# Cost-Based Optimization of Service Compositions

Philipp Leitner, *Member*, *IEEE*, Waldemar Hummer, *Student Member*, *IEEE*, and Schahram Dustdar, *Senior Member*, *IEEE*

**Abstract**—For providers of composite services, preventing cases of SLA violations is crucial. Previous work has established runtime adaptation of compositions as a promising tool to achieve SLA conformance. However, to get a realistic and complete view of the decision process of service providers, the costs of adaptation need to be taken into account. In this paper, we formalize the problem of finding the optimal set of adaptations, which minimizes the total costs arising from SLA violations and the adaptations to prevent them. We present possible algorithms to solve this complex optimization problem, and detail an end-to-end system based on our earlier work on the PREvent (prediction and prevention based on event monitoring) framework, which clearly indicates the usefulness of our model. We discuss experimental results that show how the application of our approach leads to reduced costs for the service provider, and explain the circumstances in which different algorithms lead to more or less satisfactory results.

**Index Terms**—Service composition, service-level agreements, adaptation, optimization

## 1 Introduction

Service-based applications have seen tremendous research activity in the last years, with many important results being generated around the world [1]. This global interest is justified by the ever increasing services industry, which is still only starting to explore the potential that new paradigms like Everything-as-a-Service (XaaS) or cloud computing provide [2]. However, to fully realize this potential, research and industry alike need to focus more strongly on nonfunctional properties and quality issue of services (generally referred to as QoS). In the business world, QoS promises are typically defined within legally binding service-level agreements (SLAs) between clients and service providers, represented, e.g., using WSLA [3]. SLAs contain service-level objectives (SLOs), i.e., concrete numerical QoS objectives, which the service needs to fulfill. If SLOs are violated, agreed upon monetary consequences go into effect. For this reason, providers generally have a strong interest in monitoring SLAs and preventing violations, either by using post mortem analysis and optimization [4], [5], or by runtime prediction of performance problems [6], [7]. We argue that the latter is more powerful, allowing to prevent violations before they have happened by timely application of runtime adaptation actions [8], [9], [10].

However, preventing SLA violations is, in general, not for free. For instance, some alternative services usable in a composition may provide faster response times (thereby improving the end-to-end runtime of the composite service, and reducing the probability of violating runtime-related SLOs), but those services are often more expensive than slower ones. Therefore, there is an apparent tradeoff between preventing SLA violations and the inherent costs of doing so. We argue that this tradeoff is currently not covered sufficiently in research. Instead, researchers assume that the ultimate goal of service providers is to minimize SLA violations, completely ignoring the often significant costs of doing so (e.g., [9], [10]).

In this paper, we contribute to the state of the art by formalizing this tradeoff as an optimization problem, with the goal of minimizing the total costs (of violations and applied adaptations) for the service provider. We argue that this formulation better captures the real goals of service providers. Additionally, we present possible algorithms to solve this optimization problem efficiently enough to be applied at composition runtime. We evaluate these algorithms within our PREVENT (prediction and prevention based on event monitoring) framework [8].

The remainder of this paper is structured as follows: In Section 2, we motivate our work and present an illustrative example, which will guide us through the rest of the paper. Following in Section 3, we present our earlier work on prevention of SLA violations. In Section 4, we formalize the problem of cost-based optimization of service compositions. We explain possible algorithms to solve this problem efficiently in Section 5, which are experimentally evaluated in Section 6. Finally, we compare our work with the most important related scientific approaches in Section 7, and conclude the paper in Section 8.

## 2 Motivation

In this paper, we use the scenario depicted in Fig. 1 (in BPMN [11] notation) to motivate and explain our approach.

This scenario considers the case of a manufacturer of industry products. These products are constructed on-demand by assembling various parts, some of which can be

• *The authors are with the Information Systems Institute, Distributed Systems Group, Vienna University of Technology, Argentinierstrasse 8/184-1, 1040 Vienna, Austria.*
*E-mail: {leitner, hummer, dustdar}@infosys.tuwien.ac.at.*

*Manuscript received 4 Feb. 2011; revised 16 Aug. 2011; accepted 13 Sept. 2011; published online 1 Nov. 2011.*
*For information on obtaining reprints of this article, please send e-mail to: tsc@computer.org, and reference IEEECS Log Number TSC-2011-02-0010.*
*Digital Object Identifier no. 10.1109/TSC.2011.53.*

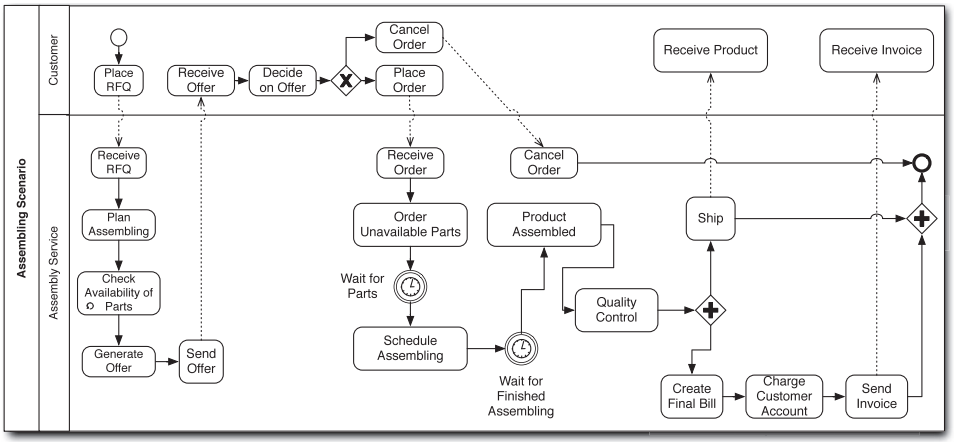



Fig. 1. Motivating scenario.

produced in-house by the manufacturer, while others need to be ordered from external suppliers. The manufacturing process depicted in Fig. 1 consists of two segments: First, the customer sends a request for quotation (RFQ), which the manufacturer responds to with an offer (consisting of estimated price and delivery time for the finished product), Second, the customer can then order this product to the offered conditions. For reasons of brevity, we concentrate on the two roles "Customer" and "Assembly Service" in the figure, even though the manufacturer interacts with many different external partners (e.g., suppliers of parts, shippers, credit card companies) to implement the described functionality. Since the manufacturer's business is based entirely on a service-based notion, the manufacturing process is implemented as a service composition, i.e., activities in the process are mapped to one or more invocations of (web) services.

With its key customers, the manufacturer has some established SLAs. We provide a list of typical SLOs in Table 1. Note that these objectives can be of quantitative (SLOs #1 to #4) or of qualitative (SLO #5) nature.

All SLOs have some target values and penalties for violating these targets associated (see Table 2). Therefore, the manufacturer has a strong interest in complying to these SLOs, as long as the costs of doing so do not exceed the benefit. The manufacturer may apply a number of runtime adaptations to the process. We sketch some example adaptation actions in Table 3. The columns + and − refer to SLOs in Table 1, and indicate that the respective action has a positive (+) or negative (−) impact on this SLO. Note that these actions and impacts are just of examplatory nature, that is, while for some business cases outsourcing may reduce costs and increase the process duration (and error rate), this does not necessarily hold for all processes. Additionally, applying these actions generally also has some associated costs, which need to be taken into account (for instance, express shipping is more expensive than regular shipping). As we can see, for the manufacturer there is a tradeoff between the three dimensions duration, costs, and quality, which is well known in many fields of engineering.

Since the manufacturer business process is implemented as a service composition, applying these adaptations

TABLE 1
Service-Level Objectives

| # | SLO Name | Description |
|---|---|---|
| 1 | **Time to Offer** | Time between receiving the RFQ and responding with an offer (in working days). |
| 2 | **Order Fulfillment Time** | Time between receiving the order and finishing the process (in working days). |
| 3 | **Process Lead Time** | Time between initializing the process and finishing it (excluding activities at customer side) in working days. |
| 4 | **Cost Compliance** | Cost overrun with regard to the offer in % of the offer. |
| 5 | **Product as Specified** | Product is exactly as specified. |

TABLE 2
Target Values and Penalties

| # | Target Value | Costs of Violation |
|---|---|---|
| 1 | $<= 2$ | Implicit costs - customer will choose a different manufacturer if offer is not received in time. |
| 2 | $<= 5$ | Manufacturer grants 5% discount per 1 day delay, 20% max discount, not additive with SLO#3. |
| 3 | $<= 6$ | Manufacturer grants 5% discount per 1 day delay, 20% max discount, not additive with SLO#2. |
| 4 | $<= 5$ | Manufacturer cannot charge more than the offer plus 5%. |
| 5 | n/a | If wrong product is delivered, manufacturer needs to produce and ship the specified product within 7 working days and grant a 5% discount. |

TABLE 3
Possible Adaptation Actions

| # | Adaptation Action | + | - |
|---|---|---|---|
| 1 | Use faster shipper or faster shipping option, e.g., express shipping. | #2, #3 | #4 |
| 2 | Order more parts instead of producing them in-house. | #2, #3 | #4 |
| 3 | Generate offer with higher priority. | #1, #3 | - |
| 4 | Outsource assembling and quality control. | #4 | #2, #3, #5 |
| 5 | Skip quality assurance or do it less thoroughly. | #2, #3, #4 | #5 |
| 6 | Add an additional quality assurance step. | #5 | #2, #3, #4 |

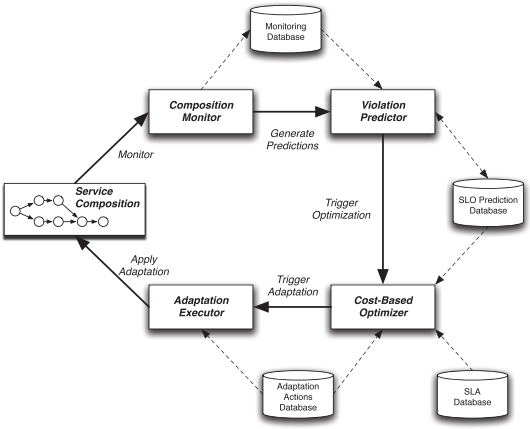



Fig. 2. Overall framework.

