Key Performance Indicators for Cloud Computing SLAs

Stefan Frey, Claudia Lüthje, Christoph Reich
Furtwangen University
Cloud Research Lab
Furtwangen, Germany
{stefan.frey, claudia.luethje, christoph.reich}@hs-furtwangen.de

***Abstract*—Reducing IT costs by using cloud computing is tempting for many companies. As cloud rapidly is gaining momentum as alternative mean of providing IT ressources, the need for regulated service qualities increases. To attract companies to outsource their services to clouds, providers need to offer Service Level Objectives specified in SLAs for their customers. The content of such Service Level Objectives is a key reason for the successful usage of cloud computing and consists of Key Performance Indicators. Due to the dynamic character and complex nature of the cloud environment, creating SLAs for the cloud can be very difficult. This paper proposes selected KPIs for cloud SLAs and describes possible Service Level Objectives, as well as how they should be monitored.**

Keywords*—*Cloud Computing; KPI; SLA; QoS

I. INTRODUCTION

After an initial hype, cloud computing has established itself as adequate means of providing resources on demand. By now cloud computing provides a practical alternative to, locally hosted resources for companies. The main benefits of cloud computing are the cost savings through its "pay-per-use" model, low investment costs and its rapid implementation of innovations. According to a market analysis by the Gartner Group [1], the IT budgets of german companies has been reduced by 2.7% in 2011. The study also shows that companies will increasingly rely on outsourcing their IT to the cloud to save costs in the future. At present, most cloud computing providers only offer generic Service Level Agreements (SLA). Thereby guarantees for QoS characteristics like, bandwidth, data backup, etc. are given on the best-effort principal. Companies require QoS, monitoring and control of the cloud services at any time, as stated in the "Architecture of Managing Clouds" [2], Study Group Report of Cloud Computing [3], and others.

For cloud computing, the quality and reliability of the services become an important aspect, as customers have no direct influence on the services. Therefore Service Level Agreements are fundamental to an effective cloud utilization and especially business customers need them to ensure risks and service qualities are prevented respectively provided in the way they want. For this purpose, the expected service qualities are documented legally binding in contracts between provider and customer. Due to significant variation in consumer needs, SLAs have to be created individually by a negotiation process. The confirmed SLAs serve as a basis for compliance and monitoring of the QoS. Due to the dynamic cloud character, the QoS attributes must be monitored and managed consistently [4].

In order to describe the QoS, metrics and key performance indicators (KPI) are used. These must exactly represent the actual service expectations and requirements, and correspond to both customer as well as provider. In addition to this QoS attributes representation, an SLA includes a general section, in which roles and responsibilities, costs, etc. are listed. The aim of this paper is to propose various possible KPIs for cloud SLAs to facilitate an assist customers in the negotiation and generation of SLAs for cloud services. In addition, a general insight on SLA content and structure as well as monitoring and management is given. After discussing related work in Section II, Section III will give a brief introduction into SLA content und management. Following Section IV presents the Service Level Objectives for cloud computing and the corresponding KPIs. The conclusion is drawn in Section V.

II. RELATED WORK

As the usage of cloud service by companies continues to grow, the need for SLAs is increasing. NIST [5] has pointed out the necessity of SLAs, SLA management, definition of contracts, orientation of monitoring on Service Level Objects (SLOs) and how to enforce them. A basic discussion of SLA management and cloud architectures can be found in Service Level Agreements for Cloud Computing [6], but it is mainly concerned about SLA definitions and negotiations.

In recent years, a significant amount of research has been performed on the standardization and creation of machine-readable formats. There are two major specification for describing SLAs, WSAL [7] and WS-A [8]. The Web Service Agreement Language (WSAL) [7] was developed by IBM with the focus on performance and availability metrics. It has been mainly developed for Web services and the usage in other fields is questionable. It shows significant shortcomings regarding content as it was focused mainly on technical properties. WS-Agreement (WS-A) [8]. was developed by the Open Grid Forum in 2007. The newest update, which is based on the work of the European SLA@SOI project, was done in 2011. Although it has been enhanced within the SLA@SOI project [9], the development is unclear, because the SLA@SOI project developed its own format SLA(T), which is supported by the European IT industry.

