Cloud SLAs: Present and Future

Salman Baset
sabaset@us.ibm.com

Agenda

- Why consider SLAs?
- Key components of a cloud SLA
- Cloud SLAs of Amazon, Azure, Rackspace, Terremark, Storm on Demand
  - Storage and compute
- Highlights of the comparison
- Future of Cloud SLAs

Why consider cloud SLAs?

- Understand what is promised to the customer
- Build solutions around it
- Propose new SLAs and offerings
- Differentiate services from the competitor

Key components of a cloud SLA

- Service guarantee
  - metrics a provider strives to meet over a time period, e.g., **availability** (99.9%), **response time** (less than 50ms for all transactions), **fault resolution time** (within one hour of problem detection), **zeroing out VM disk**
- Service guarantee time period
  - duration over which a service guarantee should be met, e.g., 99.9% availability in a **month**, response time less than 50ms in a **month**
- Service guarantee granularity
  - resource scale over which a service guarantee is specified, e.g., 99.9% availability **per VM or per data center (including its software stack)**, response time **per transaction or average**
  - **resource group**, e.g., aggregate uptime of all instances 99.9%

Key components of a cloud SLA

- Service guarantee exclusions
  - instances excluded from service guarantee exclusion, e.g., 99.9% availability **excluding scheduled maintenance, patching, customer abuse**

- Service violation measurement and reporting
  - how is the service violation measured and who reports it?

- Service credit
  - amount credited to the customer or applied towards future payments when an SLA is violated.
  - automatic credit or credit upon reporting.

Cloud providers considered

- Compute
  - Amazon EC2
  - Azure Compute
  - Rackspace Cloud Servers
  - Terremark vCloud Express (Verizon)
  - Storm on Demand
  - SCE+
- Storage
  - Amazon S3
  - Azure Storage
  - Rackspace Cloud Files

Amazon

- Data center (region), availability zones
- EC2 compute services
  - Hourly, reserved, spot instances within an availability zone in a region
  - All covered by EC2 SLA
- Storage service
  - S3: blob storage and retrieval (1 B to 5 TB)
  - Remote disks (Elastic block store) for EC2 instances
  - Simple Table
  - Only blob storage and retrieval (S3) covered by storage SLA

- EC2 SLA
  - Availability
  - Per data center instead of per VM
    - SLA is met if new or replacement VMs within data center can be launched 99.95% of the time
    - Data center unavailability measured in contiguous intervals of five minutes
  - No VM performance guarantee
  - 10% of customer bill if availability less than 99.95%
- S3 SLA
  - Number of completed transactions
  - No performance guarantee

Microsoft Azure

- Azure Compute
  - Three roles, web role, worker role, VM role (beta)
  - Compute SLA applicable only to web and worker
  - Fault and update domain
    - Fault domain is a single point of failure. Can be a single machine, but can also be a rack, details not specified in SLA.
    - Update domain: which VMs simultaneously receive patches
- Azure Storage
  - Blob storage similar to S3
  - Structured data storage
  - Queuing service, and remote disks (Azure drive)
  - All backed by SLA
- Azure Compute SLA
  - Connectivity guarantee per role
  - Uptime guarantee per role
    - Patching and maintenance excluded
  - No performance guarantee
- Azure Storage SLA
  - Maximum processing time per transaction, data transfer time not included
  - Excluded transaction list: pre-authentication failures, abusive, creation or deletion of tables, containers, queues.

Rackspace

- Cloud Servers
  - Instances purchased on hourly basis
- Cloud Files
  - Files back up service

Availability	Credit amount
100-99.9%	0%
99.89%-99.5%	10%
99.49%-99.0%	25%
98.99%-98.0%	40%
97.99%-97.5%	55%
97.49%-97.0%	70%
96.99%-96.5%	85%
< 96.5%	100%

- Cloud Servers SLA
  - Per VM (implied from SLA)
  - 100% guarantee for data center network, HVAC, physical network
  - Excluding scheduled maintenance
    - Announced 10 days in advance
  - Physical server failure
    - Repair within an hour of problem identification
    - VMs migrated within 3 hours due to overload (offline migration)
- Cloud Files
  - 99.9% availability, completed transactions
  - Unavailable
    - Data center network is down
    - Service returns a 500-599 http response within two 90s intervals
  - Scheduled maintenance
    - Announced 10 days in advance

