[Eck08] *Julian Eckert;* **Resource Planning for Distributed Service-Oriented Workflows.** In: 3rd International Conference on Software and Data Technologies (ICSOFT 2008), INNSTICC PRESS, Portugal, July 2008. Seite 38-45.

RESOURCE PLANNING FOR DISTRIBUTED SERVICE-ORIENTED WORKFLOWS

Julian Eckert
Technische Universität Darmstadt, Multimedia Communications Lab, Merckstr. 25, 64285 Darmstadt, Germany
julian.eckert@KOM.tu-darmstadt.de

Keywords: Distributed Workflows, Resource Planning, Service-oriented Computing, Service Composition, Quality of Service

Abstract: Collaborations between enterprises cause the need for cross-organizational workflows that can be realized by adopting the Service-oriented Architecture paradigm. In order to cope with the challenge to ensure several Quality of Service (QoS) demands during workflow composition, performance evaluation and execution management of service-oriented workflows became quite important in order to avoid performance degradation. In a distributed workflow scenario with services from external partners, resource planning of services is crucial to ensure that the workflow execution remains feasible and that Service Level Agreement (SLA) violations due to overload are avoided.
My research focuses on the development of a holistic resource planning approach that facilitates optimal compositions of services to workflows depending on various customer demands and request priorities, different pricing models, and several QoS requirements. Besides a worst-case and an average-case performance analysis including optimization models, I am working in my research towards a detailed resource planning approach in order to ensure that all incoming execution requests to a distributed service-oriented workflow can be served at minimal costs.

1 RESEARCH PROBLEM

Nowadays, cost-efficient business processes are crucial concerning competitive markets and deregulation of markets. These processes have to be adapted to changing environments and to varying business needs. Besides cost-efficiency, also performance issues of business processes and workflows, as an IT supported business process, are important regarding its competitiveness.

In order to support performance issues, enterprises require a continuous business process management that supports business process intelligence. Further, the capability of a flexible composition of a business process consisting of several services has to be supported. Not only intra company workflows have to be addressed, also cross-organizational workflows become more and more important due to the growing amount of enterprise collaborations. Concerning cross-organizational workflows in which multiple parties and even external partners are involved, flexible business processes can be achieved by integrating internal legacy systems, as well as by coupling external business partners. A Service-oriented Architecture (SOA) as an architectural blueprint facilitates the realization of such cross-organizational, distributed workflows and facilitates the required process flexibility. In order to meet customer expectations, Quality of Service (QoS) issues play an important role concerning an effective business process management (Almeida and Menascé, 2002). Competitive enterprises have to plan their business processes in advance to be able to adapt them to changing business needs. Workflows may use Web services as an open standard to realize business processes (Alonso et al., 2004). In order to avoid the risk of poor workflow performance in a distributed service-based scenario, resource planning, business process management, and performance analysis are important. Figure 1 shows the considered scenario of my research with

various workflow requestors, one intermediary who gathers all workflow requests and prioritizes requests, and several service providers.

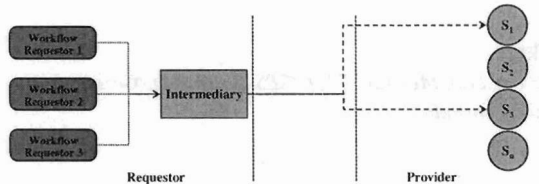

Figure 1: Research scenario

Due to possible unexpected peaks and workloads a continuous performance analysis of a service-oriented workflow becomes important in order to be able to plan the workflow execution in advance. Beside the QoS-aware composition of one workflow execution, performance analysis concerning all workflow requests becomes more and more important. Each workflow requestor requests a specific amount of workflow executions with specific deadlines and with specific QoS requirements. The intermediary has to invoke appropriate services within two steps. On the one hand the workflow controller has to select the services which fulfill the required functionality of the specific task of the process. On the other hand non-functional properties such as QoS or costs have to be considered as well in order to optimize the overall performance properties of all workflow requests within a specific response time.

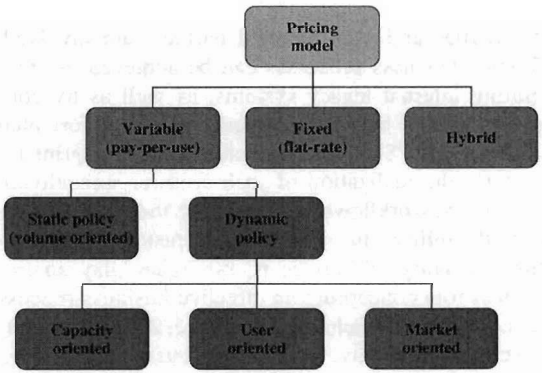

Figure 2: Pricing models for services

Further, the intermediary has to handle several pricing models of the service providers as depicted in Figure 2. This implies that the intermediary has to handle a complex service invocation problem based on several request properties, request deadlines, QoS requirements, and pricing models.