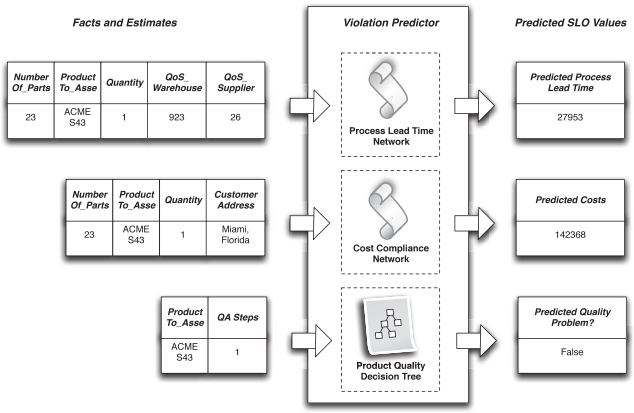



Fig. 3. Predicting SLOs using machine learning.

essentially boils down to adapting the service composition. This can be done by either adapting the data flow of the composition (e.g., to use a different shipping option), by invoking different base services, or by changing the structure of the composition itself. In our previous work, we have already shown how such adaptations can technically be applied [8], [9]. However, in these previous papers, the question of how the service provider can actually select these actions has not been discussed. Selecting the cost-optimal set of adaptations to prevent predicted violations results in an optimization problem, i.e., minimizing the total costs of all SLA violations plus all costs arising from the adaptation. This problem needs to be solved very efficiently, as the optimization has to be repeated at runtime for every composition instance that is predicted to violate one or more SLOs. Discussing this optimization problem is the main contribution of this paper.

## 3 Background

To provide some background information for this paper, we now present the PREVENT framework, which forms the basis for the research discussed here. Generally, PREVENT is a closed-loop system [12] for self-optimizing service compositions. PREVENT is based on the existing SOA runtime environment VRESCO [13]. As we have sketched in Fig. 2, the PREVENT framework consists of the seminal steps "monitor," "analyze," "plan," and "execute," as defined in the vision of autonomic computing [14]. We have previously presented our initial version of the PREVENT framework in [8].

Generally, the idea of PREVENT is to use event-based monitoring of composition data to generate runtime predictions of SLA violations before they have happened. Based on these predicted violations, adaptation actions are triggered with the goal of preventing the violation. In this paper, we focus on the implementation of the *Cost-Based Optimizer* component in Fig. 2, which we have not discussed so far in our earlier work. For every composition instance, this component receives estimations of concrete SLO values from the *Violation Predictor* component, and decides (based on these estimations as well as on knowledge of standing SLAs and available adaptation actions), which adaptations should be applied to a composition instance. In the following, we refer to this decision procedure as *cost-based optimization*. We use the term *optimization time* as the point in time during a composition instance's execution at which cost-based optimization happens. The interested reader may download our current version of the PREVENT prototype.[1]

### 3.1 Prediction of SLOs

Generally, the PREVENT approach to prediction of SLA violations is based on the idea of predicting concrete SLO values based on monitoring data. We distinguish three different types of information. *Facts* represent data that can already be measured at optimization time. *Unknowns* are the opposites of facts. They represent data that are entirely unknown at optimization time. Evidently, unknown data cannot be used in the prediction. *Estimates* are a kind of middle ground between facts and unknowns, in that they represent data that are not yet available, but can in some way be estimated. This is often the case for QoS data, because techniques such as QoS monitoring [15] can be used to get an idea of, e.g., the response time of a service before it is actually invoked. The *Violation Predictor* uses both facts and estimates from previously monitored historical service executions to train a machine learning function (we use multilayer artificial neural networks [16] for quantitative SLOs and C4.5 decision trees [17] for qualitative SLOs), which can then be used to produce a numerical estimation of the SLO values at runtime. More details about our approach to prediction of SLOs can be found in our earlier work [6].

We have sketched this machine learning-based implementation of the *SLO Predictor* in Fig. 3. One model is trained per SLO that needs to be predicted (even though the same model can be used if this SLO is used in multiple customer SLAs), and every model is trained from different data. Please see below for a discussion of how to identify which data to use for each SLO. Apparently, some historical executions of the service composition are necessary to bootstrap the training. The concrete amount of instances that are necessary depend both on the expected quality of prediction (more historical information in tendency improves the prediction quality) and on the size and complexity of the service composition. This prediction approach cannot be used while no historical information is available.

1. http://sourceforge.net/projects/vresco/.

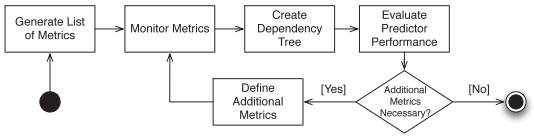



Fig. 4. Schematic dependency analysis process.

If this is the case, one could still use an alternative prediction approach, which is not data based, e.g., [18]. However, a detailed discussion of this remains part of our future work.

For understanding the remainder of the paper, it is important to keep in mind that the *SLO Predictor* essentially implements a set of estimator functions, which can be used for any partially known instance of the composition (i.e., an instance, whose facts and estimates are partially known, for instance a half-finished instance) to generate an estimation of the SLO value when the instance is finished. We will use these estimator functions in our modeling in Section 4.

### 3.2 Identification of Factors of Influence

As input to the machine learning-based *SLO Predictor* approach, we need to identify the most significant metrics that influence the SLO compliance of the composition. We refer to these metrics as the factors of influence of the service composition. Factors of influence are rarely obvious, even to domain experts. Hence, we have devised a process called dependency analysis, which can be used by business analysts to identify factors of influence. We summarize this process here, to the extent that is necessary for understanding the core contribution of the current paper.

Dependency analysis is a semiautomated process. We rely on the domain knowledge of a human business analyst, but support her with automation and knowledge discovery tools to ease repetitive tasks. The high-level process is sketched in Fig. 4. As a first step, the business analyst needs to define an (initial) list of potential factors of influence. These include both domain-specific metrics, which need to be defined manually, and typical QoS metrics, which can be automatically generated (e.g., for every used service, we generate response time and availability metrics). For every potential factor of influence, a monitor is defined or generated, which specifies how this metric can be measured from a running instance. Second, a data set containing these factors needs to be generated, either by simulating the composition in a web service test environment (e.g., Genesis2 [19]) or by monitoring real executions with monitoring of all potential factors of influence enabled. Using this data set, a dependency tree can be generated, as discussed in [5]. The dependency tree is essentially a decision tree, containing the factors that best explain SLO violations in the composition. The third step is then to use these factors to try and train a prediction model from the identified factors of influence. If this prediction model has a sufficiently high training data correlation against the measured data set (i.e., if the predictions generated with the predictor are highly correlated with the actual measured values), we can accept these factors and influence and use them in the *SLO Predictor* for the SLO. If the correlation is not sufficient, the business analyst needs to identify the reason for the lacking performance. Generally, the analyst will define additional potential factors of influence, and repeat from the second step.

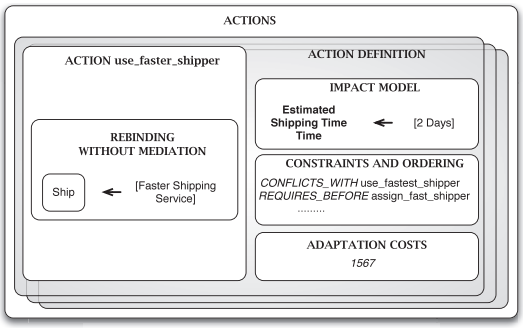



Fig. 5. Definition of adaptation actions.

