ETSI TR 103 125 V1.1.1 (2012-11)

CLOUD;
SLAs for Cloud services

Reference
DTR/CLOUD-0012-SLArequirements

Keywords
CLOUD, SLA

ETSI

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - NAF 742 C
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° 7803/88

Contents

Intellectual Property Rights

IPRs essential or potentially essential to the present document may have been declared to ETSI. The information pertaining to these essential IPRs, if any, is publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI Web server (http://ipr.etsi.org).

Pursuant to the ETSI IPR Policy, no investigation, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Foreword

This Technical Report (TR) has been produced by ETSI Technical Committee CLOUD (CLOUD).

Introduction

The present document aims to review previous work on SLAs including ETSI guides from TC USER, contributions from EuroCIO, European research projects (FP7) etc., and to derive potential requirements for work on cloud specific SLA standards.

1 Scope

The present document aims to review previous work on SLAs including ETSI guides from TC USER and contributions from EuroCIO, etc. and to derive potential requirements for cloud specific SLA standards.

2 References

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the referenced document (including any amendments) applies.

Referenced documents which are not found to be publicly available in the expected location might be found at http://docbox.etsi.org/Reference.

NOTE: While any hyperlinks included in this clause were valid at the time of publication ETSI cannot guarantee their long term validity.

2.1 Normative references

The following referenced documents are necessary for the application of the present document.

Not applicable.

2.2 Informative references

The following referenced documents are not necessary for the application of the present document but they assist the user with regard to a particular subject area.

[i.1] ETSI EG 202 009-1: "User Group; Quality of telecom services; Part 1: Methodology for identification of parameters relevant to the Users".

[i.2] ETSI EG 202 009-2: "User Group; Quality of Telecom Services; Part 2: User related parameters on a service specific basis".

[i.3] ETSI EG 202 009-3: "User Group; Quality of telecom services; Part 3: Template for Service Level Agreements (SLA)".

[i.4] NIST Special Publication 800-145: "The NIST Definition of Cloud Computing".

[i.5] FG Cloud TR: "Focus Group on Cloud Computing Technical Report Part 1: Introduction to the cloud ecosystem: definitions, taxonomies, use cases and high-level requirements", version 1.0 (02/2012).

[i.6] ITU-T Recommendation G.1000: "Communications Quality of Service: A framework and definitions".

[i.7] ETSI ETR 003: "Network Aspects (NA); General aspects of quality of service and network performance in digital networks, including ISDN".

[i.8] "Resource Infrastructure in the NEXOF Reference Architecture".

[i.9] Web Service Level Agreement (WSLA) Language Specification, version 1.0 Revision wsla-2003/01/28, IBM Corporation.

[i.10] ETSI TR 102 997: "CLOUD; Initial analysis of standardization requirements for Cloud services".

[i.11] ETSI TR/CLOUD-0011-UserRec (work in progress).

[i.12] Cloud Standards Customer Council: "Practical Guide to Cloud Service Level Agreements", Version 1.0, April 10, 2012.

[i.13] NIST Special Publication 500-292: "IST Cloud Computing Reference Architecture".

3 Abbreviations

For the purposes of the present document, the following abbreviations apply:

CSCC	Cloud Standards Customer Council
EG	ETSI Guide
ETR	ETSI Technical Report
FG	Focus Group
IaaS	Infrastructure as a Service
IT	Information Technology
ITU-T	International Telecommunication Union - Telecommunication Standardization Sector
NA	Network Aspects
NEXOF	NESSI Open Framework
NIST	National Institute of Standards and Technology
NP	Network Performance
PDF	Portable Document Format
QoS	Quality of Service
SLA	Service Level Agreement
TC	ETSI Technical Committee
TR	Technical Report
WSLA	Web Service Level Agreement

4 Background

Service Level Agreements (SLAs) are widely used by telecommunications operators and service providers for a wide range of usage-based and subscription-based communication services. SLAs are well understood in this community and there is extensive experience of their structure, content and implication for both service providers and service consumers. They form the basis for contracts and define the expectations that a customer can have of the service to be provided and the remedies available in case of breach. SLAs are perhaps historically less common in IT, where service provision models are more recent features of the market. IT outsourcing is an exception to this but SLAs in these scenarios tend to be bespoke and designed to cover long-term transfers of operational responsibility from a customer to a third party provider.

