Journal of Computational Science 4 (2013) 232–241

Contents lists available at SciVerse ScienceDirect

Journal of Computational Science

journal homepage: www.elsevier.com/locate/jocs

Bounding models families for performance evaluation in composite Web services

Serge Haddad[a], Lynda Mokdad[b,*], Samir Youcef[c]

[a] LSV, ENS de Cachan, Cachan, France
[b] LACL, University of Paris-Est, Créteil, France
[c] LORIA-INRIA-UMR 7503, Nancy, France

ARTICLE INFO

Article history:
Received 31 July 2011
Received in revised form 14 October 2011
Accepted 24 November 2011
Available online 10 January 2012

Keywords:
Composite Web services
Continuous time Markov chains
Stochastic coupling

ABSTRACT

One challenge of composite Web service architectures is the guarantee of the Quality of Service (QoS). Performance evaluation of these architectures is essential but complex due to synchronizations inside the orchestration of services. We propose methods to automatically derive from the original model a family of bounding models for the composite Web response time. These models allow to find the appropriate trade-off between accuracy of the bounds and the computational complexity. The numerical results show the interest of our approach w.r.t. complexity and accuracy of the response time bounds.

1. Introduction

A Web service essentially denotes an application (or program) accessible via Internet standard protocols [3]. The basic protocol used to access Web services is SOAP (*Simple Object Access Protocol*), an XML (*eXtensible Markup Language*) based protocol that allows a service customer to invoke services [6]. The Web services, called elementary Web services, such as described by WSDL (*Web Service Description Language*), are conceptually limited to relatively simple functionalities modeled through a collection of simple operations without control flow. For certain types of applications, it is necessary to combine a set of elementary Web services to obtain more complex one, called aggregated or composite Web services, in order to meet customer requirements [4,1]. This aggregation is possible using for example BPEL (*Business Process Execution Language For Web Services*) standard which is the result of merging previous composition languages such like WSFL (*Web Service Flow Language*) and XLANG (*XML Business Process Language*) [5].

The Web services are executable on more platforms than the previous technologies such like CORBA (*Common Object Request Broker Architecture*) and RMI (*Remote Method Invocation*), however they raise new requirements like the dynamic adaptability and the insurance of QoS (*Quality of Service*). The latter criterium is defined as a combination of several attributes which can be qualitative (e.g. security) and quantitative (e.g. response time [9,10]). Their increasing complexity requires the development of methods and tools in order to monitor and evaluate their QoS. In fact, the QoS degradation can lead to serious consequences including a significant economic impact.

In this paper, we focus on the composite Web service (CWS) response time computation, where the requests are decomposed into sub-queries to different elementary Web services and then merged into a final result. In our previous study [11], we have considered the BPEL constructors directly supported by this standard. The control patterns considered here are not directly supported by BPEL:

- parallel invocation of a constant number of elementary Web services merged by a federation component (see Fig. 1),
- parallel invocation of a variable number of elementary Web services merged by a federation component.

Under the assumption of Markovian elementary service and merging times, the modeling of a composite Web service yields a Markov chain with $O(n^2)$ states, where n is the number of invoked elementary Web services. The particular structure of this Markov chain allows us to establish recurrence equations which lead to a computational complexity time of order $O(n^2)$. It is because of the structure of the obtained Markov chain (see Section 4). In the open systems such as peer to peer environment, the number of invoked services can lie between 10^3 and 10^6. So, their exact analysis becomes difficult and often intractable. Using the stochastic comparison [7,8] and more precisely the coupling process technique, we propose a generic transformation of the studied Markov chain which guarantees that the response time of the new Markov chain is an upper bound of the initial Markov chain response time. We

[*] Corresponding author.
E-mail address: lynda.mokdad@univ-paris12.fr (L. Mokdad).