Although much research has been done in the direction of SLA formats, the contents of SLAs remain a further field for investigations. The fact that SLAs are always very scenario specific makes it difficult to generalize their contents. KPIs,

 ISBN: 978-1-61208-292-9

as a central component of service level objectives, are increasingly offered in KPI libraries [10]. However, these are mosty of rudimentary content and are not suitable for implementation.

III. SLA

Service Level Agreements (SLAs) specify the promised respectively the expected performance characteristics between service providers and customers. Thereby, all legally relevant information and services are established. The most important part of a SLA is the exact description of the service quality (service level). The following section illustrates the prerequisites, content and structure of Service Level Agreements.

The creation of Service Level Agreements provides certain requirements to customers and providers. Customers need to be able to meet certain requirements in order to successfully define SLAs, which are listed briefly here. A customer must:

- Understand the roles and responsibilities that are regulated by the SLA.
- Be able to describe precisely and specific the service to be controlled by the SLA.
- Know the requirements of the controlled services, and define the matching key figures.
- Specify service levels based on the critical performance characteristics of the service.
- Understand the process and procedures of regulated service.

These requirements are necessary so that the customer is able to put in the correct SLAs values, and to understand implications of his decisions. Furthermore, a SLA should fulfill the following tasks:

- Describe the services accurately.
- Specify the service quality to be provided in detail.
- Describe detailed the key performance indicators, metrics and service levels.
- Breakdown transparently all the costs.

A. SLA Life Cycle

The life cycle of a service level agreement involves several steps for a successful use of SLAs [11]. There are different views on whether the definition phase of the SLA is one of its life cycle or not, since this can also be counted among the preconditions (see [12] and [13]). Figure 1 shows the SLA life cycle. The individual phases are briefly described:

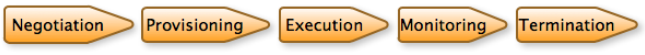

Fig. 1: SLA Life Cycle

The preconditions for this life cycle is the definition of an initial SLA template based on which the negotiation phase is started. In the negotiation phase, the deliverable and service levels and the costs are negotiated with the provider. While in the provisioning phase, the entry into force of the agreement is marked by the signatures of both partners. Here, the provided services are provisioned and the agreements are communicated and fitted into the organizations. During the execution phase the customer uses the service according to his notions. Parallel to this, the monitoring phase the runtime data is checked and assessed against the service levels. If needed, corrective actions are executed and reports and documentation are created for the partners. The final termination phase marks the end of the usage by the customer and initiates the decommission of the service.

B. SLA Content

The structure of service level agreements are generally very scenario specific and can not be easily generalized. However, there are some basic elements that should be present in every SLA. The following remarks are not intended to be used to create an universal pattern for SLAs, but rather give a guideline for most current contents of SLAs.

The contents of a SLA can be divided into the following four categories: (see [14]) *agreement-related elements, service-related elements, document-related elements and management-related elements.*

The agreement-related elements contain the basic rules of the agreement and include, among others, the subject of SLAs, objectives, partners, as well as the scope, entry into force, duration and termination of SLAs. Often these elements are shown in practice in the form of a preamble or introduction. The subject of SLAs introduction here describes the content and context as well as a description and demarcation of the services being controlled by the SLA. The objectives of the SLAs reflect the specific objectives of both parties and serve, among other things, as a basis for future success control.

