# CLOUD COMPUTING

## SERVICE LEVEL AGREEMENT (SLA)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
IIT KHARAGPUR

## What is Service Level Agreement?

- A formal contract between a Service Provider (SP) and a Service Consumer (SC)
- SLA: foundation of the consumer's trust in the provider
- Purpose : to define a formal basis for performance and availability the SP guarantees to deliver
- SLA contains Service Level Objectives (SLOs)
  - Objectively measurable conditions for the service
  - SLA & SLO: basis of selection of cloud provider

2

## SLA Contents

- A set of services which the provider will deliver
- A complete, specific definition of each service
- The responsibilities of the provider and the consumer
- A set of metrics to measure whether the provider is offering the services as guaranteed
- An auditing mechanism to monitor the services
- The remedies available to the consumer and the provider if the terms are not satisfied
- How the SLA will change over time

3

## Web Service SLA

- WS-Agreement
  - XML-based language and protocol for negotiating, establishing, and managing service agreements at runtime
  - Specify the nature of agreement template
  - Facilitates in discovering compatible providers
  - Interaction : request-response
  - SLA violation : dynamically managed and verified
- WSLA (Web Service Level Agreement Framework)
  - Formal XML-schema based language to express SLA and a runtime interpreter
  - Measure and monitor QoS parameters and report violations
  - Lack of formal definitions for semantics of metrics

4

## Difference between Cloud SLA and Web Service SLA

- QoS Parameters :
  - Traditional Web Service : response time, SLA violation rate for reliability, availability, cost of service, etc.
  - Cloud computing : QoS related to security, privacy, trust, management, etc.
- Automation :
  - Traditional Web Service : SLA negotiation, provisioning, service delivery, monitoring are not automated.
  - Cloud computing : SLA automation is required for highly dynamic and scalable service consumption
- Resource Allocation :
  - Traditional Web Service : *UDDI (Universal Description Discovery and Integration)* for advertising and discovering between web services
  - Cloud computing : resources are allocated and distributed globally without any central directory

5

## Types of SLA

- Present market place features two types of SLAs :
  - Off-the-shelf SLA or non-negotiable SLA or Direct SLA
    - Non-conducive for mission-critical data or applications
    - Provider creates the SLA template and define all criteria viz. contract period, billing, response time, availability, etc.
    - *Followed by the present day state-of-the-art clouds*.
  - Negotiable SLA
    - Negotiation via external agent
    - Negotiation via multiple external agents

6

## Service Level Objectives (SLOs)

- Objectively measurable conditions for the service
- Encompasses multiple QoS parameters viz. availability, serviceability, billing, penalties, throughput, response time, or quality
- Example :
  - "***Availability*** of a service X is 99.9%"
  - "***Response time*** of a database query Q is between 3 to 5 seconds"
  - "***Throughput*** of a server S at peak load time is 0.875"

7

## Service Level Management

- Monitoring and measuring performance of services based on SLOs
- Provider perspective :
  - Make decisions based on business objectives and technical realties
- Consumer perspective :
  - Decisions about how to use cloud services

8

## Considerations for SLA

- **Business Level Objectives**: Consumers should know *why* they are using cloud services before they decide *how* to use cloud computing.
- **Responsibilities of the Provider and Consumer**: The balance of responsibilities between providers and consumers will vary according to the type of service.
- **Business Continuity and Disaster Recovery**: Consumers should ensure their cloud providers have adequate protection in case of a disaster.
- **System Redundancy**: Many cloud providers deliver their services via massively redundant systems. Those systems are designed so that even if hard drives or network connections or servers fail, consumers will not experience any outages.

9

## Considerations for SLA *(contd...)*

- **Maintenance**: Maintenance of cloud infrastructure affects any kind of cloud offerings (applicable to both software and hardware)
- **Location of Data**: If a cloud service provider promises to enforce data location regulations, the consumer must be able to audit the provider to prove that regulations are being followed.
- **Seizure of Data**: If law enforcement targets the data and applications associated with a particular consumer, the multi-tenant nature of cloud computing makes it likely that other consumers will be affected. Therefore, the consumer should consider using a third-party to keep backups of their data
- **Failure of the Provider**: Consumers should consider the financial health of their provider and make contingency plans. The provider's policies of handling data and applications of a consumer whose account is delinquent or under dispute are to be considered.
- **Jurisdiction**: Consumers should understand the laws that apply to any cloud providers they consider.

