Cloud SLAs: Present and Future

Salman A. Baset
sabaset@us.ibm.com
IBM Research

Abstract

The variability in the service level agreements (SLAs) of cloud providers prompted us to ask the question how do the SLAs compare and how should the SLAs be defined for future cloud services. We break down a cloud SLA into easy to understand components and use it to compare SLAs of public cloud providers. Our study indicates that none of the surveyed cloud providers offer any performance guarantees for compute services and leave SLA violation detection to the customer. We then provide guidance on how SLAs should be defined for future cloud services.

1 Introduction

Cloud-based services are increasingly becoming commonplace. These services include infrastructure-as-a-service (IaaS), platform-as-a-service (PaaS), and software-as-a-service (SaaS) ([10]). Each service is typically accompanied by a service level agreement (SLA) which defines the minimal guarantees that a provider offers to its customers. The lack of standardization in cloud-based services implies a corresponding lack of clarity in the service level agreements offered by different providers.

In this paper, our goal is to systematically compare the SLAs of cloud providers and to provide guidance on how SLAs should be defined for cloud-based services in future. We break down a cloud SLA into easy to understand components (which we refer to as anatomy of cloud SLA) and use it to compare the service level agreements of Amazon [1], Rackspace [13], Microsoft Windows Azure [21], Terremark vCloud Express [18], and Storm on Demand [16]. All of the considered cloud providers offer IaaS services, and some offer PaaS services. By breaking down SLA into different components, we are able to highlight the similarities and differences among the SLAs of considered cloud providers. A common aspect of the considered SLAs is that none of the cloud providers offer any performance guarantees for the compute services. Moreover, no cloud provider automatically credits the customer for SLA violations, and leaves the burden of providing evidence for any such violation on the customer.

The rest of this paper is organized as follows. Section 2 discusses the anatomy of a typical cloud SLA. Section 3 gives an overview of cloud services offered by the considered cloud providers. Section 4 describes the SLAs of the considered cloud providers. Section 5 highlights the key aspects of our comparison, and discusses what is missing from the considered SLAs. Section 6 provides guidance on defining SLAs for future cloud-based services, and the challenges involved in defining performance based SLAs.

2 Anatomy of a Typical Cloud SLA

A typical SLA of a cloud provider has the following components.

- **Service guarantee** specifies the metrics which a provider strives to meet over a service guarantee time period. Failure to achieve those metrics will result in a service credit to the customer. Availability (e.g., 99.9%), response time (e.g., less than 50 ms), disaster recovery, and fault resolution time (e.g., within one hour of detection) are examples of service guarantees. Some service guarantees can be on a per action basis, such as zeroing out a VM disk when it is deprovisioned.

- **Service guarantee time period** describes the duration over which a service guarantee should be met. The time period can be a billing month or time elapsed since the last claim was filed. The time period can also be small, e.g., one hour. The smaller the time period, the more stringent is the service guarantee.

- **Service guarantee granularity** describes the resource scale on which a provider specifies a service guarantee. For example, the granularity can be on a per service, per data center, per instance, or per transaction basis. Similar to time period, the service guarantee can be stringent if the granularity of service guarantee is fine-grained.

  Service guarantee granularity can also be calculated as an aggregate of the considered resources, such as instances or transactions. For example, aggregate uptime of all running instances must be greater than 99.95%. However, such a guarantee implies that some instances in the aggregate SLA computation can potentially have a lower percentage uptime than 99.95% while still meeting the aggregate SLA. As a consequence, aggregate SLA computation leaves provider the wiggle room to better manage its offered services.

- **Service guarantee exclusions** are the instances that are excluded from service guarantee metric calculations. These exclusions typically include abuse of the system by a customer, or any downtime associated with the scheduled maintenance.

- **Service credit** is the amount credited to the customer or applied towards future payments if the service guarantee is not met. The amount can be a complete or partial credit of the customer payment for the affected service.

- **Service violation measurement and reporting** describes how and who measures and reports the violation of service guarantee, respectively.

