MANAGING VIOLATIONS IN SERVICE LEVEL AGREEMENTS

Omer Rana
School of Computer Science/Welsh eScience Centre
Cardiff University, UK
o.f.rana@cs.cardiff.ac.uk

Martijn Warnier, Thomas B. Quillinan and Frances Brazier
Department of Computer Science,
VU University Amsterdam, The Netherlands
warnier@cs.vu.nl
tb.quillinan@few.vu.nl
frances@cs.vu.nl

Dana Cojocarasu
Norwegian Research Center for Computers and Law
University of Oslo, Norway
d.i.cojocarasu@jus.uio.no

Abstract A Service Level Agreement (SLA) represents an agreement between a service user and a provider in the context of a particular service provision. SLAs contain Quality of Service properties that must be maintained by a provider. These are generally defined as a set of Service Level Objectives (SLOs). These properties need to be measurable and must be monitored during the provision of the service that has been agreed in the SLA. The SLA must also contain a set of penalty clauses specifying what happens when service providers fail to deliver the pre-agreed quality. Although significant work exists on how SLOs may be specified and monitored, not much work has focused on actually identifying how SLOs may be impacted by the choice of specific penalty clauses. The participation of a trusted mediator may be necessary to resolve conflicts between involved parties. The main focus of the paper is on identifying particular penalty clauses that can be associated with an SLA.

Keywords: Service Level Agreements, Violations, Penalty Clauses, WS-Agreement

1. Introduction

A Service Level Agreement (SLA) represents an agreement between a client and a provider in the context of a particular service provision. SLAs may be between two parties, for instance, a single client and a single provider, or between multiple parties, for example, a single client and multiple providers. SLAs generally specify performance related properties, generally referred to as Quality of Service (QoS) terms, that must be maintained by a provider during service provision. These properties need to be measurable and must be monitored during the provision of the service that has been agreed in the SLA – and are referred to as Service Level Objectives (SLOs). The SLA must also contain a set of penalty clauses when service providers fail to deliver the pre-agreed quality. Although significant work exists on how SLOs may be specified and monitored [10], not much work has focused on actually identifying how SLOs may be impacted by the choice of specific penalty clauses. The participation of a trusted mediator may be necessary to resolve conflicts between involved parties. Automating this conflict resolution process clearly provides substantial benefits. Different outcomes are possible. These include monetary penalties, impact on potential future agreements between the parties and the enforced re-running of the agreed service. While it may seem reasonable to penalise SLA non-compliance, there are a number of concerns when issuing such penalties. For example, consider a service provider violation in a multi-provider SLA: determining whether the service provider is the only party that should be penalised, or determining the type of penalty that are applied to each party would be required. Enforcement in the various legal systems of different countries can be tackled through stipulating a 'choice of law clause', that is, a clause indicating expressly which countries' laws will be applied in case a conflict between the provider and the client would occur. Specific 'legal templates' [4] can be used to further refine such clauses. This paper focuses on identifying particular penalty clauses that can be associated with an SLA and on identifying how penalty clauses impact the choice of SLOs. The next section discusses the types of violations that can be used in SLAs. Section 3 discusses the type of penalties that can be used. An example from resource sharing in an electronic market (based on work in the CATNETs project [8]) is presented in Section 4 and a mapping to the WS-Agreement specification is proposed in Section 5. The paper ends with discussions and conclusions.

2. Types of Violations

An SLA can go through a number of stages once it has been specified. Assuming that the SLA is initiated by a client application, these stages include: discovering providers; defining the SLA; agreeing on the terms of the SLA (in addition to the penalties if the SLOs are not met); monitoring SLA violations;

terminating an SLA; enforcement of penalties for SLA violation. Monitoring plays an important role in determining whether an SLA has been violated, and determining the particular penalty clause that should be invoked as a consequence.

Monitoring SLA violations begins once an SLA has been defined. A copy of the SLA must be maintained by both the client and the provider. It is necessary to distinguish between an 'agreement date' (forming of an SLA) and an 'effective date' (subsequently providing a service based on the SLOs that have been agreed). For instance, a request to invoke a service based on the SLOs may be undertaken at a time much later than when the SLOs were agreed. During provision it is necessary to determine whether the terms agreed in the SLA have been complied with during provision. In this context, a monitoring infrastructure is used to identify the difference between the agreed upon SLO and the value that was actually delivered during service provisioning – which is 'trusted' by both the client and the provider.