The service-related elements represent those elements which describe the regulation of a service. These must be specified individually for each service. The content is basically to describe who, when, where, and what services are provided. The description of the service should be generally understandable. The description of the quality of a service is the central role of the SLA. The negotiated quality of service is defined by Key Performance Indicators (KPIs), which is the basis for the Service Level Objectives" (SLOs). These indicators include a label next to the calculation or metric, and a reference area and measurement point. Similarly here, the cost of services to be provided are defined.

Document-related elements include administrative and editorial elements, which play a minor role inside a SLA and are mainly there to improve the handling, understanding and readability. These elements are, e.g., version, the date of last modification, revision history, table of contents, the index or glossary. These elements increase the readability by underpinning the context and explain the background.

The management-related elements include the aspects that have to do with the administration and control of SLAs. These represent a very important section of the contents of a SLA, since both the customer notification and the procedure in case of problems or failures to meet the service levels are regulated.

 ISBN: 978-1-61208-292-9

Furthermore, penalties and compensation in case of damage which may occur due to deviations from service levels are regulated.

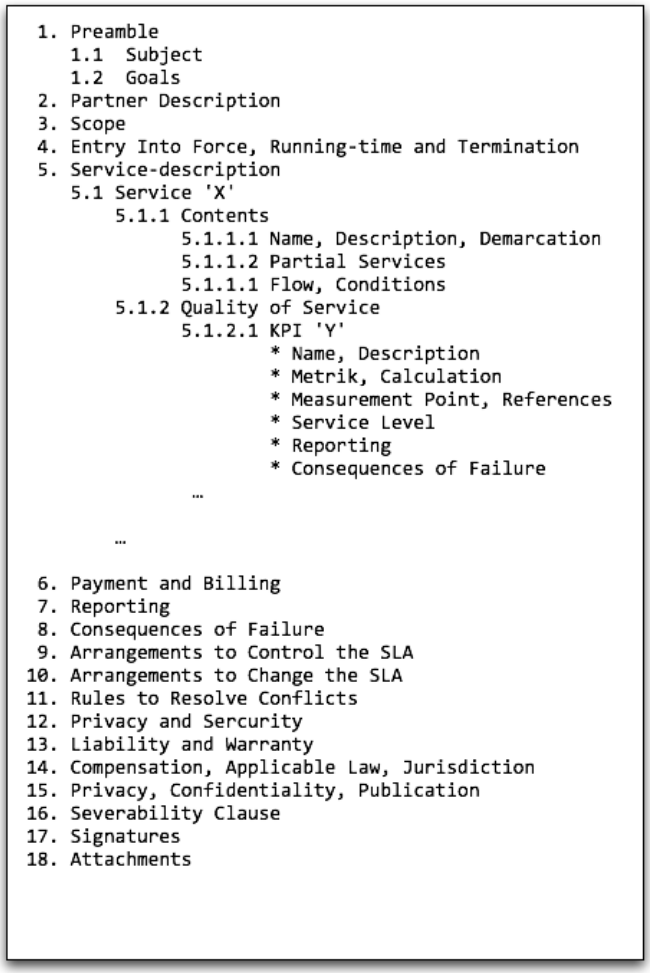

Fig. 2: SLA Structure

Based on the presented elements, an exemplary structure of an SLA can be created. This can be seen in Figure 2 above. Here, it is clear that the service descriptions, or service level objectives are the central aspect of each SLA. These and their contents are described in more detail in the following sections. Likewise, it comes clear that even small SLAs mean large administrative overhead and the creation is a lot of work.

IV. Service Level Objectives

Service Level Objectives (SLOs) are a central element of every service level agreements (SLA), which include the negotiated service qualities (service level) and the corresponding Key Performance Indicators. SLOs contain the specific and measurable properties of the service, such as availability, throughput or response time and often consist of combined or composed attributes. SLOs should thereby have the following characteristics: [15]

- Achievable / attainable
- Repeatable
- Measurable
- Understandable
- Significant
- Controllable
- Affordable
- Mutually acceptable
- Influential

A SLOs should always contain a target value or service level, a metric and corresponding measurement period, as well as the type and location of the measurement. For this purpose, KPIs with associated service level values are stated. The KPIs contain information about the measurement process, place and unit as well. A valid SLO specification might, for instance, look like this: *The IT system should achieve an availability of 98% over the measurement period of one month. The availability represents thereby the ratio of the time in which the service works with a response time of less than 100ms plus the planned downtime to the total service time, measured at the server itself.* From such a description, the actual performance values can be compared with the reference values of the SLOs and the achievement is calculated. Based on this, further measures can carried out to for correction if necessary.