3 Cloud Providers Considered

We briefly give an overview of cloud services offered by Amazon [1], Rackspace [13], Microsoft Azure [21], Terremark vCloud Express [18], and Storm on Demand [16]. These providers offer IaaS and PaaS compute and storage services. The compute service comprises of a virtual machine (or instance) or CPU cycles that a customer can purchase on an hourly, monthly, or yearly basis. The storage service allows storage and retrieval of blob or structured data. We interchangeably use customer and user to refer to the clients of cloud providers.

3.1 Amazon

Amazon [1] is an IaaS provider and offers compute (EC2 [2]) and storage (S3 [5]) services. In EC2, a customer can obtain virtual machines (instances) by the hour or reserve them in advance for an entire year [3]. In addition, EC2 offers spot instances where a customer can bid for compute capacity. EC2 SLA [4] is applicable to hourly, spot, and reserved instances.

The storage service S3 provides mechanism for storing and retrieving data objects using `put()`, `get()` operations with data size ranging from one byte to five tera bytes.

Amazon also provides a remote disk capability for its virtual machines, namely, Elastic Block Store (EBS). EBS volumes are replicated within an availability zone. A data center (or region) can contain multiple availability zones. The availability zones do not have power, networking, or hardware equipment in common. EBS volumes are not backed by any SLA; however, the snapshots of EBS volumes can be stored on S3, which as mentioned before is backed by a SLA. When creating an instance, the user must specify the region and availability zone in which she creates the instance.

Amazon also provides a SimpleDB [7] service which is a simplified relational database service. However, the service is still in beta at the time of writing of this paper. Among S3, EBS, and SimpleDB services, only S3 is backed by an SLA [6].

3.2 Windows Azure

Windows Azure [21] is a PaaS and IaaS cloud provider that offers compute (Azure Compute [22]) and storage (Azure Storage [24]) services. Azure Compute comprises of three types of compute services (which it refers to as roles), namely, web, worker, and a VM. A web role provides a web based front end for an application and comprises of an IIS server [11]. A worker role is useful for generalized development. It can run Apache Tomcat and Java Virtual Machines (JVMs) and can be used to perform background processing for a web role. A VM role is similar to instances in Amazon EC2, and gives user complete control over the virtual machine. However, at this time, VM roles are only available in beta [12] and are not covered by Azure Compute SLA [23]. The compute service can only be purchased on an hourly basis, and cannot be reserved in advance for the entire year.

Azure Compute service defines the notion of a fault domain and an upgrade domain. Each compute role belongs to a fault domain and an upgrade domain. A fault domain comprises of a single point of failure and is at least a physical machine, but may also be a rack of machines; the precise details of what comprises a fault domain are not available. An upgrade domain defines which compute roles can simultaneously receive the software or operating system updates. A fault domain may span several update domains. Likewise, an update domain may also span several fault domains.

Azure also provides Azure Storage [24], an S3 like storage service, which can be used for storing and re-

trieving blob and structured data. It also provides a queuing service and remote disks (known as Azure Drive). Azure storage service is backed by a SLA [25].

3.3 Rackspace

Rackspace [13] is an IaaS provider that provides compute instances similar to Amazon EC2 and VM role of Azure, which it refers to as "Cloud Servers". A customer can obtain VMs on an hourly basis which are covered by an SLA [14]. However, unlike EC2, Cloud Servers cannot be reserved in advance for the entire year. Rackspace also provides a managed service level for Cloud Servers. As part of the managed service, Rackspace is responsible for applying software and security patches for operating system and middleware.

Rackspace provides a storage service called "Cloud Files" which allows a customer to store and retrieve files in the cloud and is covered by an SLA [14]. The stored files are internally replicated by Rackspace.

3.4 Terremark vCloud Express

Terremark vCloud Express [18] is an IaaS provider similar to EC2, VM role of Azure, and Rackspace. A customer can obtain VMs on an hourly basis which are covered by an SLA [19].

Terremark does not provide a specialized storage service.

3.5 Storm on Demand

Storm on Demand [16] is an IaaS provider similar to EC2, the VM role of Azure, Rackspace, and Terremark vCloud Express. A customer can obtain VMs on an hourly basis which are covered by an SLA [17].