From a legal perspective, monitoring is a prerequisite for contract enforcement. In the present context, the consequences of breaching the agreed SLOs is a basic requirement. In addition, service clients base the reputations of, and their trust in, service providers largely on the supported monitoring infrastructure. In the context of SLAs three types of monitoring infrastructures can be distinguished: a trusted third party (TTP); a trusted module at the service provider; and a module on the client site. In most typical situations a TTP module provides all the necessary functionality for a monitoring service.

One of the main issues that the provider and the consumer will have to agree during the SLA negotiation is the penalty scheme. It is also necessary to define what constitutes a violation. Depending on the importance of the violated SLO and/or the consequences of the violation, the provider in breach may avoid dispatch or obtain a diminished monetary sanction from the client. As both the service provider and the client are ultimately businesses (rather than consumers), they are free to decide what kind of sanctions they will associate to the various types of SLA breaches, in accordance with the importance of the SLO that was not fulfilled. According to the Principles of European Contract Law [3], the term 'unfulfilment' is to be interpreted as comprising: (1) defective performance (parameter monitored at lower level); (2) late performance (service provided at the appropriate level but with unjustified delays); (3) no performance (service not provided at all). Based on these descriptions we define the following broad categories:

- 'All-or-nothing' provisioning: provisioning of a service meets all the SLOs – that is, all of the SLO constraints must be satisfied for a successful delivery of a service;

- 'Partial' provisioning: provisioning of a service meets some of the SLOs – that is, some of the SLO constraints must be satisfied for a successful delivery of a service;

- 'Weighted Partial' provisioning: provision of a service meets SLOs that have a weighting greater than a threshold (identified by the client).

Monitoring can be used to detect whether an SLA has been violated. Typically such violations result in a complete failure – making SLA violations an 'all-or-nothing' process. In such an event a completely new SLA needs to be negotiated, possibly with another service provider, which requires additional effort on both the client and the service provider. Based on this all-or-nothing approach, it is necessary for the provider to satisfy all of the SLOs. This equates to a conjunction of SLO terms. An SLA may contain several SLOs, where some (for example, at least two CPUs) may be more important than others (for example, more then 100 MB hard disk space). During the SLA negotiation phase, the importance of the different SLOs for the client must be established. Clients (and service providers) can then react differently according to the importance of the violated SLO. In the WS-Agreement specification [1], the importance of particular terms is captured through the use of a 'Business Value'. Weighted metrics can also be used to ensure a flexible and fair sanctionatory mechanism in case an SLA violation occurs. Thus, instead of terminating the SLA altogether it might be possible to renegotiate, for example, with the same service provider, the part of the SLA that is violated. Again, the more important the violated SLO, the more difficult it will be to renegotiate (part of) the SLA.

3. Penalties

The use of penalty clauses in SLAs leads to two concerns: what types of penalty clauses can be used; and how, if at all, can these be included in SLAs. The 'burden of proof' and the interest in demonstrating that the agreed SLOs have been violated lie with the main beneficiary of the service, that is, in the service client. An important issue that should be considered when designing 'penalty schemes' is that behind the imposition of any contractual sanctions lies the idea that faulty behaviour of a provider should be deterred. As such, it is always possible for the service provider to contest its liability in the unwanted result (SLA breach) and claim that a 'force majeure' situation occurred. Although the situation is impossible to be dealt with through automatic enforcement, monitoring the message exchanges among the provider and the client can indicate whether the SLA violation was the consequence of a 'misconduct' from the provider (either intentional or negligent). The parties are advised to stipulate either in the SLA or in the associated Collaboration Agreement how they choose to deal with the situation where the provider's

faulty behaviour cannot be documented, and a 'force majeure' situation did occur. A penalty clause in an SLA may consist of the following:

- a decrease in the agreed payment for using the service, that is, a direct financial sanction;
- a reduction in price to the consumer, along with additional compensation for any subsequent interaction;
- a reduction in the future usage of the provider's service by the consumer;
- a decrease in the reputation of the provider – and subsequent propagation of this value to other clients.