### 3.3 Adaptation Actions

The PREVENT *Adaptation Executor* can execute a range of different adaptations of service composition instances. Generally, we distinguish three types of adaptations: Data manipulation, service rebinding, and structural adaptation. Data manipulation actions represent the most simple type of adaptation, where the composition is in fact not changed. Instead, the data flow of the composition instance is intercepted and some datum is changed (e.g., the *priority* parameter of the service invoked as part of the "ship" activity is changed to "high priority"). Service rebinding represents the common case, where a different service is used to implement an activity in the composition, e.g., a faster shipping service is used in the activity "ship." For this type of adaptation, we differentiate between three types, one-to-one service rebinding without interface mediation (the original and the new service have identical interfaces), one-to-one service rebinding with interface mediation (the services have different interfaces, but the same number of service invocations is needed to achieve the required functionality), and substitution with subflow (the original service invocation is not only replaced with another single service invocation, but with a whole subcomposition). This adaptation is similar to the another type of adaptation, structural adaptation. In this case, not only the data or service bindings of a composition change, but the logical structure of the composition itself. This includes simpler cases like removing activities in an instance (e.g., skip the "quality control" activity) and more complex adaptations, where an entire subtree of the composition definition is replaced (e.g., outsource the assembling process to an external provider). Please refer to our earlier publications [8], [9] for details on how these actions are implemented. Most important for the remainder of this paper is to know how adaptation actions are defined in the PREVENT framework. We have sketched this in Fig. 5.

We can define any number of adaptation actions, which can be applied to an instance of the composition. Each of those definitions contains the description of the actual action, which can be any of the action types discussed above. In addition, the action definition also contains the impact model of the action, a list of constraints and ordering clauses, and the costs of applying this action. We assume

that every adaptation action has a constant, nonnegative cost. For example, the cost of using a faster shipping service is the cost of using the new service minus the costs of using the original shipping service. The impact model contains a set of impact clauses. Every impact clause represents the concrete impact that applying this adaptation action has on one concrete monitorable fact or estimate. Essentially, therefore, the clauses model updates to the data used to generate predictions (see Fig. 3). Every adaptation action can have any number of positive as well as negative impacts on any fact or estimate. This impact value can be determined in several ways: 1) based on measured history data if the corresponding advice has already been used before, for example, using data mining; 2) based on SLAs with external providers, if such SLAs exist; or 3) by using QoS aggregation techniques [20]. We assume that the impact model specifies impact clauses for all metrics, which the advice affects. Of course, impact clauses do not need to be exact (very often it will realistically be impossible to statically define an exact impact model before execution); however, more exact impact models lead to better predictions of SLOs after adaptation, which in turn leads to a better end-to-end performance of the PREVENT system.

## 4 Optimization Problem Formulation

In this section, we formalize the problem of selecting the most cost-effective adaptation actions to prevent one or more predicted SLA violations. We consider an interaction of the service composition with a given client, who has a given SLA with the composition provider. Let $I$ be the set of all possible composition instances of this client, and let $i \in I$ be concrete instances that we can monitor using the PREVENT tooling. Furthermore, let $S = \{s_1, s_2, \dots s_k\}$ be the set of SLOs defined in the relevant SLA. As part of the SLO definition, a penalty function is associated with all SLOs in $S$. Collectively, we refer to these functions as $P = \{p_{s1}, p_{s2}, \dots p_{sk}\}$. Penalty functions define the costs for the provider based on a measured SLO value, i.e., they are functions defined as $p_s : \mathbb{R} \rightarrow \mathbb{R}, s \in S$. Similarly, the measured value of an SLO $m_s$ is a function $m_s : I \rightarrow [0:1]$. We normalize SLO values to the interval $[0:1]$ to make them comparable. Putting it all together, we define the penalty function for a given SLO $s$ and instance $i$ as

$$p_s^i \stackrel{\text{def}}{=} p_s(m_s(i)).$$

Penalty functions for SLOs can take many different shapes. The most important ones are:

1. *constant* penalty (a constant payment needs to be made if a certain SLO threshold value is surpassed),
2. *staged* penalty (similar to a constant penalty, but with different levels of penalty),
3. *linear* penalty (the penalty is linearly increasing with the degree of violation), and
4. *linear* penalty with cap (the penalty is linearly increasing up to a maximum value).

Even though these functions span many different types of mathematical functions, they share two essential characteristics. First, SLA penalty functions are always monotonically increasing, i.e.,

$$\forall p_s \in P : \forall x_1, x_2 \in \mathbb{R} : (x_1 < x_2) \Longrightarrow (p_s(x_1) \leq p_s(x_2)).$$

This is evident, because the penalty for a higher degree of violation should never be smaller than the penalty for a lesser violation. Second, SLO penalty functions always have a point discontinuity in a special violation threshold point ($t_1$). Before (and including) $t_1$ the penalty is generally 0 (no violation has occurred), and beyond this point a positive penalty needs to be paid

$$(\forall s \in S : \forall x \in \mathbb{R} : (x \leq t_1 \iff p_s(x) = 0) \wedge (x > t_1 \iff p_s(x) > 0)).$$

This also means that penalty functions are generally discontinuous. Furthermore, this property signifies that there is no incentive for the service provider to apply further adaptation and improve an SLO value below $t_1$, because all further improvements do not further reduce his costs (they are already 0 for this SLO).

To prevent violations, we are able to apply a number of possible adaptations to an instance $i$. We define $A = \{a_1, a_2, \dots a_l\}$ as the set of all possible adaptation actions, and $A^* \in \mathcal{P}(A)$ ($\mathcal{P}(A)$ denotes the powerset of $A$) as the subset of adaptation actions that are selected to be applied. We assume that all adaptations have some costs associated, defined as a cost function $c : A \rightarrow \mathbb{R}$. We assume that cost functions are constant, that is, we do not consider cross-pricing models for services [21], which would lead to nonconstant costs of adaptation. Furthermore, adaptation actions, if applied, have some defined impact on the composition instance $i$. Hence, we define the transformation of $i$ to a modified instance $i'$ using the $\circ$ operator, defined as a function $\circ : I \times \mathcal{P}(A) \rightarrow I$. This is captured by the impact model, which has to be specified as part of the action definition (see Section 3).

Selecting the most cost-effective adaptation actions means finding the adaptation actions ($A^*$) that minimize the total costs for the service provider. The total costs $TC$ are defined in Equation 1 as the sum of the costs of SLA violations after adaptation ($VC$) and the costs of adaptation ($AC$).

$$TC : \mathcal{P}(A) \rightarrow \mathbb{R}, TC(A^*) = VC(i \circ A^*) + AC(A^*). \quad (1)$$

$AC$ is the sum of the costs of all applied adaptation actions (Equation 2).

$$AC : \mathcal{P}(A) \rightarrow \mathbb{R}, AC(A^*) = \sum_{a_x \in A^*} c(a_x). \quad (2)$$

$VC$ is defined as the sum of all penalty functions applied to an instance (Equation 3).

$$VC : I \rightarrow \mathbb{R}, VC(i) = \sum_{s_x \in S} p_{sx}^i. \quad (3)$$

Obviously, the goal of the service provider is to minimize $TC$. Hence, the optimization objective becomes finding the $A^*$ that minimizes $TC$ for a given instance $i$ (Equation 4).

$$TC(A^*) = \sum_{s_x \in S} p_{sx}^i + \sum_{a_x \in A^*} c(a_x) \rightarrow min! \quad (4)$$

Note that we can easily calculate $AC$ for any given $A^*$, but at optimization time, $VC$ is unknown (we do not know for sure which SLOs will be violated, with or without

```
1  # name: bab
2  # input: partial solution p,
3  #        next alive action index i,
4  #        target function v
5  # output: optimal complete solution
6
7  bab(p,i):
8    # recursion break condition
9    if(p is complete solution)
10     return p
11
12   # check if this sub-tree can be pruned
13   if(p is pruneable)
14         forall alive_actions(p) as j
15           set p(j) = 0
16     return p
17
18   # investigate solution sub-tree with p(i)=0
19   set p(i) = 0
20   s1 = bab(p,i+1)
21
22   # investigate solution sub-tree with p(i)=1
23   set p(i) = 1
24   s2 = bab(p,i+1)
25
26   # return better solution from both subtrees
27   if(v(s1) <= v(s2))
28     return s1
29   else
30     return s2
```

Fig. 6. Branch-and-Bound algorithm.

adaptation). However, the SLO Predictor provides estimations for SLOs based on instance data (see Section 3). Hence, we assume that we have estimation functions $e_s : I \rightarrow \mathbb{R}, s \in S$ available for each SLO, which estimate the concrete penalty values in advance with a reasonably small prediction error $\epsilon$ ($\forall s \in S, i \in I : |e_s(i) - p_s(i)| < \epsilon$). Replacing $VC$ with its prediction using $e_s$ leads to Equation 5, which we can solve.

$$TC(A^*) \approx \sum_{s_x \in S} e^i_{sx} + \sum_{a_x \in A^*} c(a_x) \rightarrow min! \quad (5)$$

However, not all combinations of adaptation actions are legal. Some adaptation actions are mutually exclusive (e.g., `use Shipping Service DHL` and `use Shipping Service UPS`), while others depend on each other (see our earlier work [9] for details on dependencies between adaptation actions). For simplicity, we capture these additional constraints using a penalty term $v : \mathcal{P}(A) \rightarrow \mathbb{N}$. The definition of $v$ is shown in Equation 6.

$$v(A^*) = \begin{cases} \infty & A^* \text{contains constraint violation} \\ 0 & \text{otherwise.} \end{cases} \quad (6)$$

By incorporating this penalty term, we arrive at our final target function (Equation 7).

$$TC(A^*) \approx v(A^*) + \sum_{s_x \in S} e^i_{sx} + \sum_{a_x \in A^*} c(a_x) \rightarrow min! \quad (7)$$

We have all necessary information to evaluate Equation 7 at optimization time for any set of actions $A^*$. However, finding the $A^*$ that minimizes $TC(A^*)$ is still far from trivial, because Equation 7 is discrete and cannot be optimized analytically. We present algorithms to find a (near-)optimal solution in Section 5.

