Service Quality Metrics and SLAs

Chap 16, from Thomas Erl BooK

- Service-level agreements (SLAs) are a focal point of negotiations, contract terms, legal obligations, and runtime metrics and measurements.
- SLAs formalize the guarantees put forth by cloud providers, and correspondingly influence or determine the pricing models and payment terms.

It is therefore crucial for SLAs and related service quality metrics to be understood and aligned in support of the cloud consumer's business requirements, while also ensuring that the guarantees can, in fact, be realistically fulfilled consistently and reliably by the cloud provider.

Service Quality Metrics 463

SLAs issued by cloud providers are human-readable documents that describe quality-of-service (QoS) features, guarantees, and limitations of one or more cloud-based IT resources.

SLAs use service quality metrics to express measurable QoS characteristics.

For example:

- ***Availability*** – up-time, outages, service duration
- ***Reliability*** – minimum time between failures, guaranteed rate of successful responses
- ***Performance*** – capacity, response time, and delivery time guarantees
- ***Scalability*** – capacity fluctuation and responsiveness guarantees
- ***Resiliency*** – mean-time to switchover and recovery

SLA management systems use these metrics to perform periodic measurements that verify compliance with SLA guarantees, in addition to collecting SLA-related data for various types of statistical analyses.

Each service quality metric is ideally defined using the following characteristics:

- ***Quantifiable*** – The unit of measure is clearly set, absolute, and appropriate so that the metric can be based on quantitative measurements.
- ***Repeatable*** – The methods of measuring the metric need to yield identical results when repeated under identical conditions.
- ***Comparable*** – The units of measure used by a metric need to be standardized and comparable. For example, a service quality metric cannot measure smaller quantities of data in bits and larger quantities in bytes.
- ***Easily Obtainable*** – The metric needs to be based on a non-proprietary, common form of measurement that can be easily obtained and understood by cloud consumers.

Examples of the metrics:-

Availability Rate Metric

The overall availability of an IT resource is usually expressed as a percentage of up-time. For example, an IT resource that is always available will have an up- time of 100%.

- Description – percentage of service up-time
- Measurement – total up-time / total time
- Frequency – weekly, monthly, yearly
- Cloud Delivery Model – IaaS, PaaS, SaaS
- Example – minimum 99.5% up-time

Availability rates are calculated cumulatively, meaning that unavailability periods are combined in order to compute the total downtime.

Availability (%)	Downtime/Week (Seconds)	Downtime/Month (Seconds)	Downtime/Year (Seconds)
99.5	3024	216	158112
99.8	1210	5174	63072
99.9	606	2592	31536
99.95	302	1294	15768
99.99	60.6	259.2	3154
99.999	6.05	25.9	316.6
99.9999	0.605	2.59	31.5

Outage Duration Metric

This service quality metric is used to define both maximum and average continuous outage service-level targets.

- ***Description*** – duration of a single outage
- ***Measurement*** – date/time of outage end – date/time of outage start
- ***Frequency*** – per event
- ***Cloud Delivery Model*** – IaaS, PaaS, SaaS
- ***Example*** – 1 hour maximum, 15 minute average

Service Reliability Metrics

- A characteristic closely related to availability, reliability is the probability that an IT resource can perform its intended function under pre-defined conditions without experiencing failure.
- Reliability focuses on how often the service performs as expected, which requires the service to remain in an operational and available state.
- Certain reliability metrics only consider runtime errors and exception conditions as failures, which are commonly measured only when the IT resource is available.

Mean-Time Between Failures (MTBF) Metric

- ***Description*** – expected time between consecutive service failures
- ***Measurement*** – Σ, normal operational period duration / number of failures
- ***Frequency*** – monthly, yearly
- ***Cloud Delivery Model*** – IaaS, PaaS
- ***Example*** – 90 day average

Reliability Rate Metric

Overall reliability is more complicated to measure and is usually defined by a reliability rate that represents the percentage of successful service outcomes.

This metric measures the effects of non-fatal errors and failures that occur during up-time periods. For example, an IT resource's reliability is 100% if it has performed as expected every time it is invoked, but only 80% if it fails to perform every fifth time.

- Description – percentage of successful service outcomes under pre-defined conditions
- Measurement – total number of successful responses / total number of requests

Frequency – weekly, monthly, yearly

Cloud Delivery Model – SaaS

Example – minimum 99.5%

Service Performance Metrics

- Service performance refers to the ability on an IT resource to carry out its functions within expected parameters.
- This quality is measured using service capacity metrics, each of which focuses on a related measurable characteristic of IT resource capacity.