4 Description of SLAs

We describe SLAs of compute and storage services offered by cloud providers considered in this paper.

4.1 Amazon

Amazon EC2 and S3 services are backed by distinct SLAs. Below, we describe the SLAs of these services in detail.

4.1.1 EC2 SLAs

Amazon EC2 SLA [4] is defined on a per data center (region in Amazon speak) basis instead of per instance. EC2 offers a 99.95% region availability rate (service guarantee). If a user is unable to access her instances in one region during a contiguous period of five minutes or launch replacement instances, the region is deemed to be unavailable during those five minutes. The burden of providing the evidence for region unavailability is on the user. Strictly speaking, if a user is running at least one VM which she cannot access during a five minute interval and cannot launch a replacement, she is eligible for a service credit if the credit value is above one dollar.

A customer can claim a service credit anytime the service falls below the availability SLA in the last 365 days or since the last time a service credit claim was filed by the customer. The service credit is up to 10% of a customer's bill (excluding any one-time costs) for the instances affected by the outage. Service credits are typically only applicable towards future EC2 payments. Amazon requires that the service credit claim be received from the customer within 30 business days of the last reported incident in the filed claim.

Amazon does not provide any service credit for failures of individual instances not attributable to region unavailability. This clause means that even if a region (data center) is available, but some services in that region fail such as EBS on which an instance depends, Amazon is at least legally not bound to provide a service credit, although it may provide a credit at its own discretion. For example, Amazon provided a service credit [8] for its April 2011 outage due to EBS failures. Further, Amazon does not provide any service credits if VMs suffer from any performance issues. A VM can suffer performance degradation due to co-location or hardware differences of the underlying physical machine [15].

Amazon EC2 SLA does not specify that scheduled and unscheduled maintenance are excluded from the service guarantee. EC2 SLA is defined on a data center basis, and, arguably, the data center being unavailable for scheduled maintenance is unlikely because it will impact all customers running their instances in that data center.

4.1.2 S3 SLAs

Amazon S3 SLA [6] provides storage request completion guarantee of 99.9% over a billing month (service guarantee time period). A storage request is considered failed if S3 server returns an "Internal Error" or "Service Unavailable" response to a request. These responses correspond to HTTP response codes 500 and 503. The burden of reporting request failure and providing evidence is on the customer.

S3 calculates failed requests over a five minute interval, which are then averaged over a month. The failed requests are calculated by dividing the number of requests generating an error response to the total number of requests in the five minute interval. The percentage of completed transactions in the billing month is calculated

Request type	Maximum processing time
(1) PutBlob and GetBlob (includes blocks and pages) (2) Get Valid Page Blob Ranges	Must complete within the product of 2 seconds multiplied by the number of MBs transferred in processing the request
Copy Blob	Must complete processing within 90 seconds
PutBlockList, GetBlockList	Must complete processing within 60 seconds
Table query, List operations	Must complete processing or return a continuation within 10 seconds
Batch table operations	Must complete processing within 30 seconds
(1) All Single Entity Table Operations (2) All other Blob and Message Operations	Must complete processing within 2 seconds

Table 1: Maximum processing times for Azure storage transactions [25]

by subtracting from 100% the average of failed request rates from each five minute period.

The service credit is 10% of the customer bill if completion rate is below 99.9% and 25% of the customer bill if completion rate is less than 99%. Amazon must receive the claim within 10 business days after the billing month in which the incident occurred.

Similar to EC2 SLA, Amazon S3 SLA does not exclude scheduled and unscheduled maintenance from service guarantee. Moreover, S3 service does not specify any performance guarantees on the storage requests.

4.2 Windows Azure

Azure compute and storage service are backed by separate SLAs which are described below.

4.2.1 Azure Compute SLA

Azure Compute SLA [23] provides connectivity and uptime service guarantees for its non-beta compute roles over a billing month (service guarantee time period). For Azure Compute SLA to be applicable, a customer must deploy at least two instances of a compute role type in different update domains. Recall from Section 3.2 that an update domain comprises of compute roles which receive the software or operating system updates at the same time.