During the negotiation phase, client and provider can agree on a direct financial sanction. Usually, the amount to be paid depends on the value of the loss suffered by the client through the violation (that should be covered entirely) and if agreed, a fix sum of money that has to be paid as 'fine' for the unwanted behaviour. Due to the potential difficulties in proving and documenting the financial value of the loss, during the negotiation phase the parties may choose an 'agreed payment for non performance', that is, a fixed sum of money that will have to be paid upon non-performance, regardless of the fact that no financial loss was suffered by the client. The service provider can deposit the negotiated fine in escrow with a TTP, who acts as a mediator, before the service provision commences. Escrow is a bond, deed, deposit, etc., kept in the custody of a third party, taking effect, or made available, only when a specified condition has been fulfilled[1]. On successful completion of the service provision (based on the SLA) the TTP returns the deposit to the service provider. Otherwise, the client receives the deposit as compensation for the SLA violation. Notice that a trusted monitor is required for this, as a client can never prove by itself that an SLA was (partially) violated. For automated use, a micro-payment [7] system is required – such as Paypal. Another possibility is that a client reduces its usage of services from a provider that violated an SLA. If the economic position of the client is strong enough, this can be a valid strategy. A third kind of penalty clause can lead to a change in the reputation of a provider [9, 12]. In such a system the reputation of service providers that violate SLAs will drop. In this case special care needs to be taken that the reputation of a service provider is correctly determined. Both reputation building, using dummy clients that 'praise' a service provider, and slandering reputations, where dummy clients (unjustly) complain about a service provider, form serious threats in reputation based systems. In the negotiation phase of the

[1] from *Concise Oxford English Dictionary, Revised 10 Edition.*

SLA, both service provider and client can agree on the reputation mechanism to use.

4. Resource Sharing Use Case

Consider a market of computational service providers, where each provider may use a combination of resources to meet a particular set of metrics of interest to a client. In a service market, the parameters of interest may be of three types: latency (time it takes to get a result back from the provider), execution time (total time it takes to execute a service at the provider), and execution cost (the monetary value associated with running a service by the provider). The resources (R) that may be used by the provider are defined as a four tuple – consisting of: number of CPUs (C), primary memory (M), disk storage (D), and time interval (δt) – δt represents the interval between the start time and the end time over which the resource is available. A resource provider is required to define their capacity using these four parameters. Generally a client does not care what resources are used, as long as their application performance constraints are met. Conversely, a service provider needs to identify which resources need to be used to achieve these metrics. Two types of SLAs co-exist in this scenario – an SLA between an application client and the service provider, and an SLA between a service provider and one or more resource owners. In this example we use only four parameters to characterise access to a resource – however this model can be expanded to include additional attributes that have been specified within the Common Information Model (CIM) [6]from DMTF.

The SLA between the service provider and the resource owner may be defined using the terms: $(C, M, D, \delta t) = R$ – and may be offered by a single provider, or it may be the aggregate capability of a group of providers. Properties of each R_i are published in a registry service—the resource owner being responsible for updating these values in the registry. The registry may also contain an aggregate resource description, describing the combined capability of multiple providers. After having discovered a provider to interact with, a client asks the provider for an SLA template. The template contains those parameters that the provider understands and can monitor. Depending on the type of description scheme being used, the client now adds constraints associated with parameters that have been identified in the SLA. This 'offer' is now sent to the provider—who may either agree with the request, or make a counter offer. A negotiation process is initiated, which eventually results in either an agreement or a failure. An example of an SLA in this context would be: $SLA_1 = (2, 512MB, 2GB, (20071001190000), (20071001191000))$— indicating a request for a resource with 2 CPUs, 512MB of RAM, 2GB of disk on October 1, 2007 from 19:00 to 19:10. Such a scenario also occurs in many data centre applications today [13].

The SLA between the client and the service provider is often harder to specify, as it can contain application specific terms as SLOs. As outlined in [5], given an SLO of 'average response time' to be less than 10 seconds, the configuration with CPU assignment of 20% fails to meet the SLO, but a CPU assignment of 90% meets the SLO but the system is over-provisioned (as only 50% is needed to meet the SLO). Therefore, identifying the types of provisioning that is needed to ensure that the SLO is not violated, but that excessive resources are not used to address a particular SLO requirement is important. A mapping is needed between the requirements identified in an SLA between a client and a service provider, and one between a service provider and a resource.