2 OUTLINE OF OBJECTIVES

My research is aiming at the development of a holistic resource planning approach facilitating a cost-efficient execution of workflow requests in service-based environments. The main question is how to serve a large amount of incoming workflow execution requests at minimal costs by meeting the requested QoS requirements. Further, how can potential performance bottlenecks of distributed workflows be predicted and avoided in order to facilitate the execution of all incoming workflow requests? The main topics of my research include:

- Optimization of the worst-case performance behavior of service-oriented workflows
- Optimization of the average-case performance behavior of service-oriented workflows
- Workload prediction and request prioritization of service-oriented workflows
- Development of an integrated resource planning approach facilitating the execution of several workflows
- Prototypical implementation of the architecture necessary to realize resource planning

In particular, the scenario of distributed workflows is considered, i.e., a large amount of service requestors and providers. This results in various requestor demands, a high amount of workflow execution requests, several pricing models, and several QoS requirements.

3 EXPECTED OUTCOME

In order to develop a holistic resource planning approach for distributed workflows several contributions are planned or already realized.

As a foundation Network Calculus as a system theory for deterministic queuing systems has been analyzed with respect to service-oriented workflows. The concept of arrival curves, service curves, delay bound, and backlog bound has been adapted and extended to service selection scenarios for distributed workflows. Furthermore, worst-case performance optimization models have been developed that facilitate compliance with several Service Level Agreements (SLAs) and ensure that several QoS properties are met even in case of a huge amount of incoming workflow execution requests.

Besides the worst-case observation, the average-case performance of distributed workflows has been analyzed including the development of average-case

performance optimization models by adapting findings of Queuing Theory to service selection scenarios for distributed workflows. Besides focusing on the worst-case and the average-case workflow performance in a scenario with several workflow requestors a holistic workload estimation and prediction approach has to be developed in order to be able to plan the invocation of all resources (services) at minimal costs. Furthermore, a detailed resource planning approach including optimization models and heuristics has to be developed that determines optimal execution plans and that make decisions which services have to be invoked at which step in the process and how many service have to be parallelized in order to be able to serve all incoming workflow execution requests from several service requestors.

The developed approach should be able to determine worst-case as well as average-case performance behavior and to plan the service invocation based on specifically developed forecasts of incoming workflow execution requests by a detailed resource planning approach. The developed optimization models, heuristics, and algorithms will be evaluated by extensive large-scale simulations.

4 RESEARCH METHODOLOGY

The development of a holistic resource planning approach requires the application of more than one single research methodology. Thus, my research applies several research methodologies to the aforementioned resource planning problem. The following list gives an overview of the contribution and the research methodology / approach used:

- Conceptual work to analyze the research problem and to design the architectural extension necessary as well as the performance estimation and workload analysis
- Web service measurements in order to be able to determine worst-case service curves as well as to determine the overall worst-case behavior
- Analytical evaluation and optimization of the worst-case and the average-case performance behavior as well as analytical models for the detailed resource planning approach
- Simulation to evaluate the service selection and resource planning approach for the considered scenario

By combining all mentioned research methodologies it becomes possible to treat all research questions addressed before.

5 STATE OF THE RESEARCH

In this section the major contributions of my research are presented in summary, i.e., on the one hand the worst-case workflow performance optimization approach and on the other hand the average-case workflow performance optimization approach. In addition, an outlook on an architectural extension as an enhancement to WSQoSX.KOM (Berbner et al., 2007) is presented.

5.1 Worst-case workflow performance optimization

Due to Web service performance observations it is possible to determine a worst-case behavior of Web service calls with a specific latency and a specific execution rate. By applying results from Network Calculus (Boudec and Thiran, 2001) to this scenario it is possible to describe the behavior of a Web service with a service curve as well as the behavior of incoming workflow execution requests with an arrival curve.

Packet switched network	Workflow Management System
Path through the network	Workflow
Node in the network	Web service
Packet	Request
Throughput	Rate at which requests can be processed
End-to-end delay	Time until a workflow is completed

Table 1: Analogies between a packet switched network and the considered WFMS

In Table 1 the analogies of a packet switched network and a Workflow Management System (WFMS) are depicted. The adaptation of Network Calculus to service-oriented workflows enables the analysis of the worst-case performance behavior of the workflow by the convolution of the service curves of the invoked services. The convolution of the aggregated service curve with the arrival curves determines the shape of the executed workflow requests. In Eckert et al. (Eckert et al., 2007) (Eckert et al., 2008a) the optimization of the worst-case performance behavior due to several bounds and objective functions is shown in a way that several QoS requirements are met and that the workflow execution remains feasible and cost-efficient.

5.2 Average-case workflow performance optimization

For the analysis of the execution capacity of Web service workflows and the planning of the workflow control, Queuing Theory can be used to describe the average performance behavior of a workflow. In this context resource planning and performance measurement are crucial to ensure that the workflow execution remains feasible and SLA violations due to overload are avoided.

In general, the principle of queuing models is well studied in the literature (Harverkort, 1998). For the application of Queuing Theory to the considered scenario we assume that the workflow execution requests are distributed in a Poisson manner with a specific arrival rate λ and that the request interarrival times are exponentially distributed. Considering several parallel incoming Poisson processes the resulting process is also a Poisson process. At the Web services the service times are negatively exponentially distributed with the mean service time μ that denotes the rate at which the jobs are processed. With the help of Burke's theorem and the concept of the Feed Forward Queuing Networks (FFQN) it is possible to determine the overall system behavior as depicted in Figure 3 with the assumption that the arrival rate is always smaller than the service rate of each invoked Web service in order to avoid overload at the Web services.