Unlike Amazon EC2, which provides availability SLA on a per data center basis, Azure SLA is calculated as an aggregate over the deployed roles. Azure SLA defines two service guarantees, namely, external network connectivity and uptime which are calculated on a monthly basis. The connectivity service guarantee is defined as the aggregate time since all the Internet facing roles have been started minus the five minute intervals during which any role does not have connectivity, divided by the aggregate time since roles have been started. Like Amazon EC2, Azure calculates downtime for its compute roles[1] in increments of five minute intervals.

The uptime service guarantee is defined as the aggregate time since roles have been deployed and started minus the time across all role instances which do not run for more than two minutes without corrective action being initiated, divided by the aggregate time since roles have been started. Any performance or availability issues due to regular platform upgrades and patches are excluded from the uptime service guarantee calculation.

The service credit is 10% of the customer bill if connectivity and uptime percentage is below 99.95% and 99.9%, respectively, and 25% if less than 99.9%. The onus for reporting a SLA violation and providing evidence is on the customer. Microsoft requires that a customer notifies it of the incident within five business days following the incident in order to be eligible to file a claim. Then, Microsoft must receive the claim within a month of the billing month in which the incident occurred.

4.2.2 Azure Storage SLA

Azure Storage SLA [25] defines service guarantee as percentage of completed transactions in a billing month. A request is considered failed if the maximum time to process the request exceeds the time specified in the service guarantee. Table 1 lists the maximum processing time for various transactions. The maximum processing time does not include the time taken to transfer data in or out of Azure Storage service.

Azure Storage calculates failed requests over one hour interval by dividing the total number of failed requests to the total storage requests. The percentage of completed transactions within a billing month is calculated by subtracting from 100% the average of failed request rates from each one hour period in the billing month.

[1] non-beta compute roles only include web and worker roles

Similar to Azure Compute, the onus for reporting an SLA violation is on the customer. Microsoft requires that a customer notifies it of the incident within five business days following the incident in order to be eligible to file a claim. Then, Microsoft must receive claim within a month of the billing month in which the incident occurred. The service credit is 10% of the customer bill if number of completed transactions are below 99.9% and 25% of the customer bill if less than 99%.

Similar to S3 SLA, Azure Storage SLA excludes any transactions from SLA computation that are beyond its reasonable control, and that result from customer's fault or abuse of the system. Unlike S3 SLA, Azure storage SLA gives detailed examples of excluded transactions such as pre-authentication failures, abusive transactions, creation or deletion of containers, tables or queues, or flooding requests not obeying back off principles.

4.3 Rackspace

Rackspace provides a compute service, namely, Cloud Servers, similar to EC2 and also provides a storage service for storing and retrieving files, namely, Cloud Files.

4.3.1 Cloud Servers SLA

Rackspace guarantees that its data center network, HVAC, and power will be 100% available in a billing month, excluding the scheduled maintenance. The scheduled maintenance does not exceed sixty minutes in any calendar month and must be announced at least 10 days in advance to the customer. If a scheduled maintenance of sixty minutes occurs every month (30 days), the maximum uptime percentage of a VM can never exceed 99.86%.

If a physical server running the virtual machine fails, Rackspace guarantees that it will be repaired within an hour of problem identification. Further, if VMs need to be migrated due to server overload, it will not take more than three hours. The SLA does not specify if Rackspace performs live or offline migration of a VM.

Rackspace computes SLA violations in increments of 30 minutes for data center network, HVAC, and power, and in one hour increments for downtime associated with physical servers or migration. If data center network, HVAC, power, or physical servers are down, or if VMs need to be migrated, the service credit starts from 5% of the customer bill up to 100% of the customer bill for affected compute instances. The implication of Cloud Servers SLA is that Rackspace provides service guarantee on a per virtual machine (instance) basis.