To choose the correct KPIs for a service a wide knowledge of the service and its usage is required. To give an insight into possible cloud-specific KPIs, the most common ones are listed briefly below without going into much detail. The following KPIs provide specifically for cloud computing selected guarantees but also may overlap in part with traditional KPIs, as the essential services requirements do not differ from other general services [16].

A. General Service KPIs

Service Level Agreements must always be tailored to the service to be controlled. Nevertheless, there are some KPIs, which rules can be used in various SLA. These KPIs represent the basic needs of each service to run efficiently. These include, for example the availability, security aspects, service times and helpdesk, as well as monitoring and reporting. These are basic requirements for every purchased service.

1) Basic Services: The basic services include the availability which is defined at the time the service is usable + the maintenance time relative to total time. Deemed usable here is if the system can handle request within a specified response time. Also included are the KPIs Mean Time Between Failure and Mean Time To Repair, which specify the time intervals at which to expect failures and how long it takes to repair them.

2) Security: Security KPIs regulate for example which software version levels shall be used, how long it should take until an update is implemented, as well as the scope and frequency of security audits. Other important KPIs control the encryption of data, the use and timeliness of anti virus software and the isolation and logging.

 ISBN: 978-1-61208-292-9

3) Service and Helpdesk: Service and Helpdesk KPI control including the times at which assistance is provided, which support methods are applied or how many calls are received per week. Similarly, the qualification of the support personnel and the duration is given to problem solving.

4) Monitoring: Monitoring KPIs to define in which values are determined intervals to monitor and how to handle the resulting reports. The arrangements of these KPIs can be reused in the other categories.

B. Network Service KPIs

Particularly for cloud computing, the network has a strong meaning, as all provided resources and services are available through a network. Here, the network has to be considered both as pure transmission medium for other services as well as independent service itself. For the KPIs described here, the entry point of the provider network is usually chosen as measured point, as the guarantees of the provider refer only to this area.

Round Trip Time: Time of a network packet to travel from sender to receiver and back. Specifies how long the transmission of one packet needs within the network limits. Usually measured in milliseconds.

Response Time: Time taken by a request until the arrival of the response at the requesting interface. Here the time for the processing of the request is included as opposed to the pure orbital period of the round trip time. The type of the request and the behavior of the processing has to be concretely defined for this.

Packet Loss: Percentage of lost packets in the total of transmissions. Formula:

$$\frac{Number \quad of \quad lost \quad packets}{Number \quad of \quad total \quad packets \quad * \quad 100} \tag{1}$$

The value of this indicator should kept as low as possible since for example an a loss rate of 5% to 10% significantly affects the quality of VoIP applications [17]

Bandwidth: Gross capacity of the connection. Amount of data which cloud be transmitted within a time unit. Here, not the actual capacity is specified but the rated maximum capacity.

Throughput: Number of transmitted data per time unit. Only the pure transmitted data is taken into account, thus the capacity available to the user is specified. Measured in Mbit/s or / Gbit/s

Network Utilization: Proportion of the throughput to the bandwidth. Here, it can be seen how busy the connection is. Formula:

$$\frac{Throughput}{Bandwidth \quad * \quad 100} \tag{2}$$

Latency: Time interval between submitting a packet and arrival at its destination. Is usually considered together with *Jitter:* The difference in the latency of a packet and the average / minimum / maximum run time. The run time variations are problematic especially in real-time applications, since packages may arrive too late or too early.

