Towards Management of SLA-Aware Business Processes Based on Key Performance Indicators

Branimir Wetzstein, Dimka Karastoyanova, Frank Leymann

Institute of Architecture of Application Systems, University of Stuttgart, Germany
{firstname.lastname}@iaas.uni-stuttgart.de

Abstract. It is increasingly important that Service Level Agreements (SLAs) are taken into account when business processes are exposed as services in a Service Oriented Architecture. SLAs define expected service behavior and non-functional properties of the service. The fact that the service provider has to offer certain guarantees concerning SLA properties has an impact on the business process lifecycle. In this paper we introduce a stepwise approach for management of SLA-aware service compositions based on process performance requirements specified as Key Performance Indicators. The approach is based on the process lifecycle known from Business Process Management and comprises a modeling, configuration and execution phase. We incorporate existing work on SLA modeling, QoS aggregation, and QoS-based service selection, and identify several problems specific to SLA-aware business processes.

1 Introduction

Service Oriented Architecture (SOA) allows realizing business processes by recursively aggregating services into orchestrations [1]. The business process is then itself exposed as a service to consumers. The standard language for service orchestration based on Web services is BPEL [2].

An important aspect in the business process lifecycle is management of the performance of business processes. Performance requirements on business processes are specified as Key Performance Indicators (KPIs) with target values which are to be achieved in a certain analysis period. Typical KPIs are process duration, process cost, but also domain dependent business process metrics such as “number of purchase orders processed in full and on time”. KPIs are monitored at process execution time using Business Activity Monitoring technology.

In addition to KPIs, it is increasingly important, for instance in business process outsourcing scenarios, that Service Level Agreements (SLA) are taken into account when designing processes that are exposed as services to customers. When a business process is exposed as a service to customers, the provider has to specify non-functional properties of the service it exposes, in particular its technical QoS characteristics such as response time, throughput and availability. In order to define expected service behavior and quality of the interaction with regard to the non-functional properties, service provider and service consumer create a contract called

SLA. An SLA defines service level parameters (e.g. average response time) and service level objectives (SLOs) which are guarantees of concrete values of service level parameters (e.g. average response time < 5 min) [3]. In addition, the SLA can define penalties in case of violations. While KPIs pose requirements on the performance of the business process in general, SLOs specify constraints only on those KPIs which are exposed and offered to the service consumer.

Before the service provider can offer certain SLOs to the service consumer, he has to know or at least be able to estimate the performance characteristics of his service. When a service stands for a business process implemented as a service orchestration its performance depends on the SLOs of the orchestrated services. The overall SLOs of the business process can be computed by aggregating the SLOs of the orchestrated services. In addition, the IT infrastructure on which the process is deployed has certain QoS properties which have to be taken into account when calculating the overall SLOs of processes.

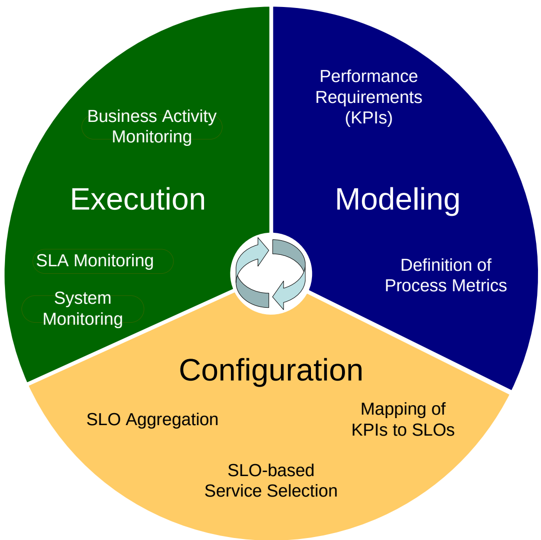

Figure 1: SLA-aware Business Process Lifecycle

In this paper, we introduce a stepwise approach towards management of SLA-aware service compositions based on performance requirements specified as KPIs. The approach consisting of three major phases (Figure 1): (i) modeling, (ii) configuration and (iii) execution. In the modeling phase the performance requirements on the process are gathered and KPIs with target values are specified. In the configuration phase KPIs are mapped to dependent SLOs of partner services and IT infrastructure. Appropriate partner services and IT infrastructure are selected and the overall SLOs of the process are calculated and assigned to its service interface. After process deployment, customers can discover the processes and negotiate and agree on concrete SLOs. During process execution SLAs are monitored. Overall process performance is monitored using Business Activity Monitoring (BAM) technology.