It is necessary when specifying an SLO to also specify the penalty that would be incurred by a provider if the SLO was not met. Often a gradual structure of penalties is defined, whereby SLO violations incur fines, and a certain number of violations within a particular time period (such as a week or a month), gives a client the right to terminate access to the service. A penalty identifies the compensation that would be made to a service client if the SLO has been violated. Examples of penalty clauses that may be associated with an SLA between a service provider and a resource owner may be as follows [2]:

- If 90% of the number of requested CPUs, and 90% of requested memory have been delivered, then these SLOs have not been violated. For provisioning below 90% of CPU and memory, and for each percent, the provider must incur a penalty of α monetary units.

- If 90% of the number of requested CPUs and 90% of the requested RAM and 80% of the requested disk have not been delivered, then for each deviation from 90% (for CPU and RAM) and 80% for disk, the penalty to the provider is β monetary units.

For an SLA between a client and a service provider, a service execution time may be used as the SLO, then the penalty clause would be written as:

- If 90% of the execution times are not in the 2 second range, then for each deviation from the 98% of between 2 and 5 seconds, the penalty to the provider is β monetary units, and for each percent of the 98% of execution times more than 5 seconds, the penalty is γ, and for other percents that are more than 5 seconds, the penalty is α monetary units.

A service provider must evaluate the penalty it would incur from the client if a resource owner was not able to achieve their SLOs.

5. Mapping to WS-Agreement

The WS-Agreement specification [1] provides an XML schema to represent the top-level structure of an agreement between two parties. This includes concepts such as an agreement identifier, guarantee terms in an agreement etc. A simple protocol is provided which allows offers, acceptance and rejection of an agreement to also be captured. An 'Agreement Factory' is used as an interface to create a new instance of an agreement, with the use of 'creation constraints' as an optional description of the types of agreements that a provider is willing to accept. An important factor in this discussion is the use of the 'Business Value' (BV) and 'Preference' specification made available in WS-Agreement. A BV allows a provider to assess the importance of a given SLO to a client. Similarly, a provider may indicate to a client the confidence that a provider has in meeting a particular SLO. Based on the specification, a BV may be expressed using a `penalty` or `reward` type. The penalty is used to indicate the likely compensation that will be required of a provider if the SLO with which the penalty is associated is not met. We may weigh the importance of an SLO with reference to other SLOs that constitute an agreement. Notice that a BV list consists of both a penalty *and* a reward – to enable a provider to assess the risk/benefit of violating a particular SLO. `Preference` is used in the BV list to provide a more detailed sub-division of a business value for different alternatives that may exist. Essentially, `Preference` allows a service provider to consider different possible alternatives for reaching the same overall SLO requirement. For instance, in the example of section 4, if a client requests access to a particular number of CPUs, it is possible to fulfil this requirement based on CPUs from one or more resource owners. `Preference` allows the provider to chose between the available options to improve its own revenue or meet other constraints that it has (provided this is not prohibited by the service provision agreement or other agreements between the parties involved).

A `Penalty` in WS-Agreement may be associated with one or more SLOs, and occurs when these SLO(s) are violated. According to the WS-Agreement specification, assessment of a violation needs to be monitored over an `AssessmentInterval` – which is defined either as a time interval or some integer count. Essentially, this means that a penalty can only be imposed if an SLO is violated within a particular time window, or if a certain number of service requests/accesses fail. `ValueUnit` identifies the type of penalty – in this case a monetary value – that must be incurred by the service provider if the violation occurs. In the current WS-Agreement specification, the concept of a `ValueExpr` is vague – being an integer, float or a 'user defined expression'. This implies that a user and provider may determine a dynamic formula that dictates the penalty amount depending on the particular context in which the WS-Agreement is being used.

```
<wsag:Penalty>
    <wsag:AssesmentInterval>
        <wsag:TimeInterval>xs:duration</wsag:TimeInterval> |
        <wsag:Count>xs:positiveInteger</wsag:Count>
    </wsag:AssesmentInterval>
    <wsag:ValueUnit>xs:string</wsag:ValueUnit>
    <wsag:ValueExpr>xs:any</wsag:ValueExpr>
</wsag:Penalty>
```

In WS-Agreement the ability to also specify a `Reward`, in addition to a penalty, provides an incentive mechanism for a provider to meet the SLO. Based on the example in Section 4, a penalty clause for the SLA between the client and the service provider would be as indicated below – specifying that four incorrect invocations of a service would lead to a penalty of $500.

```
<wsag:AssesmentInterval>
     <wsag:Count>4</wsag:Count>
 </wsag:AssesmentInterval>
<wsag:ValueUnit>US Dollar</wsag:ValueUnit>
<wsag:ValueExpr>500</wsag:ValueExpr>
```

The extent to which terms and conditions specified in WS-Agreements are legally binding is currently the subject of research [4]. One basic element is that agreements need to be confirmed by both parties. As such, penalties in a WS-Agreement, for example, cannot be one-sided. The WS-Agreements needs to be confirmed by the client. The lack of this confirmation makes WS-Agreement restricted in the context of legal perspective, as explored by Mobach et al. [11].

