The fundamental knowledge of System Design — (4) — System Availability

*System Availability = **Availability = Uptime ÷ (Uptime + downtime)***

It is the fourth series of the fundamentals knowledge of system design. You can read my previous articles.

The fundamental knowledge of System Design — (1)

Today, I will share the fundamental knowledge of system design.

medium.com

The fundamental knowledge of System Design — (2)

Please clap and share this article if you like it.

medium.com

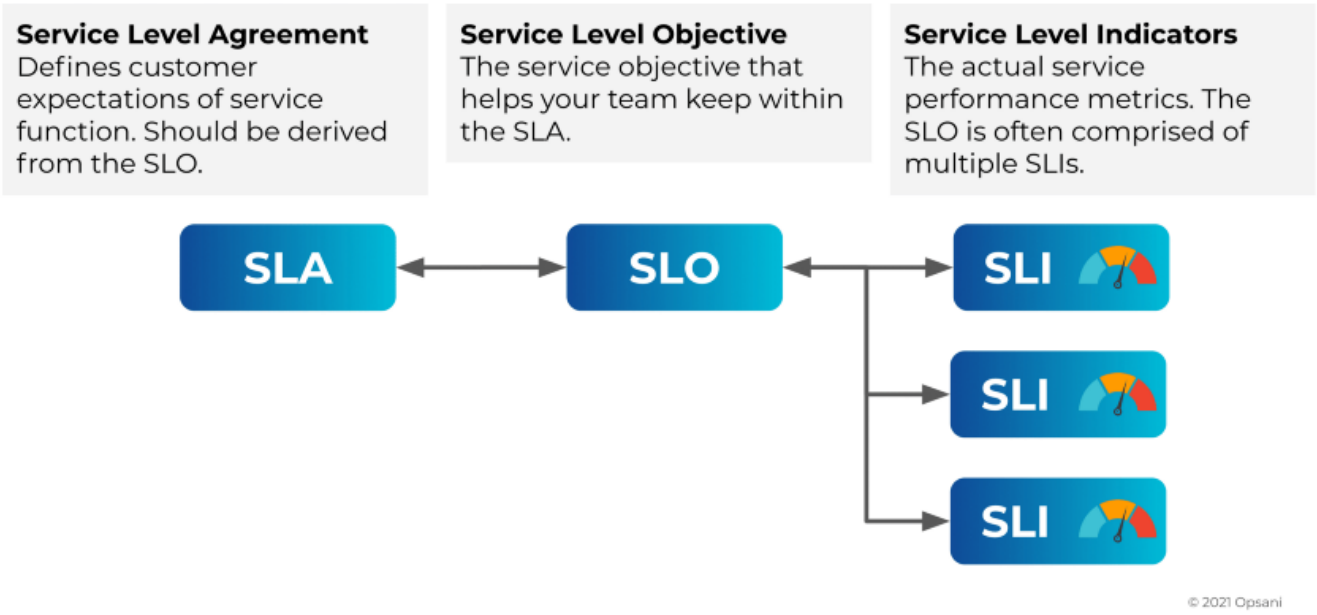
The fundamental knowledge of System Design — (3)

Servers are the core of today's computational world. Server performance depends on throughput and latency. Generally…

medium.com

	SLI	SLO	SLA
Definition	A quantifiable measure of reliability	A target reliability level of a service.	A legal contract that if breached, will have financial penalties.
Example	The ratio of valid requests loaded in < 400 ms	99% of requests served in < 400 ms over a 28-day rolling window	Monthly average round-trip transmission time of 500 ms or less for 98% of requests
Who Sets it?	SREs (site reliability engineers)	Product owner, SRE and Ops team, developers, and customers.	Business Development and legal team alongside IT and DevOps teams.

https://www.blameless.com/sre/service-level-objectives



https://opsani.com/resources/site-reliability-engineering-service-level-agreement-terms-explained-sla-slo-sli/

SLI = Service Level Indicator

It is the most important metric for business.

- Uptime of the service
- Number of transactions
- Latency
- Error rate
- Throughput
- Response time
- Durability

SLO = Service Level Objective

It is built around SLI. It refers to a target value or target range of service level. Usually a percentage and tied to a time frame.

90% (1 nine of Uptime) = 10% downtime, which means 3 out of the last 30 days

99% (2 nines of uptime) = 1% downtime, or 7.2 hours of downtime in the last 30 days

99.9% (3 nines of uptime) =0.1% downtime, or 43.2 minutes of downtime in the last 30 days

SLA = Service Level Agreement

An agreement is issued by enterprises to customers.

