Towards Measuring the Degree of Fulfillment of Service Level Agreements

Frank Schulz
SAP Research Karlsruhe
SAP AG
Karlsruhe, Germany
frank.schulz@sap.com

Abstract— In service oriented architectures (SOA), the non-functional properties of services have been recognized to be highly important in addition to the functionality of services as a means to differentiate services according to quality considerations. Service level agreements (SLAs) are formalized contracts between service providers and service consumers that are used to define quality of service (QoS) properties. The violation of an SLA by the service provider typically results in a penalty to compensate the service consumer. In order to avoid such situations, the service provider needs to recognize critical service instances and to take appropriate countermeasures before a violation happens. Therefore a measure for quantifying the danger of SLA violation is needed as part of a service level management system. This paper proposes a concept for the definition and evaluation of such a metric that takes into account the underlying structure of the SLA as well as the available options for monitoring service quality parameters. Hence it becomes possible to obtain detailed information of the status of service fulfillment at runtime and to identify critical service instances. The methodology is exemplified with the availability property.

Keywords- service oriented architecture, service level agreement; quality of service; key performance indicator; monitoring

I. INTRODUCTION

Service oriented architectures have emerged as a highly successful approach to build modular software systems [1]. In their realization with Web Services, they form the conceptual basis for many e-commerce solutions and for innovative approaches to deliver software applications to customers, like Software-as-a-Service (SaaS) and On-Demand computing. Similarly, the combination of existing services to new value-added services leads to the formation of agile service networks that represent new forms of collaboration between companies or institutions. In business contexts, where consumers of such services rely on them to operate their own business, the quality of the service offers becomes vital [2].

Service Level Agreements are the common approach to define levels of service quality. An SLA constitutes a formal agreement negotiated between a service provider and a service consumer about the consumption of a service. It contains a definition of the service (e.g. by linking to a WSDL description of the service), the duties of the involved parties, the agreed guarantees, and optionally additional information like penalties in case of failure to comply with the defined service level objectives.

Before accepting a Service Level Agreement, the service provider has to ensure that sufficient capabilities for fulfilling the agreement will be available. As soon as a Service Level Agreement has been negotiated and accepted and the corresponding service is operational, the service provider needs to guarantee that service delivery complies with the agreement. Therefore a continuous monitoring of all services at runtime is needed and critical services have to be discovered quickly. The management of a service provider’s infrastructure with the help of SLAs is an active research area [3,4].

Currently there are two main approaches for the management of service operations at runtime. One direction is based on the monitoring of SLA violations [5]. In case of a SLA violation, strategies like the provisioning of additional resources can be applied in order to remedy the situation. This approach is based on the binary information whether a given SLA is fulfilled or violated. The second direction uses low-level monitoring for assuring proper operation of system resources [6]. It results in detailed key performance indicators (KPIs); however the connection to consumer requirements via SLAs is missing.

This paper aims at filling this gap. By quantifying the degree of SLA fulfillment, it becomes possible to obtain a fine-granular status of all service instances and their corresponding consumer relations. Thus, the service landscape can be closely monitored on the level of SLAs, and critical service instances with imminent danger of SLA violation can be identified. This information is needed as the basis for further decisions within the service level management. Possible actions include the provisioning of additional resources or the cancellation of some service instances with low penalty.

The main contributions of this paper are (1) a classification of key performance indicators with respect to aggregation properties, (2) a definition of the degree of fulfillment based on aggregated KPIs and (3) a method for combining these atomic data according to the logical structure of the SLA.

Section 2 discusses related work. In section 3, technical foundations for our approach are presented, and section 4 contains the details of the new concept. Section 5 concludes with a summary and outlook.
II. RELATED WORK

In current literature there seems to be no assessment of SLA fulfillment degree or SLA violation danger with the only exception of [7]. In that work, the authors propose to characterize a SLA as green (i.e. all SLA conditions are fulfilled), yellow (SLA conditions are met, but “indicators come near to the prescribed thresholds”) or red (SLA conditions are not satisfied). Based on this definition, the SLA violation danger is defined as ratio of cumulated time of service operation with yellow SLA and cumulated time with yellow or green SLA. No further details of the approach are given, in particular the KPIs within the SLA are not considered. Our work can be seen as an independent approach to the definition of SLA fulfillment that takes into account the internal structure of SLAs and provides concrete details for the calculation of a SLA fulfillment degree.

Another related concept, though with a different goal, is introduced in [8] with the notion of soft probabilistic contracts for allowing flexibility in SLA compliance. In [9] the authors complement their approach with a statistical method for soft contract monitoring.

III. PRELIMINARIES

For working with a concrete SLA structure, the WS-Agreement specification is used. It defines a formal syntax that provides a structure for any SLA document, whereas domain-specific KPIs for the intended application have to be included at a lower level of the SLA document.

