Service Quality Metrics and SLA Models

Introduction

- 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.
- SLAs set cloud consumer expectations and are integral to how organizations build business automation around the utilization of cloud-based IT resources.
- The guarantees made by a cloud provider to a cloud consumer are often carried forward, in that the same guarantees are made by the cloud consumer organization to its clients, business partners, or whomever will be relying on the services and solutions hosted by the cloud provider.

Service Quality Metrics

- 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.

Service Quality Metrics (2)

- 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.

Service Availability 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

Table 16.1. Sample availability rates measured in units of seconds.

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

Service Availability Metrics – 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 - MTBF

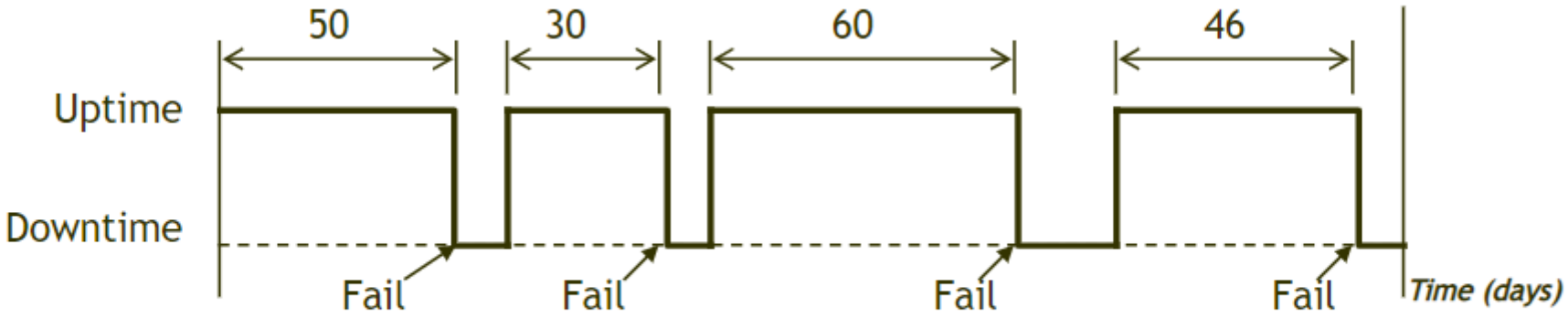
- Reliability focuses on how often the service performs as expected, which requires the service to remain in an operational and available state. Mean-Time Between Failures (MTBF) Metric include:
- Description – expected time between consecutive service failures
- Measurement – Σ, normal operational period duration / number of failures
- Frequency – monthly, yearly
- Cloud Delivery Model – IaaS, PaaS
- Example – 90 day average”

MEAN TIME BETWEEN FAILURE (MTBF)

The average time between successive failures. It is used for **repairable** systems when **failure rate** is assumed to be **constant** (random failure).

$$MTBF = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Example:



MTBF = (50+30+60+46) / 4 = 46.5 days

Service Reliability Metrics – 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

- Network capacity metric – Mbps/Gbps
- Storage device capacity metric – Gbytes/Terabytes/etc.
- Server capacity metric – CPU, RAM, storage
- Web app capacity metric – requests per mins (SaaS)
- Instance starting time metric – time to initialize new instance (IaaS, SaaS)
- Response time metric – million seconds (SaaS)
- Completion time metric – time required to complete a sync. task

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.
- Metrics include
  - Storage scalability (horizontal) metric – in respond to increased workload
  - Server scalability (horizontal) metric – in respond to increased workload
  - Server scalability (vertical) metric – in respond to workload fluctuations

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.
- 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.

Service Resiliency Metrics (2)

- 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