A customer must contact Rackspace within 30 days following the downtime and provide the problem evidence in order to receive a service credit. However, it is unclear how a customer can provide evidence for a specific problem such as HVAC, power, or network failure. Perhaps, Rackspace maps the customer's evidence to specific problems and determine the service credit accordingly.

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%
Less than 96.5%	100%

Table 2: Cloud Files service credit

4.3.2 Cloud Files SLA

Rackspace provides a 99.9% request completion rate and Cloud Files servers availability guarantee in a billing cycle [14]. The service is considered unavailable if data center network is down, or if the service returns an error response (HTTP 500-599 status code) to a request within two or more consecutive 90 second intervals, or if average download time for a 1-byte document exceeds 0.3 seconds.

The unavailability due to scheduled maintenance is excluded from the availability calculations. Similar to Cloud Servers SLA, the scheduled maintenance period does not exceed 60 minutes and must be announced 10 days in advance.

A customer must contact the Rackspace within 30 days following the downtime and provide the problem evidence in order to receive a service credit. The service credit amount is described in the Table 2.

4.4 Terremark vCloud Express

Terremark vCloud Express only provides IaaS compute services. Below we describe the SLA of this service.

4.4.1 Compute SLA

Terremark vCloud Express provides an uptime service guarantee of 100% for its data center. The service is deemed unavailable if the data center infrastructure or network is down, or if the user cannot access the web console for a duration of 15 minutes (service guarantee time period). The unavailability period of 15 minutes is three times more than the EC2 and Azure Compute

	Amazon EC2	**Azure Compute**	**Rackspace Cloud Servers**	**Terremark vCloud Express**	**Storm on Demand**
Service guarantee	Availability	Availability	Availability	Availability	Availability
Service granularity guarantee	Data center	Aggregate across all roles	Per instance*	Data center + management stack	Per instance*
Infrastructure scheduled maintenance	Unclear if excluded	Included in service guarantee	Excluded from service guarantee	Unclear if excluded	Excluded from service guarantee
OS/software patches on compute instances	N/A	Excluded from service guarantee if managed	Excluded from service guarantee if managed	N/A	Excluded from service guarantee if managed
Service guarantee time period	365 days or since last claim	Billing month	Billing month	Calendar month	Calendar month
Service credit	10% of CB if $<$ 99.95%	10% of CB if $<$ 99.95% 25% of CB if $<$ 99%	5% of CB for every 30 minutes of downtime up to 100%	$1 for 15 minute downtime up to 50% of CB	1000% for every hour of downtime up to CB
Service violation reporting onus	Customer	Customer	Customer	Customer	Customer
Service violation incident reporting	N/A	5 days of incident occurence	N/A	N/A	N/A
Service violation claim filing	within 30 business days of the last reported incident in claim	within one billing month	within 30 days of downtime	within 30 days of the last reported incident in claim	within 5 days of incident in question
SLA publish date	October 23, 2008	April 9, 2010	June 23, 2009	August 31, 2009	Unknown
Credit applied towards future payments only	Yes	No	No	Yes	No

Table 3: Compute SLA comparison. CB is an abbreviation for customer bill. * implied from SLA.

unavailability period of five minutes and as a result its service guarantee is less stringent than EC2 and Azure.

Terremark vCloud Express defines a service credit of one dollar for every 15 minutes of downtime up to 50% of the usage fees on a monthly basis. Similar to EC2, a customer is responsible for reporting the downtime within 30 days of the last day reported in unavailability claim and providing supportive evidence. The service credit is typically applied towards future vCloud Express payments.

4.5 Storm on Demand

Like Terremark vCloud Express, Storm on Demand only provides IaaS compute services. Below we describe the SLA of this service.

4.5.1 Compute SLA

Storm on Demand guarantees a 100% uptime for network and power infrastructure of its data center. However, scheduled maintenance of network and power infrastructure or physical machines is not included in the uptime guarantee. Further, unlike Rackspace Cloud Servers SLA, the maximum scheduled downtime within a month is not part of the SLA. Moreover, if the customer lets Storm on Demand manage the software and operating updates, any downtime associated with these updates is also excluded from the uptime guarantee.