## 5 Algorithms

We will now discuss different approaches for finding solutions to this problem. These algorithms are implemented in the *Cost-Based Optimizer* component. Optimization is always triggered by a predicted violation of at least one SLO, and receives as input a list of monitored facts and estimates of the current instance.

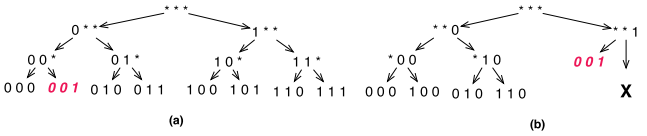



Fig. 7. Pruning of solution trees.

### 5.1 Branch-and-Bound

Branch-and-bound is a very general deterministic algorithm for solving optimization problems. The high-level idea of this approach is to enumerate the solution space in a "smart" way so that at least some suboptimal solutions can be identified and discarded prematurely, i.e., before they have been fully constructed and evaluated. We use a binary encoding to represent solutions, i.e., every solution is represented as a binary vector, and an adaptation action with index $j$ is applied *iff* the solution vector is 1 at index $j$. For example, the solution vector 00110100 encodes that the third, fourth, and sixth adaptation action should be applied. Evidently, $2^{|A|}$ different solutions exist for each optimization problem, where $|A|$ is the number of possible adaptation actions (but not all combinations need to be legal). For solutions that are still being constructed we allow a third symbol, "$*$", representing an action that is still undecided (alive). We refer to solutions, which contain at least one alive action, as partial, and solutions, which do not contain any alive actions, as complete. Therefore, the vector $001101 * 0$ is a partial solution, where the last-but-one action is alive.

We describe our general Branch-and-Bound algorithm in Fig. 6. The algorithm is easy to understand. What is the most important is the implementation of Line 13, the rules for pruning the search tree (i.e., for prematurely discarding solutions). In our Branch-and-Bound approach, we prune a partial solution in two cases: 1) if the partial solution already contains at least one conflict, or 2) if the partial solution already prevents all SLA violations (the penalty function $p_s$ is 0 for all $s \in S$) without applying any more actions. Case 1 is trivial, because the target function value for all solutions in such a subtree will always be $\infty$. Case 2 lends itself to more discussion. Remember the assumption that every action has nonnegative costs, and that we described SLA penalty functions as nonnegative functions. Therefore, we can assure that for any solution where all penalty functions are 0, the additional application of more actions can never improve the target function value. Hence, these partial solutions cannot be improved by applying more actions, and the remaining solution subtree can be pruned.

In Listing 6, we simply iterated over all actions in the order they appeared in the solution vector (in every step, we always just investigate the next action, see Lines 18 and 22). In general, this approach is suboptimal. Even though the order in which we investigate actions has no impact on the quality of our solution (the algorithm is deterministic, i.e., we will always find the global optimum eventually), the order may have an impact on the number of solutions we are able to prune. This is illustrated in Fig. 7. Assume the

following simple scenario: There is only one SLO and three possible adaptations. Only adaptation 3 is able to prevent the violation of the SLO. Actions 1 and 2 have costs but no relevant influence. There are no conflicts between actions. Hence, the optimal solution vector is 001. In Fig. 7a, we strictly followed the algorithm in Listing 6 and investigated the actions in the order they appear in the solution vector. Since the only "useful" action is investigated last, we extend the whole solution tree without any pruning (the worst case, equivalent to full enumeration). Now, in Fig. 7b, we investigate the actions in reverse order (from back to front). Now, the "useful" action is investigated first, and a large part of this solution tree can be pruned according to pruning Case 2.

Therefore, we can conclude that it is beneficial to investigate actions in a specific order that maximizes the number of solutions that can be pruned. We specify two possible criteria for this ordering: 1) the *impact* of an action on the SLOs (actions with higher total impact should be investigated first), and 2) the *utility* of an action (actions with higher utility should be investigated first). We will now define those two orderings.

Based on the set of historical process instances that we have already used to train the *Violation Predictors*, we can calculate an estimation of impact and utility of each action as follows: We define the set of available historical process instances as $H = \{h_1, h_2, \ldots h_q\}$, with $H \subseteq I$. We refer to the number of historical instances as $q = |H|$. Now, we are able to calculate an estimation of the overall impact of an adaptation action $a$ on a SLO $s$ as $\Delta_{a,s}$ (Equation 8). Simply put, the impact is the arithmetic mean of the difference between SLO value with and without applying the adaptation to each historical instance.

$$\Delta_{a,s} = \sum_{h \in H} \frac{m_s(h) - m_s(h \circ \{a\})}{q}. \tag{8}$$

Note that we have already defined in Section 4 that SLA penalty functions are monotonically increasing. Hence, higher impact values are generally good. However, the impact value may also be negative (i.e., $m_s(h) < m_s(h \circ \{a\})$). In this case, this action has a negative impact on one of the SLOs, which is reasonable and realistic. For instance, an adaptation that reduces the process lead time can very well have a negative impact on the SLO cost compliance. Based on $\Delta_{a,s}$, we can now define the total impact of each action as the sum of its impact on all SLOs. Furthermore, we can define the utility of an action as its total impact divided by its costs. Now, we are able to improve the Branch-and-Bound algorithm trivially: Instead of investigating the actions in the order they are specified in the solution vector, we now investigate them either in the order of their impact $\Delta_a$ (impact-based sorting), or in order of their utility $u_a$ (utility-based sorting). We will evaluate and discuss both alternatives in Section 6, and compare them to Branch-and-Bound with randomly ordered actions.

### 5.2 Local Search

While the Branch-and-Bound algorithm discussed above has the advantage of always finding the optimal set of actions for any composition instance, the execution time of the algorithm increases exponentially with the number of available actions. Even though we can reduce the runtime using impact- or utility-based sorting of actions, the complexity still remains exponential. Hence, there is an evident need to find strong heuristics, i.e., nondeterministic algorithms that find "good" (even if not necessarily optimal) solutions in polynomial time.

```
# name: grasp_init
# input: number of start solutions n,
         RCS max size r
# output: set of start solutions

grasp_init(n, r):
  G = {} // empty set of start solutions
  repeat n times:
    pa = empty_partial_solution
    while( VC(pa) > 0 ):
      rcs = construct_rcs(pa, r)
      if( empty(rcs) )
        break
      a = random(rcs)
      pa(a) = 1
    add(G, pa)
  return G
```

Fig. 8. GRASP construction heuristic.

A simple heuristic that is often used to very good ends is Local Search. Local Search is a metaheuristic, i.e., final solutions are constructed by iteratively improving a start solution. The general idea is that in each iteration the algorithm searches a specified *neighborhood* for better solutions than the current one. If at least one such solution is found, the algorithm progresses to the next iteration with one of the better solutions (typically, the best one in the neighborhood, equivalent to steepest descent). If no better solution can be found in the neighborhood, the algorithm has converged to a local optimum and is terminated. Usually, this algorithm is repeated multiple times with different starting solutions (because different starting solutions can lead to different local optima). This kind of algorithm typically depends on the definition of: 1) a suitable neighborhood and 2) a senseful selection of starting solutions. We use the following neighborhood definition: A complete solution vector is in the neighborhood of an original solution if the two solutions represented as binary vectors have a Hamming distance of 1, i.e., if they differ in exactly 1 bit.

For selecting the start solutions, we use two different approaches. The first and primitive one is to select $n$ start solutions with $m$ bits set to 1 at random. Alternatively, we propose to use an algorithm commonly referred to as GRASP [22] (greedy randomized adaptive search procedure). GRASP is essentially a variation of local search, in which the start solutions are constructed using a greedy heuristic. The idea is that GRASP can converge to a better solution than a simple local search because the start solutions are already better than random start solutions. However, some attention needs to be paid to using a greedy construction heuristic that actually generates start solutions, which are both of reasonable quality and at the same time widely spread over the search space.