Terremark vCloud Express

- Compute
  - VMs purchased on hourly basis
- No storage service
- Compute SLA
  - 100% uptime guarantee for data center
  - Unavailable: data center infrastructure or network is down or user cannot access the web console for 15 minutes
  - No performance guarantee, customer responsible for detecting SLA violation

Storm on Demand

- Compute
  - VMs purchased on hourly basis
- No storage service
- Compute SLA
  - 100% uptime guarantee per instance
  - Infrastructure and patch maintenance excluded from service guarantee
  - 1000% for every hour of downtime – may not exceed customer bill

Compute SLA Comparison

	Amazon EC2	Azure Compute	Rackspace Cloud Servers	Terremark vCloud Express	Storm on Demand
Service guarantee	Availability (99.95%) 5 minute interval	Role uptime and availability, 5 minute interval	Availability	Availability	Availability
Granularity	Data center	Aggregate across all role	Per instance and data center + mgmt. stack	Data center + management stack	Per instance
Scheduled maintenance	Unclear if excluded	Includ. in service guarantee calc.	Excluded	Unclear if excluded	Excluded
Patching	N/A	Excluded	Excluded if managed	N/A	Excluded
Guarantee time period	365 days or since last claim	Per month	Per month	Per month	Unclear
Service credit	10% if < 99.95%	10% if < 99.95% 25% if < 99%	5% to 100%	$1 for 15 minute downtime up to 50% of customer bill	1000% for every hour of downtime –
Violation report respon.	Customer	Customer	Customer	Customer	Customer
Reporting time period	N/A	5 days of occurrence	N/A	N/A	N/A
Claim filing timer period	30 business days of last reported incident in claim	Within 1 billing month of incident	Within 30 days of downtime	Within 30 days of the last reported incident in claim	Within 5 days of incident in question
Credit only for future payments	Yes	No	No	Yes	No

Storage SLA comparison

	Amazon S3	Azure Storage	Rackspace CloudFiles
Service guarantee	Completed transactions (with no error response 500 or 503)	Completed transactions within stipulated time	Completed transactions, data center availability
Granularity	Per transaction	Per transaction	Per transaction
Guarantee time period	Billing month	Per month	Per month
Service credit	10% if < 99.9% 25% if < 99%	10% if < 99.9% 25% if < 99%	10% if < 99% 100% if < 96.5%
Violation report responsibility	Customer	Customer	Customer
Reporting time period	N/A	5 days of incident occurrence	N/A
Claim filing timer period	Within 10 business days following the month in which incident occurred	Within one billing month of incident occurring	Within 30 days following unavailability
Credit only for future payments	Yes	No	No

Highlights of comparison

- Weak uptime guarantees for compute
  - Data center, per instance (only implicit)
- No performance guarantees for compute
- Customer detects SLA violation
  - Does not work for enterprise SLAs
  - Verizon detects SLA violation for its dedicated Internet enterprises
- Service credit
  - Partial credit, no automatic refund, and applied for future payments
- SLA violation reporting time period
  - 5 – 30 days
- Storage SLA: performance vs. request completion
- SLA jargon
  - 100% uptime, but qualified with scheduled maintenance

Future of cloud SLAs

- Service guarantee
  - More than just uptime or performance, e.g., ticket resolution time, zeroing out a VM disk.
- Service guarantee granularity
  - the finer the guarantee, the more stringent the SLA, e.g., data center uptime (coarser) > VM uptime > CPU cycles.
  - aggregate SLAs leave provider more wiggle room to manage resources.
- Service guarantee time period
  - the smaller the time period, the more stringent the guarantee, e.g., CPU cycles over 10 hours vs. CPU cycles over 5 minutes, ticket response time less than 10 minutes.
- Service violation detection and credit
  - enterprise provider must detect SLA and automatically credit the customer for premium services
- Standardization of SLAs
  - structured representation of SLAs
- Oversubscription
  - VM quiescing and migration algorithm should be tied to SLA