Storm on Demand defines a service credit for service guarantee violations due to network and power infrastructure as follows: if a server is unreachable for one hour, the customer will receive a service credit of 10 hours. In other words, a customer can receive a service credit of up to 1000% of the actual amount associated with the downtime. However, the total compensation amount may not exceed customer's monthly recurring charge. Moreover, the hourly charges applicable for service credit do not include any additional service charges such as backup or additional IP's.

It is the responsibility of a customer to file a credit claim within five business days of the incident in question and provide any supportive evidence.

5 Highlights of SLA Comparison

Table 3 and Table 4 show a tabular comparison of compute and storage service SLAs of the cloud providers considered in this paper. We highlight several aspects from this tabular comparison and from SLA description in Section 4.

5.1 Weak Uptime Guarantees for Compute

The considered cloud providers offer only weak uptime guarantees for running VMs and do not explicitly specify uptime guarantee on a per instance basis. Amazon EC2 and Terremark vCloud Express only offer uptime guarantee on a per data center basis instead of per instance basis. Arguably, the chance of a data center being unavailable is much lower than per instance. Rackspace Cloud Servers and Storm on Demand implicitly provide SLA guarantee on per instance basis. Azure Compute provides uptime guarantees as an aggregate over all instances instead of per instance basis.

5.2 No Performance Guarantees for Compute

None of the considered cloud providers offer any performance guarantees for compute instances. As a consequence, a customer can only hope for its instances to receive the provisioned CPU, memory, network, and disk resources. Lack of performance guarantees may be unacceptable in enterprise clouds.

5.3 Customer Should Detect SLA Violation

The considered cloud providers leave the burden of detecting SLA violation on the customer which may be unacceptable for enterprise. Verizon, an enterprise Internet connectivity provider, detects SLA violations for its dedicated Internet service [20].

5.4 Service Credit

Service credits for SLA violation offered by Amazon EC2, S3, and Terremark vCloud Express can only be applied towards future payments of respective cloud services. Amazon EC2 and S3, Azure Compute and Storage, and Terremark vCloud Express only partially reimburse the total cost of affected services, while Rackspace Cloud Servers and Storm on Demand can provide up to 100% reimbursement. Storm on Demand is unique, in that it offers to reimburse 1000% of the cost of affected instances. However, it is relatively a new IaaS provider.

5.5 SLA Violation Reporting Time Period

Azure Compute and Storage and Storm on Demand SLAs stipulate that a customer must report the SLA violation incident within five days of the incident occurrence which is more stringent than the 30 day violation reporting and claim filing time period offered by other cloud providers. Azure does allow a customer to file a claim 30 days after reporting the incident.

5.6 SLA Jargon

An SLA is a legal document and is the sole remedy for a customer for any service violations. The lack of standardization in SLAs and the use of SLAs as a potential marketing vehicle makes it difficult to compare SLAs of different cloud providers. As an example, Storm on Demand, and Rackspace guarantee their data center network, HVAC, and power to be 100% of the time, but they qualify it with scheduled maintenance. Similarly, Azure Storage performance guarantee does not include the time it takes to transfer the data in and out of data center. For a designer of cloud application, it is important to pay attention to these details.

	Amazon S3	Azure Storage	Rackspace Cloud Files
Service guarantee	Completed transactions (with no error response)	Completed transactions (within stipulated time)	Completed transactions, availability
Service granularity guarantee	Per transaction	Per transaction	Per transaction, data center
Service guarantee time period	Billing month	Billing month	Per month
Service credit	10% of CB if $<$ 99.9%, 25% of CB if $<$ 99%	10% of CB if $<$ 99.9%, 25% of CB if $<$ 99%	10% of CB if $<$ 99%, 100% of CB if $<$ 96.5%
Service violation reporting onus	Customer	Customer	Customer
Service violation incident reporting	N/A	5 days of incident occurrence	N/A
Service violation claim filing	within 10 business days following the month in which the incident occurred	within one billing month	within 30 days following unavailability
SLA publish date	October 1, 2007	November 12, 2010	June 23, 2009
Credit applied towards future payments only	Yes	No	No

Table 4: Storage SLA comparison. CB is an abbreviation for customer bill.