We have sketched the construction heuristic that we have used in our implementation of GRASP in Fig. 8. Summarizing, the algorithm constructs $n$ solutions by stepwise addition of actions selected randomly from a

TABLE 4
GA Configuration Parameters

| Name | Description | Default |
|---|---|---|
| Population Size | Number of solutions in every generation | 150 |
| Selection Pressure | Number of solutions to select for tournament selection | 2 |
| Crossover Probability | Probability per solution that crossover is applied | 0.8 |
| Mutation Probability | Probability per bit that mutation is applied | 0.02 |
| Break Condition | Condition for stopping the algorithm | No impr. in 20 iterations |

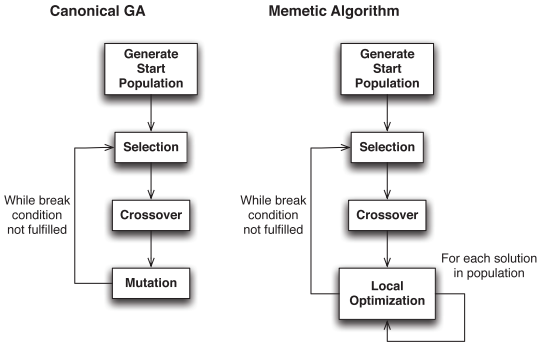



Fig. 9. Memetic algorithm.

restricted candidate set (RCS). The heuristic is based on similar concepts that we have already used in our discussion of Branch-and-Bound: The idea is to stop adding actions if either no more SLOs are violated or no senseful actions are available anymore (the RCS is empty), and to prefer adding actions which have a high utility value ($u_a$). Hence, in every step, the RCS consists of the $r$ (maximum size of the RCS) actions with highest nonnegative $u_a$, which have not yet been added and which do not lead to a conflict.

### 5.3 Genetic Algorithm

As an alternative to locality-based heuristics (local search, GRASP) we also present a solution based on the concept of evolutionary computation. More precisely, we use genetic algorithms (GAs) [23] as a more complex, but potentially also more powerful heuristic to generate good solutions to the cost-based optimization problem. The overall idea of GA is to mimic the processes of evolution in biology, specifically natural selection of the fittest individuals, crossover, and mutation. Therefore, in GA, we prefer to work on a population of solutions instead of a single one. We use the term "fit" to describe solutions with a good (low) target function value. First, we generate a random start population. For this, we use the same primitive construction scheme as discussed above for local search: We randomly apply $m$ actions in every solution. Every following iteration of the algorithm (referred to as generations) essentially follows a three-step pattern.

First, we *select* a set of solutions from the population to "survive" into the next generation. In our genetic algorithm implementation, the fittest solution (i.e., the one with the lowest target function value) is selected deterministically (elitism), while all remaining slots in the next generation population are selected using a process called tournament selection. In tournament selection, $t$ random solutions from the last generation are put into a tournament. The fittest solution of the tournament is selected into the next generation. The parameter $t$ steers the selection pressure: Low $t$ increases the time that the population takes to converge against a solution, but high $t$ increases the danger of converging against a local optimum instead of the global one.

Second, *crossover* is used to produce new solutions based on the selected ones from the last generation. The main challenge of implementing a strong crossover mechanism is to ensure that the crossover product of two fit solutions is also likely to be fit. Given the binary vector representation we use to encode solutions, we can make use of a simple one-point crossover scheme. We choose a random crossover point $cp$ from $[1 : |A| - 1]$. To construct a new child, we copy the binary vector of the first solution from the start until $cp$, and the vector of the second solution from $cp + 1$ to the end of the vector.

This simple procedure ensures that characteristics of both original solutions are preserved. However, because of the random selection of $cp$, it is possible that the child solution has a conflict, even if this was not the case for any of the parents. In this case, we remove one of the conflicting actions at random.

Third, we use *mutation* to introduce entirely new features into the population. The need for mutation can be illustrated easily: Assume that a given action $a$ is not applied in any solution in the population. Using one-point crossover as discussed above it is not possible to create any solution that uses $a$. Hence, we introduce some additional randomization. After crossover, we may randomly flip every bit in every solution in the population with a very small probability. This means that most solutions in the population are not mutated, but sometimes new actions are applied, which are not the product of crossover.

GAs are notorious for having many parameters to fine-tune the performance of the optimization. For illustrative purposes, we have summarized the parametrization options available in our implementation of GA in Table 4, including some values that we found to provide useful default parameters if applied to the cost-based optimization problem. Evidently, further customization would also be possible, for instance by using a different selection or crossover scheme. Unless stated otherwise, we will use the configuration described in Table 4 for experimentation in Section 6.

Unfortunately, this "canonical" GA implementation takes a significant amount of time to converge against a solution, because the solution space is searched solely through the (rather unguided and strongly randomized) means of crossover and mutation. One possibility to improve this aspect is to combine the canonical GA with local optimization as presented above. This leads us to an adapted algorithm, which we have sketched in Fig. 9. In literature, such combinations of GA and local search are often referred to as memetic algorithms (MA) [24].

The main changes of MA (as compared to GA) are as follows. First, a new *Local Optimization* operator is introduced after crossover. Local optimization applies

TABLE 5
Case Study SLOs

| # | SLO Name | $\mu$ | $\mu*$ | $\sigma$ | $\sigma*$ | $t_1$ |
|---|---|---|---|---|---|---|
| 1 | Order Fulfillment Time | 38811 | 35560 | 4708 | 6004 | 37000 |
| 2 | Payment Time | 4187 | 2202 | 28 | 1124 | 4150 |
| 3 | Shipping Time | 1285 | 864 | 144 | 347 | 1300 |
| 4 | Product Quality | 2.6 | 3 | 1.9 | 2.5 | 3 |
| 5 | Cost Compliance | 851 | 1149 | 212 | 521 | 1400 |

the local search algorithm as discussed above to each solution in the generation, basically reducing the population to a set of locally optimal solutions. Second, we remove the mutation operator from the algorithm (technically speaking, we set the mutation probability parameter to 0). The main reason is that given that all solutions in the population are already locally optimal, randomly mutating one bit in a solution can only lead to a worse solution. In theory, it is possible that multiple bits in a single solution are mutated at the same time, and that these mutations lead to an improvement, but this corner case is very unlikely in practice. Furthermore, the main motivation for having mutation in the first place was that it is the only way of introducing new actions in the canonical GA. This is no longer the case, because local search can do the same thing.

Generally, MA is slower than GA, because more solutions are evaluated in each generation (evidently, MA executes one local search for every solution in each generation). However, the algorithm potentially converges against a very good solution in a low number of generations. Hence, we argue that in practice MA improves on the canonical form most of the time for our problem. This will be substantiated further in Section 6.

## 6 Experimentation

In the following section, we will numerically validate the algorithms discussed in Section 5 based on an implementation of the scenario presented in Section 2. For reasons of brevity, we only summarize the experiment setup here. More details can be found in the accompanying experimentation webpage.[2] In addition, we do not explicitly evaluate the prediction quality (i.e., the SLO Predictor component) here. The interested reader may find a numerical evaluation of the prediction in [6], as well as in [8].

We have implemented the scenario from Section 2 using .NET Windows Communication Foundation[3] (WCF) technology and the VRESCO SOA runtime environment on a server running Windows Server Enterprise 2007, Service Pack 2. The machine is equipped with two 2.99-GHz Xeon X5450 processors and 32 GB of RAM. To train PREVENT, we have initialized the system with a set of 9,796 historical composition instances. These instances were created by executing the service composition repeatedly. In this historical data set, 3,660 instances have not been adapted, while one or more adaptation actions have been applied in the remaining 6,136 instances. In our experiments, we consider the case of an SLA containing up to five SLOs,

2. http://www.infosys.tuwien.ac.at/prototype/VRESCo/experimentation.html.
3. http://msdn.microsoft.com/en-us/library/ms735967(VS.90).aspx.

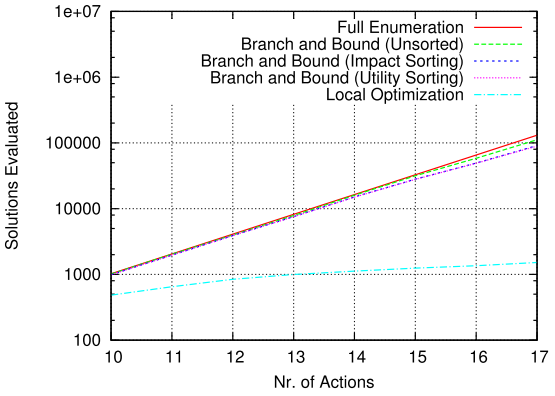



Fig. 10. Solutions evaluated for Branch-and-Bound.

similar to the previous example. Note that we have used an integer value in $[0:15]$ to represent product quality in this example, to allow for more fine-grained distinctions of different levels of product faults. In Table 5, we have sketched these SLOs and their basic statistics. $\mu$ is the mean value of the SLO without adaptation. $\mu*$ is the mean among instances to which some adaptation has been applied. $\sigma$ and $\sigma*$ are the respective standard deviations. As before, $t_1$ is the violation threshold. Furthermore, SLO 1 is associated with a staged penalty function with nine stages, SLO 2 and SLO 3 are both associated with fixed penalty functions, SLO 4 is associated with a linear penalty function with cap, and SLO 5 with a linear penalty function without cap. Additionally, we have defined 49 adaptation actions that have positive and negative influences on some or all of these SLOs. Every action has been associated with a positive cost value.