C. Cloud Storage KPIs

The term storage can be distinguished within cloud computing in two basic types. First, Storage as a service itself, that is obtained as a memory for preexisting infrastructures. On the other hand storage can be used as part of another service such as a backup or data storage for cloud services.

Response Time: Time interval between sending a request to the storage and the arrival of the response at the output interface. Usually measured in milliseconds.

Throughput: Number of transmitted data per time unit. Here, a specified amount of data is transferred to the storage and measured the needed time from a given point. The size of the data set and package sizes are important factors for the validity of this measure. Furthermore, the network and its utilization must be considered.

Average Read Speed: In contrast to the throughput, the average reading speed usually refers to an individual hard drive. This value indicates how fast data can be read from the hardware. In RAID systems or virtual storage solutions, this figure is expected to interconnected hard drives.

Average Write Speed: Just like the reading speed it refers to the write speed to the hard drive. This value thus indicates how quickly data can be written from a source to the hardware.

Random Input / Outputs per second (IOPS): Number of possible random input / output operations per second for different block sizes. The higher the IOPS value, the faster the disk. This value is also important to measure how many concurrent accesses can be handled by the system.

Sequential Input / Outputs per second (IOPS): Number of possible sequential input / output operations per second for different block sizes.

Free Disk Space Usable free capacity in % of the total capacity or remaining free space in MB, GB, or TB. This indicator can be very useful since thus it can be defined how much memory must always be at minimum available on the system.

Provisioning Type Type of provisioning where at "thin provisioning" the client gets the storage not permanently assigned but it is dynamically allocated at runtime. In contrast, the thick-provisioned storage is allocated to the customer immediately.

Average Provisioning Time Time, the provider needs to provide a defined amount of data volume growth.

D. Backup and Restore KPIs

Backup and Restore KPIs refer to both the storage, i.e., the stored data, as well as services, for example, VMs or SaaS services. Below, important KPIs are presented.

Backup Interval The time interval in which a backup is performed. Here, an exact specification is given to the provider along with the backup type and a description of the scoop.

Backup Type Definition of the backup type, e.g., full backup or incremental backup. Backup types ca relate to individual systems or whole service alliances.

 ISBN: 978-1-61208-292-9

Time To Recovery Specification of the minimum and maximum time from the failure of a storage, to the successful restore from an existing backup.

Backup Media Specifying the media where the backups are stored, such as magnetic tapes. Indication of media breaks to store backups on different media types.

Backup Archive Interval and number of archived backups. Specification of when backups are archived and how long they are kept and how these are to be terminated.

E. Infrastructure as a Service KPIs

Infrastructure as a Service refers not only to the service itself but also to the virtual machines used. For this, additional VM KPIs are specified in this section.

VM CPUs Number and type of CPUs used by the virtual machine. Additionally information about the overbooking of the provided CPU resources shall be given. Here the shared resources are allocated with more capacity than is physically available. Thus, no real physical allocation of resources takes place. Actual performance is dependent on the overall consumption of the system.

CPU Utilization Proportion of CPU resources in use to the total number of resources provided per time unit. Also the CPU queue, which indicates the number of open requests to the CPU should be considered.

VM Memory Amount and type of the provided memory. This may relate to physical memory or virtual memory. Information about the overbooking of allocated memory resources should be stated.

Memory Utilization Proportion of the memory resources used to the total amount of memory made available to the VM.

Minimum Number of VMs Guaranteed number of the provided VMs with the specified specs stated in the previous points.

Migration Time Time that is needed to move a VM from two predefined resources.

Migration Interruption Time Maximum time in which a customer has no access to migration to the resource.

Logging Retention of log data. Specifies how long log data to be stored by the provider and specification of what level to be logged. (e.g., INFO, DEBUG, etc.)