5.7 Storage SLAs: Performance vs. Request Completion

For the storage SLAs, only Azure provides a performance guarantee per transaction, i.e., processing time for a request (which excludes the data download or upload time). Amazon S3 and Rackspace Cloud Files do not provide any performance guarantees and instead only provide a request completion guarantee.

5.8 SLAs Offered By Business Internet Providers

We briefly describe the Internet Dedicated SLA offered by Verizon [20] and how it compares with SLAs offered by cloud providers. Verizon Internet SLA comprises of nine service guarantees such as availability, latency, network packet latency, ticket response time for denial of service, and time to repair.

Unlike SLAs of cloud providers considered in this paper, the onus for detecting SLA violation for certain metrics such as availability is on Verizon. However, a customer must explicitly request a service credit within 30 days of the incident occurrence. Scheduled maintenance is excluded from the service guarantee calculation. A customer account is then automatically credited for any service violation which is not the case with any of the considered cloud provider SLAs.

5.9 What is Missing from the SLAs?

The SLAs of the considered cloud providers only focus on availability or request completion rate. An enterprise cloud SLA encompasses much more than availability, request completion rate, or performance. It defines guidelines for disaster recovery, privacy, auditability, and security. The details of which service guarantees to consider for enterprise cloud SLAs can be found in cloud computing use cases paper [9]. Further, unlike the public cloud providers considered in this paper, the burden of detecting SLA violation may lie on the provider.

6 Future of Cloud SLAs

In this section, we consider how a cloud provider may define SLAs for cloud services in the future.

- **Service guarantee:** The considered cloud providers only provide uptime guarantees for IaaS services. The cloud providers may also want to offer other guarantees such as performance, security,

and ticket resolution time. Providing a performance guarantee becomes necessary if cloud providers oversubscribe the resources of physical servers to decrease the number of physical servers used and increase their utilization. The over-subscription of the physical servers implies that performance of virtual machines running on physical servers may become a concern. Further, co-location of a virtual machine with other workloads may also impact the CPU, disk, network, and memory performance of a VM. Moreover, enterprises purchasing cloud based services may demand a minimal level of performance guarantee. Therefore, it may be necessary for a cloud provider to offer performance based SLAs for its IaaS compute services with a tiered pricing model, and charge a premium for guaranteed performance.

- **Service guarantee time period and granularity:** The service guarantee time period and granularity determine how stringent is the underlying service guarantee. A service guarantee is stringent if the metric is performance based for a fine-grained resource over a small time period, e.g. 99.9% of memory transactions in a five minute interval must complete within one micro second. Such a stringent guarantee can be loosened by aggregating the service guarantee over a group of resources (e.g., aggregate uptime percentage of all instances must be greater than 99.5%). Providers can use a combination of service guarantee granularity and service guarantee time period to price their services appropriately. For enterprise and mission critical workloads, a cloud provider may have no choice but to provide finer service guarantees.

- **Service violation detection and credit:** None of the considered providers automatically detect SLA violation and leave the burden of providing the violation proof on the customer. This aspect may not be acceptable to customers with mission critical or enterprise workloads. A cloud provider can differentiate the pricing of its offering if it automatically detects and credits the customer for SLA violation. However, the tooling cost to automatically measure, record, and audit SLA metrics can be a concern.

- **Outcome based SLAs:** The cloud providers considered in this paper offer IaaS and PaaS services. Using these services, a customer can deploy her own applications in the cloud. However, in the future, cloud providers may offer outcome based services on top of cloud, where a provider delivers a complete solution for a customer using cloud. For outcome based services, a cloud provider needs to define SLAs for the promised outcomes and how those SLAs map to the underlying IaaS and PaaS infrastructure it provides.

- **Standardization of SLAs:** The lack of standardization in cloud SLAs makes it difficult for a customer to effectively compare them. As cloud services mature, and as the vision of utility computing is realized, the standardization of SLA is likely to take center stage. Structured representation of SLAs (e.g., in XML) may be necessary for standardized SLAs.