The rest of the paper is organized as follows. In Section 2 we describe how performance requirements are specified based on KPIs. In Section 3, we then show how KPIs are mapped to lower level SLOs of partner services and IT infrastructure. Section 4 deals with monitoring of KPIs and SLAs at process runtime. Section 5 summarizes the paper and outlines our future work.

2 Specifying Performance Requirements

The first step of our approach is the specification of business process performance requirements. This specification is typically performed by business people based on business goals in a certain time period. When specifying performance requirements for a business process with service levels one can distinguish between:

– *SLO requirements*: These are SLOs we want our service composition to be able to offer to the service consumers. Since SLOs are guarantees on values of SLA parameters, the first step in this phase is to identify the SLA parameters we want to offer. SLA parameters can be based on technical QoS metrics like for example availability and throughput [4, 5], but also on domain-specific business process performance metrics (so called key performance indicators (KPIs)), e.g. "percentage of purchase orders which can be processed on time and in full" (see, e.g., SCOR [6] for examples of KPIs in the supply-chain domain) or "percentage of incidents resolved on time and without callback" within the IT service management domain. After definition of SLA parameters, SLO requirements, i.e. target values of SLA parameters, are defined based on how the service provider wants to position himself compared to competition (e.g. higher availability and better response time but more expensive service), or they can be predefined based on already existing needs of customers, policies and regulations.
– *Internal business process performance requirements*: These are internal requirements towards business process performance based on time, cost, and quality perspectives that are considered important for the business performance of an enterprise as a whole, and are not exposed to process customers. Rather they are meant for internal use and performance optimization in the enterprise. Some internal KPIs may coincide with externalized SLOs (as in the two KPI examples above), or may be defined to measure the performance of particular SLOs ("number of SLO violations", "monetary loss resulting from SLO violations").

Every KPI definition contains a target value or a target value range (e.g. average process duration < 5 minutes), which should be achieved in a certain time period. This target value is what drives the next phases of the presented approach and is monitored later during process execution time.

After specifying the performance requirements as KPIs, corresponding process metrics are defined based on business processes, thus setting the basis for monitoring of KPIs and detection of SLA violations at process runtime. Process metrics (Figure 2) can be defined based on one process instance, or on multiple process instances, using average, max, min operators and others. Composite metrics can recursively be defined based on existing metrics (e.g., "number of successful purchase orders / number of all purchase orders"). We can define metrics based on process duration, number of instances that executed a concrete alternative path in a process model ("number of loan approvals without risk assessment"), or business object states ("number of products which were processed in time and without rework"). Related work on process performance measurement can be found in [7]; [8] gives an example on how process metric modeling for BAM can be supported.

