entire system.
  - **Reliable crossover.** In redundant systems, the crossover point itself tends to become a single point of failure. Reliable systems must provide for reliable crossover.
  - **Detection of failures as they occur.** If the two principles above are observed, then a user may never see a failure – but the maintenance activity must.

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.
* A cloud SLA's role is essentially the same as any contract -- it is a blueprint that governs the relationship between a customer and provider.
* 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.

Architecting uptime

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



Links for SLA's of various cloud providers

* Microsoft Azure SLA
* AWS SLA
* Google Cloud Platform SLA

THE END