Cloud represents a new market sector, combining IT functionality (e.g. computing, storage and data management) with wide-area networking and delivered as a set of services. Although still at an early stage in terms of adoption (and standardisation) there are many service offerings already available. These exhibit some commonality in their functional features, particularly for Infrastructure as a Service (IaaS), but the quality aspects and the terms and conditions applying to each provider's services can be complex and are defined and expressed differently. This puts the onus on potential customers to carefully analyse what is being offered, making comparison and selection of services appropriate to their needs difficult. Since the adoption of cloud services has potentially significant implications for operational and governance-related risk, lack of clarity on the details of the available service offers represents a barrier, particularly for applications with a long-term business importance to a user (as opposed to specific projects of limited scope and duration).

There is a need to define the role of the SLA in the business relationships between the various Cloud service stakeholders. Specifically, it is important to understand how each party can use the SLA to provide the context for their own decisions and operations. This is consistent with the aim of a recent guide from the Cloud Standards Customer Council ("Practical Guide to Cloud Service Level Agreements", Version 1.0, April 10, 2012) [i.12]. This is intended as "a practical reference to help enterprise IT and business decision makers as they analyse and consider service level agreements (SLA) from different cloud service providers". This document indicates the challenges faced by potential customers of cloud services and provides additional motivation for assessing the potential role of SLA standards for Cloud services. The aim is to encourage an open market in Cloud services and to reduce unnecessary barriers to adoption. This can include both general principles/definitions and specific metrics and processes.

The present document reviews existing ETSI standards for the quality of telecom services, including SLAs, in the context of the characteristics of cloud services, seeking to establish the extent to which they remain applicable and where new cloud-specific standards are required. The principal focus here is on services closely associated with physical computing, storage and networking resources (IaaS), although some of the discussion is also applicable to other cloud service models. This includes both general principles and definitions, where previous experience is highly relevant and specific metrics and processes, where additional work may be required.

5 Quality of Service (QoS)

The ETSI User Group (USER) has produced a multi-part deliverable covering the quality of telecom services. This ETSI Guide consists of three parts:

EG 202 009-1: "Methodology for identification of parameters relevant to the Users" [i.1].

EG 202 009-2: "User related parameters on a service specific basis" [i.2].

EG 202 009-3: "Template for Service Level Agreements (SLA)" [i.3].

5.1 General QoS characteristics

EG 202 009 [i.1] to [i.3] makes some assumptions about quality of service (QoS), which are generally applicable:

- QoS requirements are carefully identified from the user viewpoint so that a set of indicators and related reference values can be defined.
- Indicators can be measured and monitored with respect to these reference values to check whether the requirements are fulfilled.
- Reference values are well-defined (possibly in standards) and are included in the contract between the provider and the customer.

These assumptions remain valid for cloud services. Essentially they say that the performance of a service should be expressed in terms that the consumer of the service understands and can relate to his own processes (rather than being based on technical metrics from the provider domain), that assessment of performance is based on measurable parameters, that measurement protocols are clear to both provider and consumer, and that agreed performance against the defined parameters form part of the contractual relationship between provider and consumer.

Agreement on an appropriate assessment of the quality of a given service requires a series of actions:

- Analysis of specific user QoS requirements.
- Choice of the most appropriate indicators.
- Definition of the most suitable method of measurement and monitoring.
- Definition of the adequate indicator reference values.

These actions may be undertaken jointly by the provider and prospective consumer for bespoke or heavily customised services, while for more standard services they will be carried out by the provider and form part of the service offer. A common approach to assessing QoS, including standard metrics where possible, will allow potential users of a service to make meaningful comparisons without detailed individual analysis of each candidate service offering.

5.2 Content of services

EG 202 009 [i.1] to [i.3] makes a useful distinction between the delivery of the service itself (technical quality) and a number of additional provisions that make important contributions to the quality of the service. These additional provisions include information on the service, implementation and setup, problem resolution (alternative provision, repair/reinstatement of service), helpdesk, billing and accounting, statistics/reporting, updates, documentation, etc.

While cloud services will be functionally different from telecommunications services (and from one another), many of the additional provisions (or "service wrap") are likely to be applicable. It is clear that the quality of any service (communication or cloud) is influenced by the user experience through the full service lifecycle - from discovery and selection, through provisioning and usage to termination. The SLA has a role to play at each stage.

6 Cloud Services Market

6.1 Service characteristics and roles

It is important to understand the characteristics of cloud services and to compare them to the telecommunication services for which QoS and SLAs are already well understood. The National Institute of Standards and Technology (NIST) has identified five essential characteristics of cloud services which are generally well accepted [i.4]:

- On-demand self-service
- Broad network access
- Resource pooling
- Rapid elasticity
- Measured service

Most of these are also common characteristics of telecommunication services so we can expect the service relationship (as reflected in the SLA) to be similar.