6. Discussion & Conclusions

The use of penalties in SLAs has obvious benefits for both clients and service providers. Monetary sanctions and reputation-based mechanisms can both be used as, pre-agreed, penalties. It has been shown how the WS-Agreement specification can be used to specify penalties and rewards, in the context of a particular resource sharing scenario.

A particular focus has been discussion of the types of violations that can occur in SLOs during provisioning. Based on European legal contract law, we identify three types of violations that may lead to penalties – an 'all or nothing', 'a partial' or a 'weighted partial' violation of a contract. An observation in this work is that flagging a violations incurs a cost for the client (as well as the provider). It is therefore in the interest of the client to continue with service provision, even if some of the SLOs are not being observed fully – a trade-off discussed in this paper. A key contribution of this work is a model that demonstrates how a client may provide weighting to certain SLOs over others, the legal basis on which this model is based (as outlined in Section 3) and subsequently how this approach can be used alongside WS-Agreement.

Acknowledgements: This research is in part supported by the NLnet Foundation, `http://www.nlnet.nl` and in part is funded by the NWO TOKEN program. Part of this work is also supported by the European Commission Future and Emerging Technologies programme under the IST-2006-027004 "S3MS" and the IST-FP6-003769 "CATNETS" projects.

References

[1] A. Andrieux, K. Czajkowski, A. Dan, K. Keahey, H. Ludwig, T. Nakata, J. Pruyne, J. Rofrano, S. Tuecke, and M. Xu. Web Services Agreement Specification (WS-Agreement). *GRAAP Working Group at the Open Grid Forum*, September 2006.

[2] ARAD Automatic Real-time Decision-making, 2002. available at:`http://www.haifa.il.ibm.com/projects/software/arad/papers/ARAD-May-2%002.pdf`.

[3] M. J. Bonell. The UNIDROIT Principles of International Commercial Contracts and the Principles of European Contract Law: Similar Rules for the Same Purposes? , 1996.

[4] M. Boonk, F. Brazier, D. de Groot, M. van Stekelenburg, A. Oskamp, and M. Warnier. Conditions for Access and Use of Legal Document Retrieval Web Services. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Law (ICAIL'07)*. ACM Press, 2007.

[5] Y. Chen, S. Iyer, X. Liu, D. Milojicic, and A. Sahai. SLA Decomposition: Translating Service Level Objectives to System Level Thresholds. *HPL-2007-17*, 2007.

[6] DMTF, "Common Information Model". See Web site at: `http://www.dmtf.org/standards/cim/`. Last accessed: June 2007.

[7] R. Hauser, M. Steiner, and M. Waidner. *Micro-payments Based on IKP*. IBM TJ Watson Research Center, 1996.

[8] L. Joita, O. F. Rana, P. Chacin, I. Chao, F. Freitag, L. Navarro, and O. Ardaiz. Application Deployment on Catallactic Grid Middleware. *IEEE DS-Online*, 7(12), 2006.

[9] S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina. The eigentrust algorithm for reputation management in p2p networks. In *Proc. of the 12th Int. World Wide Web Conference*, Budapest, Hungary, May 20-24 2003. ACM Press.

[10] A. Keller and H. Ludwig. The WSLA Framework: Specifying and Monitoring Service Level Agreements for Web Services. *Journal of Network and Systems Management*, 11(1):57–81, 2003.

[11] D. G. A. Mobach, B. J. Overeinder, and F. M. T. Brazier. A WS-Agreement Based Resource Negotiation Framework for Mobile Agents. *Scalable Computing: Practice and Experience*, 7(1):23–36, 2006.

[12] P. Resnick, K. Kuwabara, R. Zeckhauser, and E. Friedman. Reputation systems. *Communications of the ACM*, 43(12):45–48, 2000.

[13] E. Wustenhoff. Service Level Agreement in the Data Center. *Sun Microsystems Professional Series*, April 2002.