A. WS-Agreement

While several proposals for formalizing service level agreements exist, e.g. [10,11], the most popular approach seems to be the WS-Agreement specification [12]. Among its advantages are the high adoption in practice, the framework structure and extensibility, and the fact that it represents an open standard for both SLA documents and SLA negotiation. WS-Agreement defines the high level format of an SLA document and allows for the inclusion of domain-specific KPIs and SLOs.

The high level structure of an SLA document according to WS-Agreement is shown in Fig. 1. The Service Terms and Guarantee Terms contain detail information; both are recursively defined with a Term Compositor construction. The Service Terms describe the service sufficiently well, e.g. by including or linking to a WSDL document or parts of it. In addition they can define variables of the service that can be monitored. Most important for this paper are the Guarantee Terms. Again, a Term Compositor structure allows for a logical combination of several Guarantee Terms via AND, NOT and XOR operators, and thus for constructing a logical formula. Each Guarantee Term contains a Service Level Objective (SLO) which is a predicate over variables defined in the Service Terms part of the WS-Agreement document. Based on the observed values of these variables during operation of the service, the Service Level Objective is evaluated; it can be fulfilled or violated. The variables can either be considered as key performance indicators (KPIs) of the service and directly compared with KPI target values, or they form the basis for constructing custom level predicates. In addition, a Guarantee Term also includes a set of Business Values, among them a Penalty term. The Penalty contains an Assessment Interval which can either be a time period or a number of service invocations, and the amount of compensation in case of a violation of the service level objective. Compensation can be monetary or otherwise, for example extended usage of the service free of charge. Other components of the Guarantee Term structure like preconditions for the validity of the Service Level Objective are of minor importance to the topic of this paper and not further discussed.

B. Key Performance Indicators

As an example, three common key performance indicators for services are presented. They will also be used in the next section.

1) Availability: The availability of a service within a given time interval can be defined as the ratio between the sum of durations during which the service could be invoked and the total duration of the time interval. For example, when during a one hour interval the service could be called all the time except for two periods of three minutes each, the downtime was \((3+3) / 60 = 10\%\) and hence the availability was 90%.

2) Response Time: The response time of a single service invocation can be defined as the time duration between service request and service response. Depending whether network latency is taken into account, the time on consumer side between sending the service request and receiving the response, or the time on provider side between receiving the service request and constructing the response can be considered. Typical service level objective for performance are defined for individual service invocations (e.g. completion time of each service call less than 2 seconds) or
for a certain quantile (e.g., for 90% of all service invocations within one month, completion time less than 2 seconds).

3) **Throughput**: The throughput or capacity of a service is the number of service invocations that can be processed by the system within a given time interval. For example, a service may respond up to 100 service requests within one hour.

### C. Properties of KPIs

As described above, a Service Level Objective can be regarded as a logical predicate over some variables related to the service. The variables themselves can be related to a single service invocation, or they can represent an aggregation over several service calls. With respect to the examples above, availability and throughput only make sense as aggregated KPIs based on a sequence of service requests. In contrast, response time can be measured both as single invocation or aggregated service property. Because a property being defined for a single invocation can be mapped to an aggregated property, only aggregated KPIs are considered in the remainder of this work.

When aggregation takes place, the duration of an assessment interval has to be specified. The complete service operation time is then divided into subsequent assessment periods. At the beginning of each assessment period, the KPI aggregation is initialized, and at the end of the assessment period, the Service Level Objective can be evaluated. The innovation of this paper is the evaluation of the aggregated KPI inside the assessment period and calculation of an appropriate degree of fulfillment.

In this work, only numerical KPIs or arithmetic expressions over numerical KPIs that can be compared with a numerical target value are considered. Arbitrary custom level predicates (which are allowed in WS-Agreement) are excluded as long as they cannot be mapped to a numerical indicator. The reason for this restriction becomes clear in the next section.

### IV. DEGREES OF SLA FULFILLMENT

This section gives details of the approach proposed in this paper. It is based on three main steps. First, a fulfillment function is used to map actual KPI values and target values to a degree of fulfillment. Second, SLO fulfillments are aggregated based on the logical structure of the SLA. Third, the degree of fulfillment may be mapped into categories for reasons of simplification.