Perhaps the most significant difference is the emphasis on on-demand self-service. This stipulates that the consumer be able to "unilaterally provision computing capabilities as needed, without requiring human interaction with each service provider". The implication is that negotiation of SLAs is less common and needs to be amenable to automation. This provides strong motivation for a machine-readable SLA, particularly as regards operational quality parameters (i.e. those that are directly monitored rather than indirectly assured).

In addition, the resource pooling characteristic extends similar considerations in telecommunication services. Cloud resource pooling refers to the service provider's use of a set of computing resources "to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand". This is similar to the sharing of network capacity in telecommunications services but involves computing resources (storage, processing, memory) as well as network. A cloud service SLA should specify what parts of a cloud service and its supporting infrastructure are shared since this can have implications for contention and fate sharing in the event of failures. Consistent ways of identifying and characterising multi-tenancy will be desirable.

Related to the issue of resource pooling is the degree of location independence associated with a cloud service. A provider of cloud infrastructure services can choose where to physically deploy its resource pools (based on, for example, cost of buildings and energy, network latency to target customers etc). Since cloud services typically involve the processing and storage of data, a consumer of cloud services may have a legitimate interest in specifying constraints on the location of the computing resources used. Legal, regulatory or business considerations may impose restrictions at the level of country, region or data centre. In the context of a cloud service SLA, this situation is similar to requirements for redundant physical paths in network services. However, constraints on physical resources are likely to be more commonly required by consumers of cloud services. Standard ways of representing resource location characteristics in SLAs are therefore desirable.

The market for Cloud services is still relatively immature. NIST has defined a Reference Architecture [i.13] that reflects the current position well. This identifies five distinct roles:

- Cloud Consumer - uses services from cloud providers
- Cloud Provider - makes services available to cloud consumers
- Cloud Carrier - provides connectivity and transport between cloud consumers and cloud providers
- Cloud Broker - manages the use and delivery of cloud services; negotiates relationships

- Cloud Auditor - conducts independent assessments of cloud services

There is the potential to introduce additional roles, and alternative reference models (e.g. ITU-T [i.5]) do so, but justification for this is not yet clear.

6.2 SLAs for Cloud services

The principal roles of interest from the point of view of Cloud service SLAs are Cloud Consumer, Cloud Provider and Cloud Carrier. Details of the Cloud Broker and Cloud Auditor roles are less well understood. For this reason they are not discussed in detail in the CSCC guide. We follow the same approach and focus specifically on the relationship between the Cloud Consumer and Cloud Provider. In addition, we make the simplifying assumption that SLAs are bipartite and independent. It is undoubtedly true that indirect relationships (e.g. those between Cloud Provider and Cloud Carriers - as described in the CSCC guide, Figure 1) may be relevant to the service experienced by a Cloud Consumer, but if a Cloud Consumer has a single business relationship (and SLA) with a Cloud Provider, how the Provider chooses to fulfil his obligations is not relevant to the Consumer.

It is important to distinguish between the content of the SLA, which is a representation of the context shared by the parties to the SLA (and defines a reference point), and the use that each party makes of this context in its own operations throughout the SLA lifecycle (publish offer/discover/select/agree/provision/use/terminate). The SLA needs to provide a complete description of the service for both Consumer and Provider to avoid uncertainty and disputes which benefit neither.

An unambiguous, clearly defined SLA has benefits to both Consumer and Provider:

- Clarity on service offer:
  - for the consumer, this allows comparison between different providers, including not using external cloud services at all (which is often the main choice to be made currently);
  - for the provider, this allows the scope and extent of a service offer to be clearly defined and used to promote services, comparing with other vendors or with in-house developments.
- Unambiguous definition of expectations and obligations on both sides:
  - for the consumer, this allows the impact on business processes to be determined, including any changes required to enable cloud services to be used effectively. This includes proper consideration of any governance, legal and regulatory considerations. Pricing should also be clearly defined so that return on investment can be determined;
  - for the provider, this allows services to be managed and prioritised effectively, particularly where many consumers are being served from a multi-tenanted infrastructure.
- Boundaries of liability:
  - clarity on liability for legal and regulatory compliance, consequential losses and remedies for failure to meet the SLA is important for both consumer and provider to manage their exposure to risk associated with the service contract.

The existence of well defined SLAs between Cloud Consumers and Cloud Providers clearly also supports the activities of Cloud Brokers and Cloud Auditors. They are able to use the context shared between Consumer and Provider to carry out their functions without being party to the SLA themselves.