10

## SLA Requirements

- **Security**: Cloud consumer must understand the controls and federation patterns necessary to meet the security requirements. Providers must understand what they should deliver to enable the appropriate controls and federation patterns.
- **Data Encryption**: Details of encryption and access control policies.
- **Privacy**: Isolation of customer data in a multi-tenant environment.
- **Data Retention and Deletion**: Some cloud providers have legal requirements of retaining data even of it has been deleted by the consumer. Hence, they must be able to prove their compliance with these policies.
- **Hardware Erasure and Destruction**: Provider requires to zero out the memory if a consumer powers off the VM or even zero out the platters of a disk, if it is to be disposed or recycled.

11

## SLA Requirements *(Contd...)*

- **Regulatory Compliance**: If regulations are enforced on data and applications, the providers should be able to prove compliance.
- **Transparency**: For critical data and applications, providers must be proactive in notifying consumers when the terms of the SLA are breached.
- **Certification**: The provider should be responsible in proving the certification of any kind of data or applications and keeping its up-to date.
- **Monitoring**: To eliminate the conflict of interest between the provider and the consumer, a neural third-party organization is the best solution to monitor performance.
- **Auditability**: As the consumers are liable to any breaches that occur, it is vital that they should be able to audit provider's systems and procedures. An SLA should make it clear how and when those audits take place. Because audits are disruptive and expensive, the provider will most likely place limits and charges on them.

12

## Key Performance Indicators (KPIs)

- Low-level resource metrics
- Multiple *KPI*s are composed, aggregated, or converted to for high-level *SLO*s.
- Example :
  - downtime, uptime, inbytes, outbytes, packet size, etc.
- Possible mapping :
  - ***Availability (A) = 1 – (downtime/uptime)***

13

## Industry-defined KPIs

- Monitoring:
  - Natural questions:
    - "who should monitor the performance of the provider?"
    - "does the consumer meet its responsibilities?"
  - Solution: neutral third-party organization to perform monitoring
  - Eliminates conflicts of interest if:
    - Provider reports outage at its sole discretion
    - Consumer is responsible for an outage
- Auditability:
  - Consumer requirement:
    - Is the provider adhering to legal regulations or industry-standard
    - SLA should make it clear how and when to conduct audits

14

## Metrics for Monitoring and Auditing

- **Throughput** – How quickly the service responds
- **Availability** – Represented as a percentage of uptime for a service in a given observation period.
- **Reliability** – How often the service is available
- **Load balancing** – When elasticity kicks in (new VMs are booted or terminated, for example)
- **Durability** – How likely the data is to be lost
- **Elasticity** – The ability for a given resource to grow infinitely, with limits (the maximum amount of storage or bandwidth, for example) clearly stated
- **Linearity** – How a system performs as the load increases

15

## Metrics for Monitoring and Auditing *(Contd...)*

- **Agility** – How quickly the provider responds as the consumer's resource load scales up and down
- **Automation** – What percentage of requests to the provider are handled without any human interaction
- **Customer service response times** – How quickly the provider responds to a service request. This refers to the human interactions required when something goes wrong with the on-demand, self-service aspects of the cloud.
- **Service-level violation rate** – Expressed as the mean rate of SLA violation due to infringements of the agreed warranty levels.
- **Transaction time** – Time that has elapsed from when a service is invoked till the completion of the transaction, including the delays.
- **Resolution time** – Time period between detection of a service problem and its resolution.

16

# SLA Requirements w.r.t. Cloud Delivery Models

| Requirement | Platform as a Service | Infrastructure as a Service | Software as a Service |
|---|---|---|---|
| Data Encryption | ✓ | ✓ | |
| Privacy | ✓ | ✓ | ✓ |
| Data Retention and Deletion | | ✓ | ✓ |
| Hardware Erasure and Destruction | | ✓ | ✓ |
| Regulatory Compliance | ✓ | ✓ | ✓ |
| Transparency | ✓ | ✓ | ✓ |
| Certification | ✓ | ✓ | ✓ |
| Terminology for Key Performance Indicators | | ✓ | ✓ |
| Metrics | ✓ | ✓ | ✓ |
| Auditability | ✓ | ✓ | ✓ |
| Monitoring | ✓ | ✓ | ✓ |
| Machine-Readable SLAs | | ✓ | |