7 Conclusion

In this paper, we break down a cloud SLA into several components, and use it for comparing SLAs of well known public IaaS providers. Our study indicates that none of the IaaS providers offer any performance based SLAs for compute services. Moreover, all cloud providers leave the burden of providing evidence for SLA violation on the customer. We then discussed how SLAs should be defined for future cloud services. We believe that customers and cloud providers will benefit from this study when purchasing or selling cloud-based services. This study will also help in clarifying and defining SLAs of existing and future cloud based services.

References

[1] Amazon. https://www.amazon.com, 2011. [Online; accessed August 2011].

[2] Amazon EC2. https://aws.amazon.com/ec2/, 2011. [Online; accessed August 2011].

[3] Amazon EC2 Reserved Instances. https://aws.amazon.com/ec2/reserved-instances/, 2011. [Online; accessed August 2011].

[4] Amazon EC2 SLA. https://aws.amazon.com/ec2-sla/, 2011. [Online; accessed August 2011].

[5] Amazon S3. https://aws.amazon.com/s3/, 2011. [Online; accessed August 2011].

[6] Amazon S3 SLA. https://aws.amazon.com/s3-sla/, 2011. [Online; accessed August 2011].

[7] Amazon SimpleDB. https://aws.amazon.com/simpledb/, 2011. [Online; accessed August 2011].

[8] Amazon SLAs Didn't Cover Major Outage. http://www.informationweek.com/news/cloud-computing/software/229403086, 2011. [Online; accessed August 2011].

[9] Cloud Computing Usecases Whitepaper. http://www.opencloudmanifesto.org/Cloud_Computing_Use_Cases_Whitepaper-4_0.pdf, 2011. [Online; accessed August 2011].

[10] P. Mell and T. Grance. The NIST Definition of Cloud Computing. NIST Special Publication 800-145, September 2011.

[11] Microsoft IIS. https://www.iis.net/, 2011. [Online; accessed August 2011].

[12] Overview of the Windows Azure VM Role. http://msdn.microsoft.com/en-us/library/gg433107.aspx, 2011. [Online; accessed August 2011].

[13] Rackspace. https://www.rackspace.com, 2011. [Online; accessed August 2011].

[14] Rackspace Cloud Servers SLA. http://www.rackspace.com/cloud/legal/sla/, 2011. [Online; accessed August 2011].

[15] J. Schad, J. Dittrich, and J.-A. Quiané-Ruiz. Runtime measurements in the cloud: observing, analyzing, and reducing variance. *Proc. VLDB Endow.*, 3:460–471, September 2010.

[16] Storm on Demand. http://www.stormondemand.com, 2011. [Online; accessed August 2011].

[17] Storm on Demand SLA. https://www.stormondemand.com/about/sla.html, 2011. [Online; accessed August 2011].

[18] Terremark vCloud Express. http://vcloudexpress.terremark.com/, 2011. [Online; accessed August 2011].

[19] Terremark vCloud Express SLA. https://community.vcloudexpress.terremark.com/en-us/product_docs/w/wiki/service-level-agreement.aspx, 2011. [Online; accessed August 2011].

[20] Verizon Internet Dedicated Service Level Agreement. http://www.verizonbusiness.com/terms/us/products/internet/sla/, 2012. [Online; accessed March 2012].

[21] Windows Azure. https://www.microsoft.com/windowsazure/, 2011. [Online; accessed August 2011].

[22] Windows Azure Compute. https://www.microsoft.com/windowsazure/features/compute/, 2011. [Online; accessed August 2011].

[23] Windows Azure Compute SLA. https://www.microsoft.com/download/en/details.aspx?displaylang=en&id=24434, 2011. [Online; accessed August 2011].

[24] Windows Azure Storage. https://www.microsoft.com/windowsazure/features/storage/, 2011. [Online; accessed August 2011].

[25] Windows Azure Storage SLA. https://www.microsoft.com/download/en/details.aspx?displaylang=en&id=6656, 2011. [Online; accessed August 2011].