- A set of common performance capacity metrics is provided in this section. Note that different metrics may apply, depending on the type of IT resource being measured.

Network Capacity Metric

- Description – measurable characteristics of network capacity
- Measurement – bandwidth / throughput in bits per second
- Frequency – continuous
- Cloud Delivery Model – IaaS, PaaS, SaaS
- Example – 10 MB per second

Storage Device Capacity Metric

- Description – measurable characteristics of storage device capacity
- Measurement – storage size in GB
- Frequency – continuous
- Cloud Delivery Model – IaaS, PaaS, SaaS
- Example – 80 GB of storage

Server Capacity Metric

- Description – measurable characteristics of server capacity
- Measurement – number of CPUs, CPU frequency in GHz, RAM size in GB, storage size in GB
- Frequency – continuous
- Cloud Delivery Model – IaaS, PaaS
- Example – 1 core at 1.7 GHz, 16 GB of RAM, 80 GB of storage

Web Application Capacity Metric

- Description – measurable characteristics of Web application capacity
- Measurement – rate of requests per minute
- Frequency – continuous
- Cloud Delivery Model – SaaS
- Example – maximum 100,000 requests per minute

Instance Starting Time Metric

- Description – length of time required to initialize a new instance
- Measurement – date/time of instance up – date/time of start request
- Frequency – per event
- Cloud Delivery Model – IaaS, PaaS
- Example – 5 minute maximum, 3 minute average

Response Time Metric

- Description – time required to perform synchronous operation
- Measurement – (date/time of request – date/time of response) / total number of requests
- Frequency – daily, weekly, monthly
- Cloud Delivery Model – SaaS
- Example – 5 millisecond average

Completion Time Metric

- Description – time required to complete an asynchronous task
- Measurement – (date of request – date of response) / total number of requests
- Frequency – daily, weekly, monthly
- Cloud Delivery Model – PaaS, SaaS
- Example – 1 second average

Service Scalability Metrics

Service scalability metrics are related to IT resource elasticity capacity, which is related to the maximum capacity that an IT resource can achieve, as well as measurements of its ability to adapt to workload fluctuations.

For example, a server can be scaled up to a maximum of 128 CPU cores and 512 GB of RAM, or scaled out to a maximum of 16 load-balanced replicated instances.

The following metrics help determine whether dynamic service demands will be met proactively or reactively, as well as the impacts of manual or automated IT resource allocation processes.

Storage Scalability (Horizontal) Metric

- ***Description*** – permissible storage device capacity changes in response to increased workloads
- ***Measurement*** – storage size in GB
- ***Frequency*** – continuous
- ***Cloud Delivery Model*** – IaaS, PaaS, SaaS
- ***Example – 1,000 GB maximum (automated scaling)***

Server Scalability (Horizontal) Metric

- ***Description*** – permissible server capacity changes in response to increased workloads
- ***Measurement*** – number of virtual servers in resource pool
- ***Frequency*** – continuous
- ***Cloud Delivery Model*** – IaaS, PaaS
- Example – 1 virtual server minimum, 10 virtual server maximum (automated scaling)

Server Scalability (Vertical) Metric

- ***Description*** – permissible server capacity fluctuations in response to workload fluctuations
- ***Measurement*** – number of CPUs, RAM size in GB
- ***Frequency*** – continuous
- ***Cloud Delivery Model*** – IaaS, PaaS
- Example – 512 core maximum, 512 GB of RAM

Service Resiliency Metrics

- The ability of an IT resource to recover from operational disturbances is often measured using service resiliency metrics.
- When resiliency is described within or in relation to SLA resiliency guarantees, it is often based on redundant implementations and resource replication over different physical locations, as well as various disaster recovery systems.
- The type of cloud delivery model determines how resiliency is implemented and measured. For example, the physical locations of replicated virtual servers that are implementing resilient cloud services can be explicitly expressed in the SLAs for IaaS environments, while being implicitly expressed for the corresponding PaaS and SaaS environments.

Resiliency metrics can be applied in three different phases to address the challenges and events that can threaten the regular level of a service:

- ***Design Phase*** – Metrics that measure how prepared systems and services are to cope with challenges.
- ***Operational Phase*** – Metrics that measure the difference in service levels before, during, and after a downtime event or service outage, which are further qualified by availability, reliability, performance, and scalability metrics.
- ***Recovery Phase*** – Metrics that measure the rate at which an IT resource recovers from downtime, such as the meantime for a system to log an outage and switchover to a new virtual server.