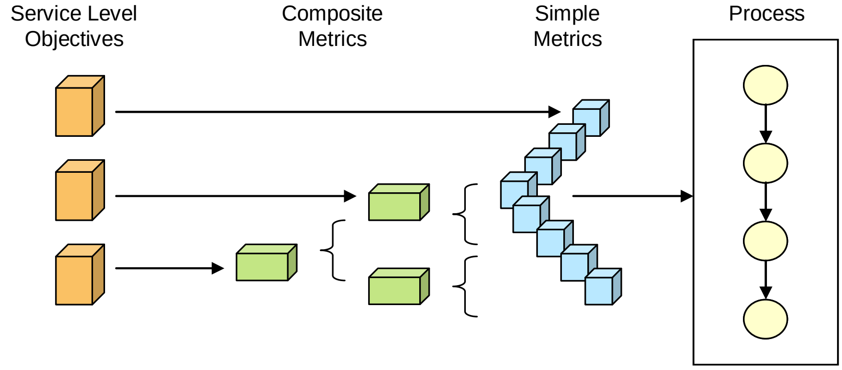

Figure 2: Process Metrics as a basis for SLOs

When defining SLA parameters based on process metrics, we can associate them to either the WSDL description of the process or, if available, an abstract BPEL definition that specifies the behavioral interface of the process. Abstract BPEL process descriptions are used in case a service provider is unwilling to disclose his complete executable business process definition to customers. An abstract BPEL process would be particularly important if there is a multi-step interaction between service consumer and service provider (Figure 3); in that case we could, for example, specify for both consumer and provider how long a certain step should maximally take. This is an example of bilateral obligations in an SLA as the service consumer also has to provide certain SLOs.

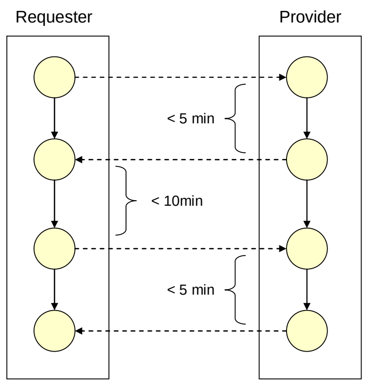

Figure 3: Bilateral obligations in multi-step interactions

3 Mapping KPIs to SLOs of Services and IT Infrastructure

In the process configuration phase we need to ensure that the service composition can actually meet the performance requirements imposed on the business process during the previous phase. This involves: (i) understanding the dependencies between KPIs of the process and performance characteristics of orchestrated partner services and used IT infrastructure; (ii) selection of partner services and corresponding SLAs in such a way that after aggregating their values, KPI target values are achievable; (iii) calculation of SLOs the provider can offer to service consumers.

3.1 Influence Factors on the Performance of a Service Composition

There are several influence factors on the performance of a business process implemented as a service orchestration.

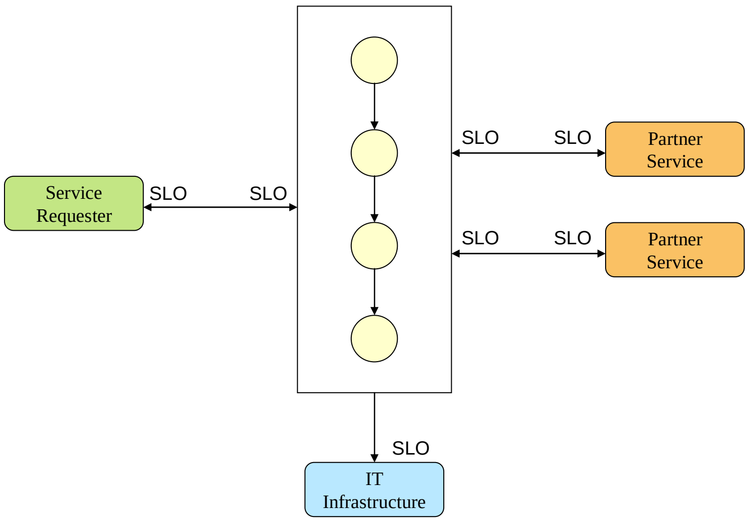

Figure 4: Influence Factors on the Performance of a Service Composition

As shown in Figure 4, following aspects have to be taken into account:

– The process is an orchestration of partner services, which influence the performance of the process. We assume that these services expose SLOs themselves.
– The process is deployed on a process engine (e.g. a BPEL engine) which runs on a concrete IT infrastructure (application server, databases, preconfigured hardware etc.) This infrastructure has properties that influence the availability, throughput, and performance of the process.
– The process is itself exposed as service to a service requester. It thus has to provide SLOs to the requester. These SLOs are dependent on SLOs of partner services and the IT infrastructure. The process can request certain SLOs from service requester, too, e.g. in case that the process has to call back the service

requester during a multi-step interaction (represented by the bidirectional arrows in Figure 4, see also Figure 3). In that case the duration of the requester process would have an influence on the overall business process duration of the provider and thus would have to be considered in the SLA.