7 QoS Assessment

7.1 Metrics, Measurement and Monitoring

ITU-T has defined a methodology for capturing the quality requirements of a user of communication services [i.6]. This Recommendation and ETR 003 [i.7] use a set of 7 general criteria (definitions from EG 202 009-1 [i.1]):

- Speed - performance criterion that describes the time interval required to perform a function or the rate at which the function is performed.
- Accuracy - faithfulness and completeness in carrying out the communication function with respect to a reference level.
- Availability - likelihood with which the relevant components of the service function can be accessed, at the instant of request, as required by the specified conditions, in particular those related to open hours, geographic coverage and resource size aspects if any.
- Reliability - ability of an item to perform a required function under stated conditions for a give time period.
- Security - ability of a service to ensure the confidentiality of the pieces of information worked out, exchanged or stored, the communication privacy, the authenticity and integrity of the information exchanged or stored as well as the protection of the user and his communication means against any type of threat.
- Simplicity.
- Flexibility - options required by the customer and offered by the provider in order to accommodate special requirements.

EG 202 009-1 [i.1] modifies this view by adding:

- Capability - ability to meet a demand of a given size under given internal conditions.
- Usability - effectiveness, efficiency and satisfaction with which specified users can achieve specified goals (tasks) in a particular environment. In telecommunications, usability should also include the concepts of learnability and flexibility; and reference to the interaction of more than one user with each other and with the terminals and the telecommunications system.
- Fidelity - as a supplement to accuracy.

Each of these generic aspects can be applied at different stages of the SLA lifecycle, and are applicable to cloud services at the level of detail of these definitions. They therefore remain useful dimensions along which to classify cloud services.

Terms in a SLA need to be quantifiable to support meaningful comparison between competing service offerings and to provide a basis for the determination of compliance. This leads to a requirement for a set of service metrics which provides a framework for measurement. The variation between different cloud service offerings, even in the IaaS service model, which is the most homogeneous of those identified by NIST, means that a standard set of universally applicable metrics is not appropriate. A flexible way of defining measurements in a consistent way is preferred.

The NEXOF-RA research project has developed a model of service metrics [i.8]. This has strong similarities with the Web Service Level Agreement (WSLA) specification [i.9] which focused on SLAs for web services. Adapting the concepts and terminology from these sources, a framework for organising the representation of technical terms suitable for use is SLAs can be defined. The components of this framework are:

- SLA Characteristic: identifies a high level aspect of the SLA which is of particular concern to the potential service consumer. It corresponds to a general property which can be described in terms of the measurable behaviour of the service (as distinct from those described in clause 7.2). This corresponds to the high level aspects presented above (e.g. Availability, Accuracy) and others, including some identified in [i.10].

- SLA Parameter: describes an observable property of a service whose value can be obtained from a source of measurement and which contributes to one or more SLA Characteristics. (An example might be Response Time). SLA Parameters should be understandable by service consumers, and potential consumers, and provide an operational definition of the SLA Characteristic in the context of a particular service.

- SLA Metric: defines a specific measurement methodology, including units, controlled conditions and measurement procedures. Each metric contributes to the evaluation of one or more SLA Parameters. SLA Metrics are essentially technical and provide a reproducible basis for the derivation of SLA Parameters. Service consumers may not have a clear understanding of the details which makes the specification and adoption of a set of standard metrics important in supporting comparison and monitoring of cloud services.

This model provides a useful framework and template for the definition of metrics.

EG 202 009-1 [i.1] provides some specific guidance to measurements of service quality which are generally applicable and naturally extensible to cloud services. In summary measurements are classified as either Objective or Subjective:

- Objective (intrusive - based on artificially generated traffic; non-intrusive - based on observing real traffic)

- Subjective (e.g. annual survey, possibly carried out by third party or following complaints)

The approach is to start with quality criteria derived from functional criteria and translate these to technical criteria, with associated metrics. At least one QoS parameter is necessary for an evaluation of each criterion. All criteria are needed for a comprehensive appraisal of a given service aspect. However a reduced set of carefully selected parameters may be used for QoS monitoring.

EG 202 009-1 [i.1] also identifies some principles for defining parameters and metrics which should serve as useful guidelines:

- beware mean values that may not reflect customer perception;

- focus on disturbance (should be 0) rather than performance (close to 100 %);

- use figures that may be consolidated (disturbance rate);

- define thresholds for quality of service targets.

Indicators and parameters to measure the QoS need to be usable by both users and providers, even if viewed from different perspectives. EG 202 009 [i.1] to [i.3] recognises that the results of monitoring are expected to be made publicly available for services offered to the general public. This is because it cannot be assumed that users will have their own measurement capability. This makes it particularly important that monitoring is based on standard parameters and metrics, with the potential for independent audit of service behaviour.