Two common metrics related to measuring resiliency are as follows:-

Mean-Time to Switchover (MTSO) Metric

- ***Description*** – the time expected to complete a switchover from a severe failure to a replicated instance in a different geographical area
- ***Measurement*** – (date/time of switchover completion – date/time of failure)/ total number of failures
- ***Frequency*** – monthly, yearly
- ***Cloud Delivery Model*** – IaaS, PaaS, SaaS
- Example – 10 minute average

Mean-Time System Recovery (MTSR) Metric

- ***Description*** – time expected for a resilient system to perform a complete recovery from a severe failure
- ***Measurement*** – (date/time of recovery – date/time of failure) / total number of failures
- ***Frequency*** – monthly, yearly
- ***Cloud Delivery Model*** – IaaS, PaaS, SaaS
- Example – 120 minute average

Cloud Usage Cost Metrics

The following sections describe a set of usage cost metrics for calculating costs associated with cloud-based IT resource usage measurements:

- ***Network Usage*** – inbound and outbound network traffic, as well as intra-cloud network traffic.
- ***Server Usage*** – virtual server allocation (and resource reservation)
- ***Cloud Storage Device*** – storage capacity allocation
- ***Cloud Service*** – subscription duration, number of nominated users,number of transactions (of cloud services and cloud-based applications)

For each usage cost metric a description, measurement unit, and measurementfrequency is provided, along with the cloud delivery model most applicable tothe metric. Each metric is further supplemented with a brief example.

Network Usage

- Defined as the amount of data that is transferred over a network connection, network usage is typically calculated using separately measured inbound network usage traffic and outbound network usage traffic metrics in relation to cloud services or other IT resources.

Inbound Network Usage Metric

- Description – inbound network traffic
- ***Measurement*** – Σ, inbound network traffic in bytes
- ***Frequency*** – continuous and cumulative over a predefined period
- ***Cloud Delivery Model*** – IaaS, PaaS, SaaS
- Example – up to 1 GB free, $0.001/GB up to 10 TB a month

Outbound Network Usage Metric

- *Description* – outbound network traffic
- *Measurement* – Σ, outbound network traffic in bytes
- *Frequency* – continuous and cumulative over a predefined period
- *Cloud Delivery Model* – IaaS, PaaS, SaaS

Example – up to 1 GB free a month, $0.01/GB between 1 GB to 10 TB per

Month Network usage metrics can be applied to WAN traffic between IT resources of one cloud that are located in different geographical regions in order to calculate costs for synchronization, data replication, and related forms of processing.

Conversely, LAN usage and other network traffic among IT resources that reside at the same data center are typically not tracked.

Intra-Cloud WAN Usage Metric

- ***Description*** – network traffic between geographically diverse IT resources of the same cloud
- ***Measurement*** – Σ, intra-cloud WAN traffic in bytes
- ***Frequency*** – continuous and cumulative over a predefined period
- ***Cloud Delivery Model*** – IaaS, PaaS, SaaS
- Example – up to 500 MB free daily and $0.01/GB thereafter, $0.005/GB after 1 TB per month.

Many cloud providers do not charge for inbound traffic in order to encourage cloud consumers to migrate data to the cloud. Some also do not charge for WAN traffic within the same cloud.

Network-related cost metrics are determined by the following properties:

- ***Static IP Address Usage*** – IP address allocation time (if a static IP is required)
- ***Network Load-Balancing*** – the amount of load-balanced network traffic (in bytes)
- ***Virtual Firewall*** – the amount of firewall-processed network traffic (as per allocation time)

Server Usage

The allocation of virtual servers is measured using common pay-per-use metrics in IaaS and PaaS environments that are quantified by the number of virtual servers and ready-made environments. This form of server usage measurement is divided into on-demand virtual machine instance allocation and reserved virtual machine instance allocation metrics.

The former metric measures pay-per-usage fees on a short-term basis, while the latter metric calculates up-front reservation fees for using virtual servers over extended periods. The up-front reservation fee is usually used in conjunction with the discounted pay-per-usage fees.