3.2 Mapping KPIs to Service Level Parameters

The first step of the configuration phase is to map KPIs to service level parameters of partner services and IT infrastructure. The goal is to find out which service level parameters influence the performance of the KPI. At this point, we are not yet interested in the concrete SLO, i.e. a guaranteed value concerning the service level parameter. Depending on the type of the metric and the process, the KPI can be mapped to different service level parameter types and different partner services and/or IT infrastructure can be selected. A detailed examination is out of scope of this paper; we will describe this issue on an example.

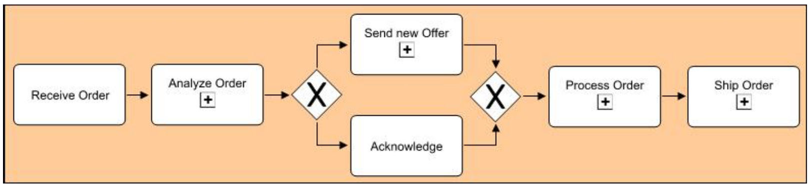

Figure 5: Purchase Order Process

Figure 5 shows a business process for processing purchase orders. After receiving an order from the customer, the order is analyzed and it is checked whether all of the order items can be delivered in full and on time. This step involves checking whether the order items are already available in the warehouse or, if not, contacting corresponding suppliers. If the order can be processed without changes an acknowledgement is sent to the customer, otherwise a new offer is sent and negotiated. Finally, the order is processed and submitted to the shipment service.

A possible KPI for that process is "% of orders delivered in full and on time > 90%", "in full" denoting that all order items should be delivered. Figure 6 shows how that KPI could be mapped to SLA parameters of orchestrated services.

The condition "in full" can be satisfied when either a supplier is able to deliver the needed items on time or they are already available in the warehouse.

The condition "on time" is mapped to the duration of the process until shipment is received by the customer. This duration can be further mapped to "delivery time" and "shipment time" which are guaranteed by SLAs with supplier and shipment service, respectively. In addition, the duration of internal activities, which also consist of several service invocations, has to be taken into account (not further detailed in Figure 6). Finally, the duration is also dependent on the availability of the IT infrastructure which the process runs on.

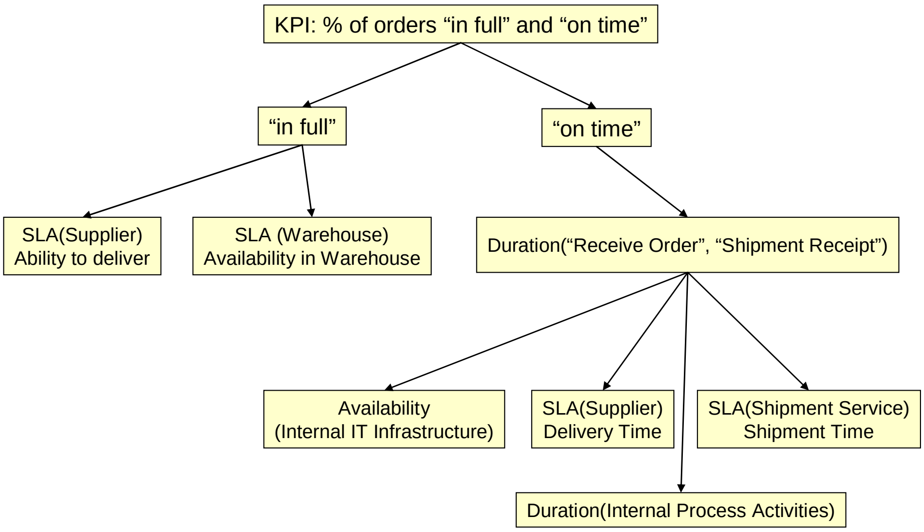

Figure 6: Mapping of KPIs to Service Level Parameters

3.3 SLO Aggregation

By mapping KPIs on service level parameters, the dependent metrics which have an influence on the KPI are identified. In the next step, it has to be ensured that the target value of the KPI can be achieved when aggregating concrete SLOs offered by partner services and IT infrastructure. Thus, one has to select partner services and IT infrastructure, aggregate their individual SLOs, and compare the aggregated value with the target value of the KPI. In the purchase order example above, for example, one would have to aggregate the expected durations of all the services in the process to come up with the overall duration of the process which can then be compared with a target value for the "on time" condition.