As a first experiment, we analyze the suitability of different variants of the Branch-and-Bound algorithm. As all of these algorithms are deterministic, we are guaranteed to find the optimal solution to any optimization problem eventually. However, the three different versions of the algorithm (Branch-and-Bound with random action sorting, with impact-based sorting, and with utility-based sorting) may differ significantly with regard to their runtime. As an independent measure of algorithm runtime, we use the number of solutions that have to be evaluated. All results concerning algorithms with randomized elements are arithmetic means of five repeated runs.

Fig. 10 plots the number of solutions depending on the number of adaptation actions that are available (up to a maximum of 17 actions, note the logarithmic scale on the $y$-axis). For reasons of comparison, we also plot local search in the figure, whose runtime grows linearly with the number of actions. It was not feasible to evaluate Branch-and-Bound for more than 17 actions.

As we can see, there is little difference between the three variants of Branch-and-Bound, and none is able to reduce the number of solutions that have to be evaluated significantly below full enumeration. The reason for this unsatisfying result is that, in this concrete optimization instance, very little combinations of actions can prevent the violation of all SLOs (the SLOs are conflicting), i.e., bounding Condition 2 cannot be applied very often. We can see that by relaxing the problem and disabling SLOs 4 and

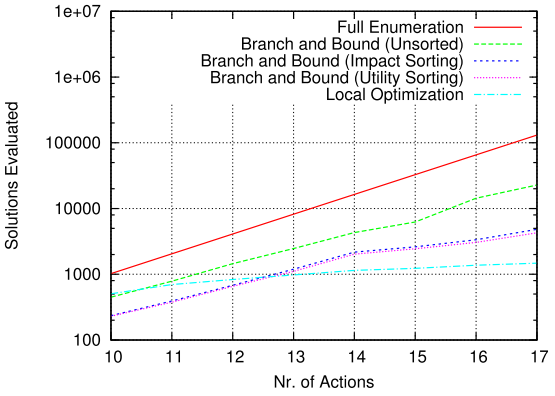



Fig. 11. Solutions evaluated without conflicting SLOs.

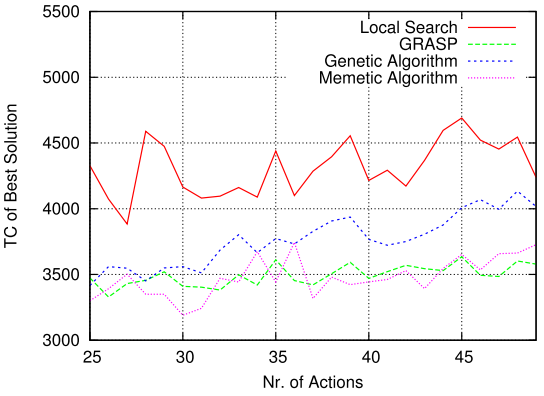



Fig. 13. Quality of solution per heuristic algorithm.

5, a significant performance boost can be achieved (Fig. 11) by both impact-based and utility-based sorting. The difference between impact-based and utility-based sorting is not significant.

However, even though smart action sorting can reduce the solution space if there are no conflicting SLOs, the number of solutions that need to be evaluated still grows exponentially with the number of available actions. Hence, solving the cost-based optimization problem deterministically is only possible for very small problems. If the set of possible adaptations grows, we need to fall back to heuristic optimization. For these algorithms, there are no guarantees about the quality of the solution. That means that we need to compare them in two dimensions. First, and similar to before, we need to look at the number of solutions that are evaluated before the algorithm produces the final result (Fig. 12), as a measure of the runtime of the algorithm. Second, we also need to take into account the quality of the best found solution (Fig. 13).

In Fig. 12, we can see that, not surprisingly, all algorithms scale much better than Branch-and-Bound (note the linear scale on the $y$-axis and compare with Fig. 11). GRASP is very efficient, and the fastest algorithm in this experiment with an almost constant runtime. The computation of local search is also reasonably efficient, but the number of solutions that have to be evaluated increases more strongly as compared to GRASP. This is because for GRASP the start solutions are already better, hence less local search steps are necessary before a solution is reached.

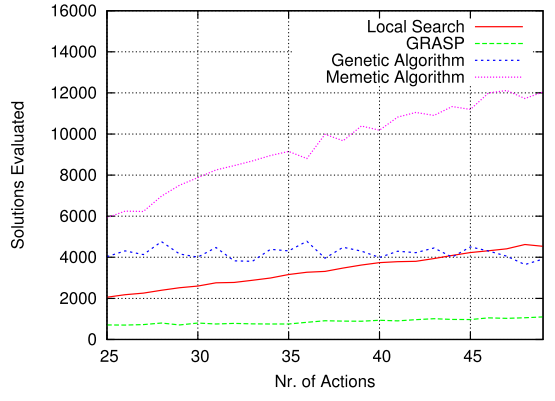



Fig. 12. Solutions evaluated per heuristic algorithm.

Note that the number of solutions evaluated for local search is directly proportional to the number of start solutions used. In this experiment, we used 25 start solutions. If we had used 50 start solutions instead, the runtime of local search would have been almost on the level of MA. GA also has a relatively constant runtime, but on much higher level than GRASP. The slowest algorithm in this experiment is MA, which is due to its unique combination of local optimization and genetic algorithm.

In Fig. 13, the algorithms are compared with regard to solution quality, measured as predicted total costs ($TC$, as defined in Section 4) for the service provider. We observe a quite clear ordering of algorithms in this experiment. GRASP and MA generally perform best. For most instances, MA is slightly better, even though this is not true for all cases. GA comes in third, and local optimization with random start solutions usually produces solutions vastly inferior to all competitors.

Drawing conclusions from these experiments, we note that Branch-and-Bound is applicable in situations, where just a small set of actions is available. In general, impact-based or utility-based sorting should be used instead of random sorting, because there is no evident disadvantage to these approaches and they may be helpful if there are no conflicting SLOs. We did not discover a significant difference in the performance of these two variants. If more actions are available, MA and specifically GRASP are interesting candidate algorithms. GRASP produces good solutions in very little time and can generally be used even for short-running compositions, where adaptation decisions need to be taken in a short time frame (below 1 second). MA is very promising in case of long-running compositions, where the time necessary to find a solution is not critical. MA often produces slightly better solutions than GRASP, but takes much more time to do so.

In a second set of experiments, we now evaluate the end-to-end effectiveness of PREVENT. That is, we analyze if the system fullfills its main promise, preventing SLA violations, and reducing the total costs for the service provider. Hence, we execute 500 instances of the scenario composition and monitor the actual total costs and violations (after adaptation). We compare these numbers with the number of violations and the total costs that the PREVENT SLO Predictor can predict after roughly half of the service composition is finished. We assume that these predictions

TABLE 6
End-to-End Results

| | SLO 1 | SLO 2 | SLO 3 | SLO 4 | SLO 5 | Total |
|---|---|---|---|---|---|---|
| **Considering Costs** | | | | | | |
| Violations Predicted/Actual | 209/129 | 442/1 | 390/42 | 75/104 | 0/41 | 1116/317 (28.4%) |
| Avg. Costs Predicted/Actual | 5207/2904 | 884/2 | 6264/558 | 840/1068 | 0/9 | 14923/8415 (56.4%) |
| **Ignoring Costs** | | | | | | |
| Violations Predicted/Actual | 223/40 | 449/0 | 245/17 | 66/218 | 0/115 | 983/390 (39.7%) |
| Avg. Costs Predicted/Actual | 5521/756 | 898/0 | 4086/216 | 756/2364 | 0/29 | 12632/22241 (176.1%) |

reflect the violations and costs that we would end up with if we did nothing at all. Since our case study is rather short running, but uses a relatively large set of adaptations, we use GRASP for cost-based optimization. The results of this experiments are depicted in Table 6.

Evidently, the usage of PREVENT fulfills its main promise. Using PREVENT the total number of SLO violations decreases to about 28 percent of the number of predicted violations. However, we can also see that PREVENT does not primarily prevent violations, but rather aims at minimizing the costs of violations. For instance, for SLO 4 and 5 the total number of violations even increases. This is because these SLOs are conflicting with the first SLOs, and SLO 1 is in general the most expensive one to violate. Hence, PREVENT happily trades violations of SLO 4 and 5 for preventing violations of SLO 1. Thereby, the total costs for the service provider can be reduced to 56 percent of the predicted costs. The lower part of the table validates the claim of the paper that it makes sense to incorporate the costs of adaptation into the decision process. To that end, we have modified the target function of the optimization in such a way that the costs of adaption are ignored. In this configuration, the total costs after adaptation are 176 percent of the predicted costs. That means that in this experiment it is in fact much more expensive for the provider to prevent adaptations (in the way that optimization ignoring costs suggests) than doing nothing at all.