V. Conclusion

The paper pointed out both the general content and specific KPIs for the creation of cloud SLAs. Thus, cloud user have now th basis for the creation of cloud SLAs. Since this is only a general overview of the contents of cloud SLAs the details and designs have to be discussed further. Particularly, in the area of measurement of the KPIs further research is needed.

Acknowledgment

This research is supported by the German Federal Ministry of Education and Research (BMBF) through the research grant number 03FH046PX2.

References

[1] Gartner Group, “Cio-prioritäten und budgets 2011.” [Online]. Available: http://www.cio.de/strategien/analysen/2262709/ [retrirved: june, 2013].

[2] Distributed Management Task Force, “Architecture for managing clouds.” [Online]. Available: http://dmtf.org [retrieved: june, 2013].

[3] ISO/IEC SC 38 Study Group, “Jtc 1/sc 38 study group report on cloud computing,” International Organization for Standardization, Tech. Rep., 2011. [Online]. Available: http://isotc.iso.org [retrirved: june, 2013].

[4] A. Keller and H. Ludwig, “The wsla framework: Specifying and monitoring service level agreements for web services,” Journal of Network and Systems Management, vol. 11, no. 1, pp. 57–81, Mar. 2003.

[5] F. Liu, J. Tong, J. Mao, R. Bohn, J. Messina, L. Badger, and D. Leaf, “Nist cloud computing reference architecture,” NIST special publication, vol. 500, p. 292, 2011.

[6] J. Happe, W. Theilmann, A. Edmonds, and K. Kearney, Service Level Agreements for Cloud Computing. Springer-Verlag, 2011, ch. A Reference Architecture for Multi-Level SLA Management, pp. 13–26.

[7] H. Ludwig, A. Keller, A. Dan, R. P. King, and R. Franck, “Web Service Level Agreement (WSLA) Language Specification, v1.0,” Jan. 2003. [Online]. Available: http://www.research.ibm.com/wsla/WSLASpecV1-20030128.pdf [retrirved: may, 2013].

[8] K. T. Kearney, F. Torelli, and C. Kotsokalis, “Sla*: An abstract syntax for service level agreements,” 11th IEEE/ACM International Conference on Grid Computing, pp. 217–224, 2011.

[9] SLA@SOI. SLA@SOI projekt website. http://sla-at-soi.eu/. [retrirved: june, 2013].

[10] MIRROR-42, “Kpi library.” [Online]. Available: http://mirror42.com [retrieved: june, 2013].

[11] W. Sun, Y. Xu, and F. Liu, “The role of xml in service level agreements management,” in Services Systems and Services Management, 2005. Proceedings of ICSSSM ’05. 2005 International Conference on, vol. 2, 2005, pp. 1118–1120.

[12] P. Hasselmeyer, B. Koller, I. Kotsiopoulos, D. Kuo, and M. Parkin, “Negotiating slas with dynamic pricing policies,” Proceedings of the SOC@ Inside07, 2007.

[13] G. R. Gangadharan, G. Frankova, and V. D’Andrea, “Service license life cycle,” in Collaborative Technologies and Systems, 2007. CTS 2007. International Symposium on, 2007, pp. 150–158.

[14] T. G. Berger, Konzeption und Management von Service-Level-Agreements fur IT-Dienstleistungen. TU Darmstadt, 2005.

[15] R. Sturm, W. Morris, and M. Jander, Foundations of Service Level Management, ser. Sams Professionals. Pearson Sams, 2000, iSBN: 978-0-6723-1743-9.

[16] S. Ran, “A model for web services discovery with qos,” SIGecom Exch., vol. 4, no. 1, pp. 1–10, Mar. 2003.

[17] K. C. Mansfield and J. L. Antonakos, Computer Networking from LANs to WANs: Hardware, Software, and Security. Boston: Course Technology, Cengage Learning, 2010, iSBN: 9781743044544.

 ISBN: 978-1-61208-292-9