1877-7503/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jocs.2011.11.003

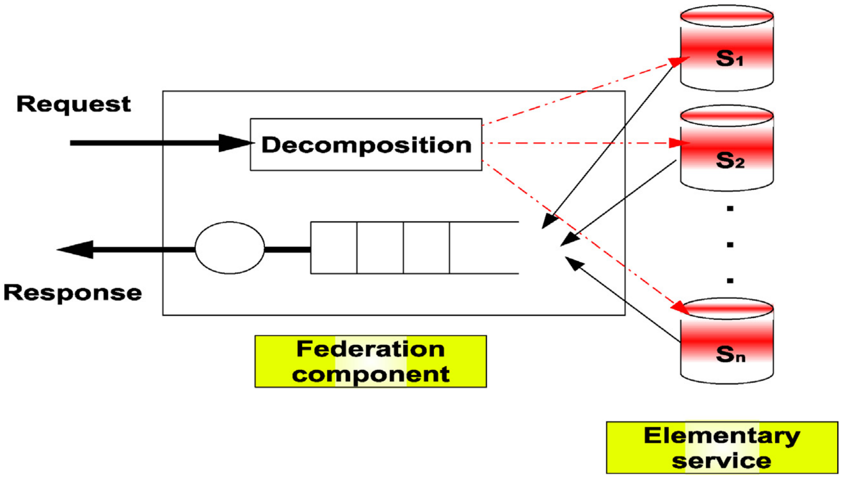

Fig. 1. Data processing in a composite Web service platform.

instantiate this transformation in three ways, where each obtained new Markov chain is parameterized by a "quantitative" parameter. Otherwise stated, we propose three families of the bounding models. By an appropriate choice of the parameter, the recurrence equation systems can be resolved with an algorithm with $O(n)$ and $O(n\sqrt{n})$ respectively space complexity and time complexity.

Moreover, we show by empirical studies that depending on the numerical values of the original Markov chain, the bound provided by any of the three bounding families can be better than the two other ones. We also characterize the three cases w.r.t. these numerical values.

We generalize our work as follows. Assume that an elementary Web service can be invoked with a constant probability. Thus the response time of a given composite Web service can be computed as a weighted sum of the elementary Web services response times where the computational cost is of order $O(n^3)$. To handle this case, we combine the upper bounds proposed for the first pattern and the Chernoff bounds in order to limit the study only to two cases: the case where all services are invoked and a probabilistic "worst case" i.e. a constant number of invoked services with a very small probability to exceed this threshold. This approach allows us to keep the same computational cost as in the first case (i.e. when the number of elementary Web services is constant). Note that Chernoff bounds give bounds on the tail distributions of the sums of independent random variables (more details are given in Section 3).

The rest of the paper is organized as follows. Section 2 presents some related works. Section 3 recalls some definitions and results related to the process coupling technique and the Chernoff bounds. In Section 4, we study the control pattern parallel invocation where the number of elementary Web services is constant. Section 5 summarizes the obtained numerical results in this case. In Section 6, we study the control pattern parallel invocation where the number of elementary Web services is variable. Section 7 summarizes the obtained numerical results in this case. Section 8 summarizes the contributions of this paper and gives some perspectives to this work.

2. Related work

In the framework of Web services performance evaluation, two approaches are generally used: benchmarking and modeling methods. In the following, we present some studies using the two approaches.

As far as performance measurement of Web services is concerned, XML specification and SOAP protocol have been studied in [21–23] by testing and measuring of SOAP-based Web services response time. A comparative study on response time and throughput with existing protocols, like RMI, RMI/IIOP or CORBA/IIOP, is presented in [21]. A critical study of XML-based protocols for Web services is presented and binary encoded protocol has been proposed instead of text XML-based ones in [22]. In [24], information about past workflow executions is collected in a log. Starting from this log a continuous Markov chain is derived, in order to compute the execution response time and the cost of this workflow.

In [10], the composite Web service response time is considered as a response time of fork and join model. This model states that a single Internet application can invoke in parallel a set of elementary Web services and gather their responses from all these launched services in order to return the results to a client. In this considered study, authors analyze the effects of exponential response times based on earlier work in [12]. An exact analysis of fork and join system is possible when the system is significantly simplified. This is the case for example when the job arrival process in the system follows a Poisson distribution with execution task having exponential distribution and the number of queues is equal to two. The exact computation response time of a such system can be found in [13–15]. An approximation technique has been proposed in the case where the number of servers is greater than two and the servers are homogeneous [15]. This last study is extended in [16]. General arrival process and services times are considered in [17]. The most general case is considered in [18]. In this work, upper and lower bounds are proposed by assuming that the response times in each queue are mutually independent. Two approximation techniques are presented: one is based on a decomposition approach and the other is based on an iterative solution method.

In order to overcome the limitations of these studies and particularly the one presented in [10], we have proposed a general model taking into account the fact that elementary Web services are heterogenous and the number of invoked services can be variable (this is the case when we use for example the BPEL multi-choice constructor) [11]. More recently, the problem of computing the distribution of the throughput time in workflow nets has been studied in [20]. In this paper, authors consider workflow with transition execution time having exponential distributions and formulas have been proposed for each refinement rule (sequence, parallel, synchronization and loop execution pattern). Response time of a Web service middleware is considered in [19], which follows a fork and join model of execution. The author proposes that while performing