- Refund service fee
- Provide free service for a period of time

Availability %	Downtime per year	Downtime per month*	Downtime per week
90% ("one nine")	36.5 days	72 hours	16.8 hours
95%	18.25 days	36 hours	8.4 hours
98%	7.30 days	14.4 hours	3.36 hours
99% ("two nines")	3.65 days	7.20 hours	1.68 hours
99.5%	1.83 days	3.60 hours	50.4 minutes
99.8%	17.52 hours	86.23 minutes	20.16 minutes
99.9% ("three nines")	8.76 hours	43.2 minutes	10.1 minutes
99.95%	4.38 hours	21.56 minutes	5.04 minutes
99.99% ("four nines")	52.56 minutes	4.32 minutes	1.01 minutes
99.999% ("five nines")	5.26 minutes	25.9 seconds	6.05 seconds
99.9999% ("six nines")	31.5 seconds	2.59 seconds	0.605 seconds

https://ophir.wordpress.com/2011/01/31/does-sla-really-mean-anything/

Case Study

Assume I have a website **http://xxx.com.** From the launch on January 1, 2022, to March 15, 2022, the requested data is as follows:

- The total number of requests from the whole of January was 500, the number of error responses was 20
- The total number of requests from the whole of February was 600, the number of error responses was 10, and the downtime was 10 minutes
- The total number of requests from the current March was 400, and the number of error responses was 15.

Then what are the SLI, SLO, and SLA I calculated?

SLI, 1 — (20+10+15)/(500+600+400) = 97%

SLO, 1 — (10/(74*24*60))=99.991%

SLA, If the service provider cannot meet the term of the agreement that the SLO does not reach 99.999%, how much is the compensation according to the signed SLA agreement.

The application

It is **the term of the agreement** under which Google has agreed to provide Google Cloud Platform to customers.

Cloud Spanner Service Level Agreement (SLA)

During the term of the agreement under which Google has agreed to provide Google Cloud Platform to Customer (as applicable, the "Agreement"), the Covered Service will provide a Monthly Uptime Percentage to Customer as follows (the "Service Level Objective" or "SLO"):

Covered Service	Monthly Uptime Percentage
Cloud Spanner - Multi-Regional Instance	>= 99.999%
Cloud Spanner - Regional Instance	>= 99.99%

If Google does not meet the SLO, and if Customer meets its obligations under this SLA, Customer will be eligible to receive the Financial Credits described below. This SLA states Customer's sole and exclusive remedy for any failure by Google to meet the SLO. Capitalized terms used in this SLA, but not defined in this SLA, have the meaning given to them in the Agreement. If the Agreement authorizes the resale or supply of Google Cloud Platform under a Google Cloud partner or reseller program, then all references to Customer in this SLA mean Partner or Reseller (as applicable), and any Financial Credit(s) will only apply for impacted Partner or Reseller order(s) under the Agreement.

https://cloud.google.com/spanner/sla

Ideally, the SLI should directly measure a specific quality of service. But, in many cases, the direct measurement may be very difficult to be observed and obtained. So, only some kind of indicator can be used. Latency is the most direct monitoring indicator. Durability is also an important metric for the data storage systems to monitor how long data can be kept intact. While 100% availability is impossible to achieve, a near 100% availability metric is an achievable goal. The operations expert often uses the number 9 to describe availability. For example, 99% availability is called “2 nines” and 99.99% availability is called “4 nines”.

The current availability indicator for Google cloud computing services is “3.5 nines” — 99.95% availability.

Choosing a target SLO is not a purely technical activity, as there are also product and business-level decisions involved here. The choice of SLI and SLO should directly reflect the product and business-level decisions. Site reliability engineers (SREs) should discuss and provide advice on feasibility and risk. That’s why it is important to understand the various indicators and limitations of the system. Only enough SLOs should be selected to cover system properties.

SLI and SLO are very useful when making decisions about system operation and maintenance.

1. Monitor and measure the SLI of the system
2. Compare SLI and SLO to decide if action is required
3. If an action needs to be performed, then it is up to decide what exactly needs to be performed in order to meet the goal
4. perform these operations

For example, if in step 2, the request latency is rising, the SLO will be exceeded within a few hours with no operations. The third step will test whether the server is not enough CPU resources, and add some CPU to spread the load. Without SLO, we don’t know if (or when) the action needs to be performed.

SLA requires the business and legal departments to choose the appropriate consequence clause. The role of the site reliability engineer is to help the business and legal departments understand the probability and difficulty of meeting the SLA’s SLO. Google guarantees that the service’s annual availability time is≥99.99%. Also, Google guarantees the first response within 1 hour of the user’s request for technical support, including phone calls, emails, etc. The term also comes with a lot of reward and compensation details.

Monthly Uptime Percentage	Percentage of monthly bill for Cloud Spanner Multi-Regional which does not meet SLO that will be credited to future monthly Customer bills
99.0% – < 99.999%	10%
95.0% – < 99.0%	25%
< 95.0%	50%

- **"Financial Credit"** means the following for Cloud Spanner Regional instances:

Monthly Uptime Percentage	Percentage of monthly bill for Cloud Spanner Regional which does not meet SLO that will be credited to future monthly Customer bills
99.0% – < 99.99%	10%
95.0% – < 99.0%	25%
< 95.0%	50%

- **"Monthly Uptime Percentage"** means total number of minutes in a month, minus the number of minutes of Downtime suffered from all Downtime Periods in a month, divided by the total number of minutes in a month.
- **"Valid Requests"** are requests that conform to the Documentation, and that would normally result in a non-error response.