On-Demand Virtual Machine Instance Allocation Metric

- ***Description*** – uptime of a virtual server instance
- ***Measurement*** – Σ, virtual server start date to stop date
- ***Frequency*** – continuous and cumulative over a predefined period
- ***Cloud Delivery Model*** – IaaS, PaaS
- Example – $0.10/hour small instance, $0.20/hour medium instance, $0.90/hour large instance.

Reserved Virtual Machine Instance Allocation Metric

- ***Description*** – up-front cost for reserving a virtual server instance
- ***Measurement*** – Σ, virtual server reservation start date to expiry date
- ***Frequency*** – daily, monthly, yearly
- ***Cloud Delivery Model*** – IaaS, PaaS
- Example – $55.10/small instance, $99.90/medium instance, $249.90/ large instance

Cloud Storage Device Usage (26/03/2021)

Cloud storage is generally charged by the amount of space allocated within a predefined period, as measured by the on-demand storage allocation metric.

Similar to IaaS-based cost metrics, on-demand storage allocation fees are usually based on short time increments (such as on an hourly basis).

Another common cost metric for cloud storage is I/O data transferred, which measures the amount of transferred input and output data.

On-Demand Storage Space Allocation Metric

- ***Description*** – duration and size of on-demand storage space allocation in bytes
- ***Measurement*** – Σ, date of storage release / reallocation to date of storage allocation (resets upon change in storage size)
- ***Frequency*** – continuous
- ***Cloud Delivery Model*** – IaaS, PaaS, SaaS
- Example – $0.01/GB per hour (typically expressed as GB/month)

I/O Data Transferred Metric

- ***Description*** – amount of transferred I/O data
- ***Measurement*** – Σ, I/O data in bytes
- ***Frequency*** – continuous
- ***Cloud Delivery Model*** – IaaS, PaaS
- Example – $0.10/TB

Note that some cloud providers do not charge for I/O usage for IaaS and PaaS implementations, and limit charges to storage space allocation only.

Cloud Service Usage

Cloud service usage in SaaS environments is typically measured using the following three metrics:

Application Subscription Duration Metric

- ***Description*** – duration of cloud service usage subscription
- ***Measurement*** – Σ, subscription start date to expiry date
- ***Frequency*** – daily, monthly, yearly
- ***Cloud Delivery Model*** – SaaS
- Example – $69.90 per month

Number of Nominated Users Metric

- Description – number of registered users with legitimate access
- Measurement – number of users
- Frequency – monthly, yearly
- Cloud Delivery Model – SaaS
- Example – $0.90/additional user per month

Number of Transactions Users Metric

- Description – number of transactions served by the cloud service
- Measurement – number of transactions (request-response message exchanges)
- Frequency – continuous
- Cloud Delivery Model – PaaS, SaaS
- Example – $0.05 per 1,000 transactions

Pricing Models

The pricing models used by cloud providers are defined using templates that specify unit costs for fine-grained resource usage according to usage cost metrics. Various factors can influence a pricing model, such as:

- Market competition and regulatory requirements.
- Overhead incurred during the design, development, deployment, and operation of cloud services and other IT resources.
- Opportunities to reduce expenses via IT resource sharing and data center optimization.

A pricing model can contain multiple price templates,whose formulation is determined by variables like:

- ***Cost Metrics and Associated Prices*** – These are costs that are dependent on the type of IT resource allocation (such as on-demand versus reserved allocation).
- ***Fixed and Variable Rates Definitions*** – Fixed rates are based on resource allocation and define the usage quotas included in the fixed price, while variable rates are aligned with actual resource usage.
- ***Volume Discounts*** – More IT resources are consumed as the degree of IT resource scaling progressively increases, thereby possibly qualifying a cloud consumer for higher discounts.
- ***Cost and Price Customization Options*** – This variable is associated with payment options and schedules. For example, cloud consumers may be able to choose monthly, semi-annual, or annual payment installments.

Price templates are important for cloud consumers that are appraising cloud providers and negotiating rates, since they can vary depending on the adopted cloud delivery model.

For example:-

- IaaS – Pricing is usually based on IT resource allocation and usage, which includes the amount of transferred network data, number of virtual servers, and allocated storage capacity.
- PaaS – Similar to IaaS, this model typically defines pricing for network data transferred, virtual servers, and storage. Prices are variable depending on factors such as software configurations, development tools, and licensing fees.
- SaaS – Because this model is solely concerned with application software usage, pricing is determined by the number of application modules in the subscription, the number of nominated cloud service consumers, and the number of transactions.

- Chapter V ends here