When it comes to QoS-based service selection, several papers [9, 10] describe how to find an optimal selection of functionally equivalent partner services based on their QoS properties and under consideration of global constraints (target values) in a service composition. QoS-based selection uses QoS aggregation [9, 11, 12] to calculate the overall QoS of the service composition by aggregating the QoS of orchestrated services. In [12] an approach is described on how SLOs, in particular, can be aggregated in BPEL processes.

QoS aggregation as part of QoS-based service selection mostly addresses only QoS properties when interacting with partner services. It does not deal with the duration of other process activities (e.g., assign activity in BPEL); it also neglects the time which is taken by the BPEL engine software to navigate through the process model. In particular in short-lasting processes, these issues have to be taken into account [13]. The IT infrastructure consisting of both hardware (single-node-server, cluster of servers, network) and software (application server, BPEL engine, database) influences in particular response time, throughput and availability of the process. If the IT infrastructure is predefined, the goal is to estimate QoS properties of the process (performance prediction) [14], for example, by using benchmarking techniques. Otherwise, one could use capacity planning techniques to estimate needed IT infrastructure based on performance requirements (capacity planning) [15].

For the overall SLO calculation we have to take both partner services and the IT infrastructure into account. However, from case to case some or all of these parameters of the computation can be fixed or variable, e.g., for some of the process tasks there might be no alternative partner services, or the IT infrastructure could be predefined. In addition, if there are alternatives for the selection of partner services or different possible infrastructure configurations, one can calculate different SLO alternatives (with different prices), the customer could later choose from.

If after calculating the overall SLOs based on available partner services and IT infrastructure the target values from the modeling phase cannot be reached, the process needs to be reengineered, through e.g. changes in the process flow and process activities, or by adapting/refining the target values in the SLOs requirements.

After calculating the SLOs of the process, an SLA template is created. Such a template consists of the definition of SLA parameters, offered SLO alternatives, requirements considering SLOs of service requester, and penalties in case of violations. An SLA parameter definition consists of process metric definitions which specify how the SLA can be monitored at runtime. In [3] SLO definitions are bound to WSDL operations. However, as argued in Section 2, it might be needed to utilize an abstract BPEL process definition in some cases.

4 Monitoring of KPIs and SLAs

After creation of the SLA template the process can be deployed to the process engine. Before being able to invoke the process, the requester and provider have to negotiate and agree on a set of SLO alternatives as specified in the SLA template. The requester can then choose one of the alternatives. The result of this process is an SLA that serves as a contract between provider and consumer. Note that a service composition can be in the role of both provider and consumer in multiple SLAs.

The SLA has to be monitored at process runtime. SLA monitoring (monitoring from service consumer point of view), which can also be done by third-parties, only observes the SLA of the two partners involved in this contract. Several approaches have already been proposed [16, 17]. In addition, service provider monitors the KPIs of his process using a BAM tool. BAM (monitoring from provider point of view) looks at the performance of the process as a whole having an overview over all partners, and enables analyzing why SLA violations took place.

When evaluating KPIs using BAM technology, the established dependencies between KPIs and SLA parameters (Section 3.2), enable analyzing "why" certain values of KPIs arose. For example, in case the "% of orders processed in full and on time" is deviating from the target value, one could analyze whether the cause lies in the supplier service, shipment service, internal process activities, or availability of process engine. In order to support this kind of analysis, the BAM tool has to aggregate events from the process engine, system monitor (providing metrics on IT infrastructure), and SLA monitor.

5 Conclusions and Future Work

In this paper we have given an overview over the steps which are needed for management of SLA-aware service compositions based on KPIs.

The first step consists of defining KPIs and SLO requirements for business processes. There are already several proposals on how to model SLOs for Web services [3, 17, 18, 19]. However, they focus mostly on technical QoS properties and do not deal explicitly with the specification of KPIs and SLOs for business processes implemented as service orchestrations, which is an interesting issue for future work.