## 7 Related Work

To the best of our knowledge, no approaches with the exact focus of this paper (cost-based optimization of service compositions) have been published so far. However, there are some areas relevant or related to this problem, which we discuss in the following.

On a fundamental level, our work is based on the notion that both atomic and composite services exhibit some measurable quality (QoS). Monitoring QoS has been an active research area for some time. Different techniques proposed in this direction include monitoring based on client feedback [25], monitoring of TCP-level metrics using network analysis techniques [15] or event-based monitoring based on event-condition-action rules [26]. We use the VRESCO event engine and event-based monitoring in a manner very similar to the approach presented in [26].

The PREVENT approach aims at autonomous optimization of service compositions with regards to SLA violations and costs of adaptation. This bears a natural resemblance to the idea of QoS optimization for service compositions, as prominently described in [27]. Later approaches tried to improve on this concept by using more efficient heuristic algorithms, e.g., H1_RELAX_IP [28] (a heuristic relaxation of integer programming), WFlow [29] (based on stochastic workflow reduction) or the immune algorithm [30]. Different authors approached the problem by combining global optimization and local selection (which can be done much more efficient than global optimization). This approach can also be considered a heuristic, because the combination with local selection does not guarantee a globally optimal solution [31]. Most comparably to our work, the authors of [32] use a genetic algorithm combined with local search to efficiently solve the QoS optimization problem. The main difference of our work to all these approaches is that we do not optimize the composition with regard to global QoS goals. Instead, our optimization goal is to minimize the costs resulting from SLA violations and adaptations. Therefore, in our work, some SLAs are allowed to be violated if it is financially desirable for the provider to do so. Hence, the optimization problem we have to solve is different.

To our work, even more important than the measurement of past quality is the prediction of future QoS. One well-known approach to establishing predictable QoS levels in a composite service is QoS aggregation, i.e., the process of calculating the quality dimensions of a composite service based on the QoS of the utilized services and aggregation functions. QoS aggregation has for instance been discussed in [20]. The concept of QoS aggregation has been extended to SLA aggregation by several authors [33], [34]. As an alternative to QoS and SLA aggregation, different authors have proposed to use various machine learning techniques to predict composition QoS from monitored runtime data [6], [7]. This approach is also the one that we use in PREVENT, as explained in Section 3. The main advantage that we see in using machine learning is that it is very easy to incorporate non-QoS data (composition instance data, such as customer identifiers or ordered products) without the need to explicitly specify aggregation rules describing how this data influence the composition performance. However, note that the contribution discussed in this paper is in principle agnostic of the actual approach used for prediction of violations.

Generally, PREVENT is a system to monitor and prevent SLA violations. In this area, some works exist, which discuss the runtime monitoring of composition quality, such as [35]. This paper is of particular interest to us, because it discusses an integrated approach toward monitoring based on events. As stated above, this is quite related to monitoring in PREVENT. These works do not attempt to explain the reasons for SLA violations, and neither do they try to prevent them. The MoDe4SLA approach [4] is a top-down approach toward identifying

these influential factors of SLA violations. Research in a similar direction, but using data mining techniques instead of top-down analysis, has also been presented in [5]. Our work is different in that we do not only try to identify which parts of a service composition cause SLA violations, but actively prevent them by applying targeted adaptation actions. Therefore, our system essentially implements the paradigm of self-adapting service compositions. This is related to the area of flexible service composition, as introduced in [36]. Flexible service compositions reoptimize their composition at runtime, to deal with unanticipated problems. Similar ideas (self-healing processes) have also been presented as part of the DISC framework [37], which implements dynamic and only partially defined processes. A different kind of self-healing processes have been discussed in [38]. In this paper, the authors present the VieDAME framework, which autonomously monitors the QoS of services used in the composition, and triggers service re-selection if the monitored QoS falls below a given threshold. This is similar to the PREVENT approach, but our system supports a wider range of adaptation actions (as discussed in our earlier work [8]). Additionally, [38] does not take the costs of adaptation into account. Another middleware for self-adapting compositions is MASC [39]. However, the authors of this paper focus more on adaptation for functional reasons, while our main goal is the optimization of nonfunctional aspects. Furthermore, the MASC system also does not explicitly incorporate costs of adaptation.

The core contribution of this paper is the notion that there generally is a tradeoff to consider between preventing SLA violations and the costs of doing so. Hence, a composite service provider is maximizing his own revenue by minimizing his total costs. Similar models have been investigated in many related areas before. For instance, Mazzucco et al. [40] describe a model for revenue maximizing in web services hosting using dynamic admission policies. Similarly, techniques to optimize application servers in a way to maximize the provider profit in distributed systems have been proposed in [41]. Other tradeoffs that have been discussed in the literature include the performance-security tradeoff [42] or the tradeoff between composition QoS and the costs of monitoring [43].

## 8 Conclusion and Future Work

For providers of composite web services, it is essential to be able to minimize cases of SLA violations. One possible route to achieve this is to predict at runtime, which instances are in danger of violating SLAs, and to apply various adaptation actions to these instances only. However, it is not trivial to identify which adaptations are the most cost-effective way to prevent any violation, or if it is at all possible to prevent a violation in a cost-effective way. In this paper, we have modeled this problem as a one-dimensional discrete optimization problem. Furthermore, we have presented both, deterministic and heuristic solution algorithms. We have evaluated these algorithms based on a manufacturing case study and have shown which types of algorithms are better suited for which scenarios.

The main current limitation is that adaptation is only considered on instance level, that is, for each composition instance separately. Aggregate SLOs, which are defined over a number of instances, are out of scope. Similarly, at the moment we do not consider "permanent" adaptations, i.e., adaptations which are done for all future instances. We believe that the PREVENT adaptation model can be extended to this kind of SLOs and actions, but new approaches to predict violations and impact models are needed to this end.

## Acknowledgments

The research leading to these results received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreements 215483 (S-Cube) and 257483 (Indenica).

## References

[1] M.P. Papazoglou, P. Traverso, S. Dustdar, and F. Leymann, "Service-Oriented Computing: State of the Art and Research Challenges," *Computer*, vol. 40, no. 11, pp. 38-45, Nov. 2007.

[2] A. Lenk, M. Klems, J. Nimis, S. Tai, and T. Sandholm, "What's Inside the Cloud? An Architectural Map of the Cloud Landscape," *Proc. ICSE Workshop Software Eng. Challenges of Cloud Computing (CLOUD '09)* pp. 23-31, 2009.

[3] A. Dan, D. Davis, R. Kearney, A. Keller, R. King, D. Kuebler, H. Ludwig, M. Polan, M. Spreitzer, and A. Youssef, "Web Services on Demand: WSLA-Driven Automated Management," *IBM Systems J.*, vol. 43, no. 1, pp. 136-158, Jan. 2004.

[4] L. Bodenstaff, A. Wombacher, M. Reichert, and M.C. Jaeger, "Analyzing Impact Factors on Composite Services," *Proc. IEEE Int'l Conf. Services Computing (SCC '09)*, pp. 218-226, 2009.

[5] B. Wetzstein, P. Leitner, F. Rosenberg, S. Dustdar, and F. Leymann, "Identifying Influential Factors of Business Process Performance Using Dependency Analysis," *Enterprise Information Systems*, vol. 4, no. 3, pp. 1-8, July 2010.

[6] P. Leitner, B. Wetzstein, F. Rosenberg, A. Michlmayr, S. Dustdar, and F. Leymann, "Runtime Prediction of Service Level Agreement Violations for Composite Services," *Proc. Third Workshop Non-Functional Properties and SLA Management in Service-Oriented Computing (NFPSLAM-SOC '09)*, pp. 176-186, 2009.

[7] L. Zeng, C. Lingenfelder, H. Lei, and H. Chang, "Event-Driven Quality of Service Prediction," *Proc. Sixth Int'l Conf. Service-Oriented Computing (ICSOC '08)*. pp. 147-161, 2008.

[8] P. Leitner, A. Michlmayr, F. Rosenberg, and S. Dustdar, "Monitoring, Prediction and Prevention of SLA Violations in Composite Services," *Proc. IEEE Int'l Conf. Web Services (ICWS '10)*, pp. 369-376, 2010.

[9] P. Leitner, B. Wetzstein, D. Karastoyanova, W. Hummer, S. Dustdar, and F. Leymann, "Preventing SLA Violations in Service Compositions Using Aspect-Based Fragment Substitution," *Proc. Int'l Conf. Service-Oriented Computing (ICSOC '10)*, 2010.