7.2 Assurance of other QoS aspects

Charging/billing, security, etc. are generally not assessable by monitoring of QoS parameters. Instead, quality is assured by approval of the systems and processes used by the provider and by adherence to published codes of practice. Verification/certification by a third party is desirable to provide confidence to consumers and to ensure consistency between providers.

8 SLA Template

A Service Level Agreement records the shared context associated with an instance of a service, agreed between its provider and a user. It is, or forms part of, the contract governing the relationship associated with use of the service. A general template for SLAs can make comparison of different service offerings easier and reduce the time taken by provider and consumer to understand and negotiate terms. This becomes particularly important where automated approaches to service discovery, negotiation and configuration are used. This is consistent with the "On-demand self service" characteristic of cloud computing.

EG 202 009-3 [i.3] proposes a general template for Service Level Agreements. This is split between general contractual terms, governance and operational procedures and service-specific issues, such as quality-related metrics.

It suggests identifying broad areas of agreement separately from details that can vary or be altered dynamically - i.e. quality metrics or flexibility points within the scope of the agreement. In practical terms, a cloud service consumer is likely to first satisfy himself that the characteristics of the service, its provider and broad principles of the SLA are acceptable. This can be done (manually) as part of normal due diligence processes. Deployment decisions based on service quality characteristics can then be made automatically, based on matching specific requirements to the defined characteristics of the service.

The proposed SLA template contains the following clauses:

- Content
- Technical features
- Geographical features, coverage
- Security aspects
- Duration of the agreement
- QoS commitments

This appears to be applicable to cloud services as well as to the communication services originally envisaged. The main point of difference is in the QoS commitments clause. This is where specific service metrics will be needed to describe cloud services. Standardisation here is desirable from the point of view of comparison (pre-commitment) and compliance auditing.

9 Discussion and Conclusion

The present document has analysed existing work on service level agreements (SLAs), mainly developed to address the requirements of communication services. Much of the existing work on SLAs can be reused for cloud services with little or no modification as it is driven more by the service-based business model than the specifics of the technical capabilities provided.

There is a need to develop a common vocabulary for describing cloud services. Consensus is beginning to emerge, particularly for the relatively homogeneous IaaS service mode but further work is needed even here. All SLAs should contain common features which result from the service provision business model. Specifications developed for telecom services (including by ETSI) can be extensively reused here. In addition, there is a need for a systematic, template approach to metrics, measurement and monitoring for cloud services. Some candidate approaches and methodologies have been identified in previous work and a proposal made in the present document for organising standard SLA parameters and metrics in a structured way. Cloud services are anticipated to be considerably more diverse than traditional telecom services. Specific metrics and parameters are required to describe cloud services in terms that are meaningful to service consumers. This has been investigated by a number of research projects, including those represented in the Framework 7 "QoS and SLAs Collaboration Working Group". Initial engagement between TC CLOUD and this initiative indicates a consistency of approach and between some of the technical solutions developed so far, although there is currently no consistently applied framework to promote reuse of results.

The present document proposes the use of a framework for service metrics, applicable to cloud services. This can be used to integrate specific developments coming (at least initially) from the research community. Priority should be given to those SLA Characteristics and Parameters which address the recommendations of potential users of cloud services.

As a first step, a set of SLA Characteristics of cloud services will be defined, closely aligned to the major areas of concern from a customer point of view, as identified, for example, in [i.10] and [i.11].

Each SLA Characteristic will then be defined according to a set of SLA Parameters appropriate to the specific service types identified. It is expected that this will be done first for Infrastructure as a Service (including network connectivity) as there is significant commonality between current market offerings. Platform as a Service and Software as a Service are more diverse and the decomposition of SLA Characteristics into SLA Parameters may currently only be feasible for some general service aspects.

The standardisation of SLA Metrics to allow the unambiguous evaluation of SLA Parameters will require technical consensus and robust specification so that results are reproducible and can support independent audit and certification of key aspects of service behaviour.

The present document recommends that SLA Characteristics and SLA Parameters should be specified in a single document, updated as required to reflect the development of the market for cloud services and the definition of SLA Metrics. SLA Metrics should be specified separately and linked to existing SLA Parameters. In this way, it will be possible to start with a top level concern for a particular type of cloud service and identify appropriate parameters and metrics that can express in a formal way the expectations of the behaviour of specific services - as required for an effective SLA.

History

Document history		
V1.1.1	November 2012	Publication