#### A. From KPIs to degree of SLO fulfillment

For an SLO of the form ‘KPI ≤ KPI\_Target’ with target KPI\_Target as the threshold value at the end of the assessment interval, and an assessment interval of duration T, let the function KPI(t) represent the cumulated indicator value for the time from start of the assessment interval until time t. Then the degree of SLO fulfillment at time t is defined as

\[
SLO(t) = \begin{cases} 
1 & \text{for } T \cdot \text{KPI}(t) \leq t \cdot \text{KPI}_{\text{Target}} \\
0 & \text{for } \text{KPI}(t) > \text{KPI}_{\text{Target}} \\
1 - \text{KPI}(t)/\text{KPI}_{\text{Target}} + t/T & \text{otherwise.}
\end{cases}
\]

A central idea is the construction of a linear target function from the threshold value KPI\_Target, namely \(t/T \cdot \text{KPI}_{\text{Target}}\). As long as the current value \(\text{KPI}(t)\) is below this target function value, a complete fulfillment of the SLO is assumed (case 1). As soon as the current value \(\text{KPI}(t)\) is larger than the target value at the end of the assessment interval, the SLO will definitely be violated, thus the fulfillment is 0 (case 2). When \(\text{KPI}(t)\) is larger than the target function, but less than the final target value, the SLO becomes critical because the linear prediction of the trend indicates a violation at the end of the assessment interval. The criticality is quantified by the difference between the ratio of KPIs and the ratio of time (case 3).

The idea is illustrated with figure 2. Assume that an SLO specifies a service availability of at least 55 minutes during assessment intervals of 1 hour duration. Hence the threshold value for a violation is \(\text{KPI}_{\text{Target}} = 5\) minutes. The current value \(\text{KPI}(t)\) is the cumulated time of non-availability within each assessment interval. In period 1, the service becomes unavailable for 3 minutes at time \(t_1 = 20\) min. Then the current SLO fulfillment degree is \(1 - 3/5 + 20/60 = 73.3\%\). As no further unavailability occurs, the SLO fulfillment reaches 100\% again at time \(t_2 = 36\) min. In assessment period 2, the unavailability of the service for 6 minutes, i.e., more than the KPI threshold of 5 minutes, results in a SLO fulfillment of 0\% at time \(t_3\), meaning SLO violation in this period.

![Figure 2: SLO fulfillment degree](image)
Three extensions are straightforward: first, an SLO of the form \( \text{KPI} \geq \text{KPI}_{\text{target}} \) can be handled in a similar way, second, an arithmetic expression over several KPIs can be used instead of a single KPI, and third, the linear target function with respect to time can be replaced. For example, a danger of SLA violation early within the assessment period may be considered more critical than the same danger near the end of the assessment period.

B. Aggregation of SLO fulfillments to SLA fulfillment

When the SLA guarantee term is a logical expression over several SLO predicates, the expression immediately translates into a formula for calculating the degree of SLA fulfillment. This transformation is achieved by the following mappings from logical operators to arithmetic operators:

- \( \text{AND} \rightarrow \min(\cdot) \)
- \( \text{OR} \rightarrow \max(\cdot) \)
- \( \text{NOT} \rightarrow \text{complement function } 1 - \text{SLO}(t) \).

Hence the degree of SLA fulfillment \( \text{SLA}(t) \) is defined as an arithmetic expression over the \( \text{SLO}(t) \) terms. For example, an SLA of the form \( \text{Availability} > 99\% (\text{SLO}_1) \) AND Average Response Time < 2 seconds (\( \text{SLO}_2 \)) results in the SLA fulfillment function

\[
\text{SLA}(t) = \min(\text{SLO}_1(t), \text{SLO}_2(t)).
\]

An important property of this approach is the fact that the current SLA fulfillment degree can be evaluated at any time \( t \) for any combination of SLO predicates defined in the SLA. It is not required that the assessment intervals of different SLO are the same or multiples of each other. This reflects the direct connection of the proposed SLA fulfillment degree with the penalties due in case of a SLA violation.

C. SLA fulfillment categories

The SLA fulfillment defined above is expressed as a number in the interval \([0,1]\), where 1 means complete fulfillment and 0 represents violation. For visualization purposes or reporting reasons, SLA fulfillment degrees can be clustered into categories. For example, a mapping may be applied where \( \text{SLA}(t) = 0 \) maps to 'red', \( 0 < \text{SLA}(t) < 1 \) maps to 'yellow' and \( \text{SLA}(t) = 1 \) maps to 'green'.

V. CONCLUSION

The degree of SLA fulfillment or the danger of SLA violation, respectively, can give important information during service operation that opens new options for service level management systems. This paper presents a unified approach for defining such a metric based on aggregated KPIs where the logical structure of the SLA guarantee terms is taken into account.

Future work will consider an extension of the presented approach with respect to dependencies between KPIs, and an experimental evaluation of the prediction of SLA violations based on the degree of SLA fulfillment. Also a dynamic visualization with drill-down capabilities to discover the relevant KPIs of a critical SLA is a relevant topic for further exploration.

ACKNOWLEDGMENT

This research was partially funded by the German Federal Ministry of Education and Research (BMBF) under promotional reference 01G09004.

REFERENCES