[10] R. Kazhamiakin, B. Wetzstein, D. Karastoyanova, M. Pistore, and F. Leymann, "Adaptation of Service-Based Applications Based on Process Quality Factor Analysis," *Proc. Second Workshop Monitoring, Adaptation and Beyond (MONA+)*, pp. 395-404, 2009.

[11] "Business Process Modeling Notation Specification," technical report, Object Management Group, 2006.

[12] M. Salehie and L. Tahvildari, "Self-Adaptive Software: Landscape and Research Challenges," *ACM Trans. Autonomous and Adaptive Systems*, vol. 4, no. 2, pp. 1-42, May 2009.

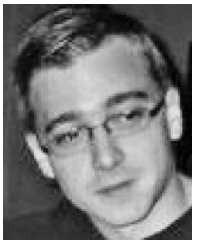
[13] A. Michlmayr, F. Rosenberg, P. Leitner, and S. Dustdar, "End-to-End Support for QoS-Aware Service Selection, Binding, and Mediation in VRESCo," *IEEE Trans. Services Computing*, vol. 3, no. 3, pp. 193-205, July 2010.

[14] J.O. Kephart and D.M. Chess, "The Vision of Autonomic Computing," *Computer*, vol. 36, no. 1, pp. 41-50, Jan. 2003.

[15] F. Rosenberg, C. Platzer, and S. Dustdar, "Bootstrapping Performance and Dependability Attributes of Web Services," *Proc. IEEE Int'l Conf. Web Services (ICWS '06)*, pp. 205-212, 2006.

[16] S. Haykin, *Neural Networks and Learning Machines: A Comprehensive Foundation,* third ed. Prentice Hall, 2008.
[17] J.R. Quinlan, "Induction of Decision Trees," *Machine Learning,* vol. 1, pp. 81-106, Mar. 1986.
[18] D. Ivanovic, M. Carro, and M. Hermenegildo, "Towards Data-Aware QoS-Driven Adaptation for Service Orchestrations," *Proc. IEEE Int'l Conf. Web Services (ICWS '10),* pp. 107-114, 2010.
[19] L. Juszczyk and S. Dustdar, "Script-Based Generation of Dynamic Testbeds for SOA," *Proc. IEEE Int'l Conf. Web Services (ICWS '10),* pp. 195-202, 2010.
[20] M.C. Jaeger, G. Rojec-Goldmann, and G. Muhl, "QoS Aggregation for Web Service Composition Using Workflow Patterns," *Proc. Eighth Int'l Enterprise Distributed Object Computing Conference (EDOC '04),* pp. 149-159, 2004.
[21] L. Xu and B. Jennings, "A Cost-Minimizing Service Composition Selection Algorithm Supporting Time-Sensitive Discounts," *Proc. IEEE Int'l Conf. Services Computing (SCC '10),* pp. 402-408, 2010.
[22] T. Feo and M. Resende, "Greedy Randomized Adaptive Search Procedures," *J. Global Optimization,* vol. 6, pp. 109-133, 1995.
[23] D.E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning.* Addison-Wesley Professional, 1989.
[24] N. Radcliffe and P. Surry, "Formal Memetic Algorithms," *Evolutionary Computing,* vol. 865, pp. 1-16, 1994.
[25] R. Jurca, B. Faltings, and W. Binder, "Reliable QoS Monitoring Based on Client Feedback," *Proc. 16th Int'l Conf. World Wide Web (WWW '07),* pp. 1003-1012, 2007.
[26] L. Zeng, H. Lei, and H. Chang, "Monitoring the QoS for Web Services," *Proc. Fifth Int'l Conf. Service-Oriented Computing (ICSOC '07),* pp. 132-144, 2007.
[27] L. Zeng, B. Benatallah, A.H.H. Ngu, M. Dumas, J. Kalagnanam, and H. Chang, "QoS-Aware Middleware for Web Services Composition," *IEEE Trans. Software Eng.,* vol. 30, no. 5, pp. 311-327, May 2004.
[28] R. Berbner, M. Spahn, N. Repp, O. Heckmann, and R. Steinmetz, "Heuristics for QoS-Aware Web Service Composition," *Proc. IEEE Int'l Conf. Web Services (ICWS '06),* pp. 72-82, 2006.
[29] T. Yu, Y. Zhang, and K.-J. Lin, "Efficient Algorithms for Web Services Selection with End-to-End QoS Constraints," *ACM Trans. the Web,* vol. 1, article 6, May 2007.
[30] J. Xu and S. Reiff-Marganiec, "Towards Heuristic Web Services Composition Using Immune Algorithm," *Proc. IEEE Int'l Conf. Web Services (ICWS '08),* pp. 238-245, 2008.
[31] M. Alrifai and T. Risse, "Combining Global Optimization with Local Selection for Efficient QoS-Aware Service Composition," *Proc. 18th Int'l Conf. World Wide Web (WWW '09),* pp. 881-890, 2009.
[32] F. Rosenberg, M.B. Müller, P. Leitner, A. Michlmayr, A. Bouguettaya, and S. Dustdar, "Metaheuristic Optimization of Large-Scale QoS-Aware Service Compositions," *Proc. IEEE Int'l Conf. Services Computing (SCC '10),* 2010.
[33] T. Unger, F. Leymann, S. Mauchart, and T. Scheibler, "Aggregation of Service Level Agreements in the Context of Business Processes," *Proc. 12th Int'l Enterprise Distributed Object Computing Conf. (EDOC '08),* pp. 43-52, 2008.
[34] I. Haq, A. Huqqani, and E. Schikuta, "Aggregating Hierarchical Service Level Agreements in Business Value Networks," *Proc. Seventh Int'l Conf. Business Process Management (BPM '09),* pp. 176-192, 2009.
[35] L. Baresi, S. Guinea, M. Pistore, and M. Trainotti, "Dynamo + Astro: An Integrated Approach for BPEL Monitoring," *Proc. IEEE Int'l Conf. Web Services (ICWS '09),* pp. 230-237, 2009.
[36] D. Ardagna, M. Comuzzi, E. Mussi, B. Pernici, and P. Plebani, "PAWS: A Framework for Executing Adaptive Web-Service Processes," *IEEE Software,* vol. 24, no. 6, pp. 39-46, Nov./Dec. 2007.
[37] E. Zahoor, O. Perrin, and C. Godart, "DISC: A Declarative Framework for Self-Healing Web Services Composition," *Proc. IEEE Int'l Conf. Web Services (ICWS '10),* pp. 25-33, 2010.
[38] O. Moser, F. Rosenberg, and S. Dustdar, "Non-Intrusive Monitoring and Service Adaptation for WS-BPEL," *Proc. 17th Int'l Conf. World Wide Web (WWW '08),* pp. 815-824, 2008.
[39] A. Erradi, P. Maheshwari, and V. Tosic, "Policy-Driven Middleware for Self-Adaptation of Web Services Compositions," *Proc. ACM/IFIP/USENIX Int'l Conf. Middleware (Middleware '06),* pp. 62-80, 2006.
[40] M. Mazzucco, I. Mitrani, J. Palmer, M. Fisher, and P. McKee, "Web Service Hosting and Revenue Maximization," *Proc. Fifth European Conf. Web Services (ECOWS '07),* pp. 45-54, 2007.
[41] D. Villela, P. Pradhan, and D. Rubenstein, "Provisioning Servers in the Application Tier for E-Commerce Systems," *ACM Trans. Internet Technology,* vol. 7, no. 1, article 7, 2007.
[42] S.S. Yau, Y. Yin, and H.G. An, "An Adaptive Tradeoff Model for Service Performance and Security in Service-Based Systems," *Proc. IEEE Int'l Conf. Web Services (ICWS '09),* pp. 287-294, 2009.
[43] Y. Zhang, M. Panahi, and K.-J. Lin, "Service Process Composition with QoS and Monitoring Agent Cost Parameters," *Proc. IEEE 10th Conf. E-Commerce Technology and the Fifth IEEE Conf. Enterprise Computing, E-Commerce and E-Services,* pp. 311-316, 2008.

**Philipp Leitner** received the BSc and MSc degrees in business informatics from Vienna University of Technology. He is currently working toward the PhD degree and is a university assistant in the Distributed Systems Group, Vienna University of Technology. His research is focused on middleware for distributed systems, especially for SOAP-based and RESTful web services. He is a member of both the IEEE and the IEEE Computer Society.



**Waldemar Hummer** received the BSc degree from the University of Innsbruck and the MSc degree from the Vienna University of Technology, both in computer science, and the BSc degree in business administration, from the Vienna University of Economics and Business. He is currently working toward the PhD degree and is a university assistant for the Distributed Systems Group, Vienna University of Technology. His primary topics of interest are in the areas of self-optimizing service-based systems, web service composition, and web data aggregation. He is a student member of the IEEE.

**Schahram Dustdar** is a full professor of computer science with a focus on Internet technologies, heading the Distributed Systems Group, Vienna University of Technology. He is also an honorary professor of information systems in the Department of Computing Science at the University of Groningen, The Netherlands. He is a senior member of both the IEEE and the IEEE Computer Society.