*Source: "Cloud Computing Use Cases White Paper" Version 4.0*

17

# Example Cloud SLAs

| Cloud Provider | Service | Type of Delivery Model | Service Level Agreement Guarantees |
|---|---|---|---|
| Amazon | EC2 | IaaS | Availability (99.95%) with the following definitions : Service Year : 365 days of the year, Annual Percentage Uptime, Region Unavailability : no external connectivity during a five minute period, Eligible Credit Period, Service Credit |
| | S3 | Storage-as-a-Service | Availability (99.9%) with the following definitions: Error Rate, Monthly Uptime Percentage, Service Credit |
| | SimpleDB | Database-as-a-Service | No specific SLA is defined and the agreement does not guarantee availability |
| Salesforce | CRM | PaaS | No SLA guarantees for the service provided |
| Google | Google App Engine | PaaS | Availability (99.9%) with the following definitions : Error Rate, Error Request, Monthly Uptime Percentage, Scheduled Maintenance, Service Credits, and SLA exclusions |

18

## Example Cloud SLAs *(contd...)*

| Cloud Provider | Service | Type of Delivery Model | Service Level Agreement Guarantees |
|---|---|---|---|
| Microsoft | Microsoft Azure Compute | IaaS/PaaS | Availability (99.95%) with the following definitions : Monthly Connectivity Uptime Service Level, Monthly Role Instance Uptime Service Level, Service Credits, and SLA exclusions |
| | Microsoft Azure Storage | Storage-as-a-Service | Availability (99.9%) with the following definitions: Error Rate, Monthly Uptime Percentage, Total Storage Transactions, Failed Storage Transactions, Service Credit, and SLA exclusions |
| Zoho suite | Zoho mail, Zoho CRM, Zoho books | SaaS | Allows the user to customize the service level agreement guarantees based on : Resolution Time, Business Hours & Support Plans, and Escalation |

19

## Example Cloud SLAs *(contd...)*

| Cloud Provider | Service | Type of Cloud Delivery Model | Service Level Agreement Guarantees |
|---|---|---|---|
| Rackspace | Cloud Server | IaaS | Availability regarding the following: Internal Network (100%), Data Center Infrastructure (100%), Load balancers (99.9%)<br>Performance related to service degradation: Server migration, notified 24 hours in advance, and is completed in 3 hours (maximum)<br>Recovery Time: In case of failure, guarantee of restoration/recovery in 1 hour after the problem is identified. |
| Terremark | vCloud Express | IaaS | Monthly Uptime Percentage (100%) with the following definitions: Service Credit, Credit Request and Payment Procedure, and SLA exclusions |

20

# Example Cloud SLAs *(contd…)*

| Cloud Provider | Service | Type of Cloud Delivery Model | Service Level Agreement Guarantees |
|---|---|---|---|
| Nirvanix | Public, Private, Hybrid Cloud Storage | Storage-as-a-Service | Monthly Availability Percentage (99.9%) with the following definitions: Service Availability, Service Credits, Data Replication Policy, Credit Request Procedure, and SLA Exclusions |

21

# Limitations

- **Service measurement**
  - Restricted to uptime percentage
  - Measured by taking the mean of service availability observed over a specific period of time
  - Ignores other parameters like stability, capacity, etc.
- **Biasness towards vendors**
  - Measurement of parameters are mostly established according to vendor's advantage
- **Lack of active monitoring on customer's side**
  - Customers are given access to some ticketing systems and are responsible for monitoring the outages.
  - Providers do not provide any access to active data streams or audit trails, nor do they report any outages.

22

## Limitations *(contd…)*

- Gap between *QoS hype* and *SLA offerings* in reality
- QoS in the areas of *governance*, *reliability*, *availability*, *security*, and *scalability* are not well addressed.
- No formal ways of verifying if the SLA guarantees are complying or not.
- Proper SLA are good for both provider as well as the customer
  - Provider's perspective : Improve upon Cloud infrastructure, fair competition in Cloud market place
  - Customer's perspective : Trust relationship with the provider, choosing appropriate provider for moving respective businesses to Cloud

23

## Expected SLA Parameters

- Infrastructure-as-a-Service (IaaS):
  - CPU capacity, cache memory size, boot time of standard images, storage, scale up (maximum number of VMs for each user), scale down (minimum number of VMs for each user), On demand availability, scale uptime, scale downtime, auto scaling, maximum number of VMs configured on physical servers, availability, cost related to geographic locations, and response time
- Platform-as-a-Service (PaaS):
  - Integration, scalability, billing, environment of deployment (licenses, patches, versions, upgrade capability, federation, etc.), servers, browsers, number of developers

24

## Expected SLA Parameters *(contd...)*

- Software-as-a-Service (SaaS):
  - Reliability, usability, scalability, availability, customizability, Response time
- Storage-as-a-Service :
  - Geographic location, scalability, storage space, storage billing, security, privacy, backup, fault tolerance/resilience, recovery, system throughput, transferring bandwidth, data life cycle management

25

# Thank You!

26