In the configuration phase KPIs are mapped to SLOs of partner services and IT infrastructure, in order to identify the dependencies. Based on this mapping, appropriate services and IT infrastructure are selected, their SLOs aggregated and compared to the target values of the KPIs. In particular, the mapping of KPIs to SLOs is to the best of our knowledge not yet adequately addressed in related work.

Finally, during process execution KPIs and SLAs are monitored. In this context, the dependency tree created during mapping of KPIs to SLOs can be used to analyze why deviations of KPI target values took place. Therefore, the business activity monitor has to integrate events from process engine, system monitor, and SLA monitor. Developing such a BAM solution is part of our future work.

Acknowledgements

The research leading to these results has received funding from the European Community’s 7th Framework Programme under the Network of Excellence S-Cube – Grant Agreement no. 215483.

References

1. Weerawarana, S., Curbera, F., Leymann, F., Storey, T., Ferguson, D.: Web Services Platform Architecture. Prentice Hall PTR (2005)
2. Web Services Business Process Execution Language Version 2.0 – Committee Specification. Published via Internet (Jan 2007)
3. Ludwig, H., Keller, A., Dan, A., King, R.P., Franck, R.: Web Service Level Agreement (WSLA). Technical report, IBM (2003)
4. Mani, A., Nagarajan, A.: Understanding Quality of Service for Web Services. IBM (2002)
5. Cardoso, J., Sheth, A.P., Miller, J.: Workflow Quality of Service. ICEIMT (2003)
6. Supply Chain Operations Reference Model, SCOR Version 8.0 (2007) http://www.supply-chain.org
7. zur Muehlen, M.: Workflow-based Process Controlling: Foundation, Design and Application of workflow-driven Process Information Systems. Logos, Berlin (2004)

8. Wahli, U.; Avula, V.; Macleod, H.; Saeed, M.; Vinther, A.: Business Process Management: Modeling through Monitoring Using WebSphere V6.0.2 Products. IBM Redbook SG24714801 (2007)
9. Zeng, L., Benatallah, B., Ngu, A.H.H., Dumas, M., Kalagnanam, J., Chang, H.: QoS-Aware Middleware for Web Services Composition. IEEE Trans. Software Eng. 30(5) (2004)
10. Jaeger, M.C., Muehl, G.; Golze, S.: QoS-Aware Composition of Web Services: An Evaluation of Selection Algorithms. OTM Conferences (2005)
11. Jaeger, M.C., Rojec-Goldmann, G., Muehl, G.: QoS Aggregation in Web Service Compositions. EEE (2005)
12. Unger, T.: Aggregation von QoS und SLAs in BPEL Geschaeftsprozessen Diplomarbeit Nr. 2305, Universitaet Stuttgart (2005)
13. Rud, D.; Kunz, M.; Schmietendorf, A.; Dumke, R.: Performance Analysis in WS-BPEL-Based Infrastructures. In Proc. “23rd Annual UK Performance Engineering Workshop” (UKPEW 2007), pp. 130-141 (2007)
14. Marzolla, M.; Mirandola, R.: Performance Prediction of Web Service Workflows, Proc. QoSA'07, Software Architectures, Components and Applications Third International Conference on Quality of Software Architectures, QoSA 2007, Medford, MA, USA, July 11-13, (2007)
15. Menascé, D. A.; Almeida, V. A. F.: Capacity planning for web services: metrics, models, and methods. Prentice Hall (2002)
16. Sahai, A., Machiraju, V., Sayal, M., van Moorsel, A.P.A., Casati, F.: Automated SLA Monitoring for Web Services. DSOM (2002)
17. Keller, A.; Ludwig, H.: The WSLA Framework: Specifying and Monitoring Service Level Agreements for Web Services. Technical report, IBM (2002)
18. Sahai, A., Durante, A., Machirajau, V.: Towards Automated SLA Management for Web Services. Technical report, HP (2002)
19. Andrieux, A. et al.: Web Services Agreement Specification (WS-Agreement). Technical report, GGF (2004)