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

<table>
<thead>
<tr>
<th>Availability</th>
<th>Credit amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>100-99.9%</td>
<td>0%</td>
</tr>
<tr>
<td>99.89%-99.5%</td>
<td>10%</td>
</tr>
<tr>
<td>99.49%-99.0%</td>
<td>25%</td>
</tr>
<tr>
<td>98.99%-98.0%</td>
<td>40%</td>
</tr>
<tr>
<td>97.99%-97.5%</td>
<td>55%</td>
</tr>
<tr>
<td>97.49%-97.0%</td>
<td>70%</td>
</tr>
<tr>
<td>96.99%-96.5%</td>
<td>85%</td>
</tr>
<tr>
<td>&lt; 96.5%</td>
<td>100%</td>
</tr>
</tbody>
</table>

• **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
<table>
<thead>
<tr>
<th>Service guarantee</th>
<th>Amazon EC2</th>
<th>Azure Compute</th>
<th>Rackspace Cloud Servers</th>
<th>Terremark vCloud Express</th>
<th>Storm on Demand</th>
</tr>
</thead>
<tbody>
<tr>
<td>Availability</td>
<td>(99.95%)</td>
<td>Role uptime and availability, 5 minute interval</td>
<td>Availability</td>
<td>Availability</td>
<td>Availability</td>
</tr>
<tr>
<td>5 minute interval</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Granularity</th>
<th>Data center</th>
<th>Aggregate across all role</th>
<th>Per instance and data center + mgmt. stack</th>
<th>Data center + management stack</th>
<th>Per instance</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Scheduled maintenance</th>
<th>Unclear if excluded</th>
<th>Includ. in service guarantee calc.</th>
<th>Excluded</th>
<th>Unclear if excluded</th>
<th>Excluded</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Patching</th>
<th>N/A</th>
<th>Excluded</th>
<th>Excluded if managed</th>
<th>N/A</th>
<th>Excluded</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Guarantee time period</th>
<th>365 days or since last claim</th>
<th>Per month</th>
<th>Per month</th>
<th>Per month</th>
<th>Unclear</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Service credit</th>
<th>10% if &lt; 99.95%</th>
<th>10% if &lt; 99.95% 25% if &lt; 99%</th>
<th>5% to 100%</th>
<th>$1 for 15 minute downtime up to 50% of customer bill</th>
<th>1000% for every hour of downtime –</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Violation report respon.</th>
<th>Customer</th>
<th>Customer</th>
<th>Customer</th>
<th>Customer</th>
<th>Customer</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Reporting time period</th>
<th>N/A</th>
<th>5 days of occurrence</th>
<th>N/A</th>
<th>N/A</th>
<th>N/A</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Claim filing timer period</th>
<th>30 business days of last reported incident in claim</th>
<th>Within 1 billing month of incident</th>
<th>Within 30 days of downtime</th>
<th>Within 30 days of the last reported incident in claim</th>
<th>Within 5 days of incident in question</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Credit only for future payments</th>
<th>Yes</th>
<th>No</th>
<th>No</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
</table>

Compute SLA Comparison
### Storage SLA comparison

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