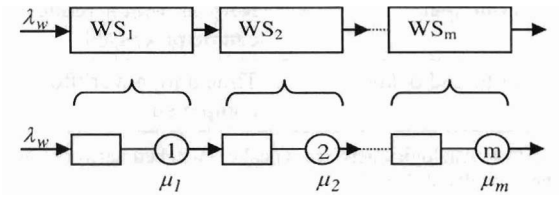

Figure 3: FFQN model for workflows

An optimization model for the maximization of the service utilization of the invoked Web services is shown in Eckert et al. (Eckert et al., 2008b). Further, this model includes constraints as costs and overall average response time in order to guarantee that the requests are processed within a specific time period. With this approach it is possible to handle an average-case scenario. In addition, the approach facilitates to determine which services have to be invoked at which step in the workflow in order to guarantee specific SLA requirements for the workflow requestors at minimal costs.

5.3 Architectural extensions

Our past research aims at service selection models and heuristics for the optimization of QoS parameters for one workflow request (Berbner et al., 2006a) and (Berbner et al., 2006b). The question is how to develop an architecture that is capable to handle several incoming workflow execution requests with several QoS requirements and with prioritized requests. For this purpose it is mandatory to extend the existing architecture WSQoSX.KOM with components as a *Workload Predictor* and a *Resource Planner*.

A Workflow Predictor should monitor incoming execution requests, and achieve a forecast for future workflow execution requests with quantitative forecasting methods as linear regression or nonlinear methods. The Resource Planning Component has to determine optimal service compositions and has to check whether the execution capacity of some services is reached, should replace services by other services and has to specify which services have to be invoked in parallel in order to be able to serve all incoming requests.

6 STATE OF THE ART

The main topic *service composition* has been widely studied in the literature in the past. Many approaches only focus on service composition problems for one workflow request and do not consider request prioritization as well as several cost models.

Several approaches for the QoS-aware composition of services with the help of ontologies and artificial intelligence planner exist (Naseri and Towhidi, 2007). Further, those composition problems are also solved with the help of genetic algorithms (Canfora et al., 2005). A predictive QoS model to compute the QoS for workflows in terms of performance, cost, and reliability is shown by Cardoso et al. (Cardoso et al., 2004). An approach for the adaptation of Queuing Theory to composite Web services and the implications of the capacity planning process can be found in Peng et al. (Peng et al., 2004). Due to various workflow requestors, task deadlines as constraints can be introduced. An approach for the QoS-optimization in such service-based systems is shown by Orleans and Furtado (Orleans and Furtado, 2007).

7 CONCLUSION AND OUTLOOK

In a cross-organizational workflow scenario with various workflow requestors, an intermediary, and sev-

eral service providers, it has to be guaranteed that all workflow requests can be served at minimal costs with specific QoS guarantees. Therefore, it is mandatory to develop a holistic resource planning approach that handles all incoming workflow execution requests and make decisions which services have to be invoked at which step in the workflow and which services have to be parallelized.

The aim of my presented research in this paper is to develop such a holistic approach. During my work, I contributed or currently contribute the following:

- Analytical optimization models concerning the worst-case performance behavior of service-oriented workflows
- Analytical optimization models concerning the average-case performance behavior of service-oriented workflows
- A workload forecast and prioritization model concerning workflow requests in a distributed environment
- Mechanisms and algorithms for an efficient resource planning
- An architectural extension facilitating the implementation of a holistic resource planning in a distributed workflow scenario

In particular, the architectural enhancements as well as the detailed resource planning approach have to be further developed. Finally, the developed optimization models, heuristics, and algorithms have to be evaluated by extensive large-scale simulations.

ACKNOWLEDGEMENTS

This work is supported in part by E-Finance Lab Frankfurt am Main e.V. (http://www.efinancelab.com).

REFERENCES

Almeida, V. A. F. and Menascé, D. A. (2002). Capacity Planning: An Essential Tool for Managing Web Services. *IT Professional*, 4(4):33–38.

Alonso, G., Casati, F., Kuno, H., and Machiraju, V. (2004). *Web Services - Concepts, Architectures and Applications*. Data-Centric Systems and Applications. Springer-Verlag.

Berbner, R., Spahn, M., Heckmann, O., and Steinmetz, R. (2007). WSQoSX A QoS Architecture for Web Service workflows. In *5th International Conference on Service-Oriented Computing (ICSOC 2007), Demo Track*.

Berbner, R., Spahn, M., Repp, N., Heckmann, O., and Steinmetz, R. (2006a). An Approach for Replanning of Web Service Workflows. In *Proceedings of the 12th Americas Conference on Information Systems (AMCIS'06)*.

Berbner, R., Spahn, M., Repp, N., Heckmann, O., and Steinmetz, R. (2006b). Heuristics for QoS-aware Web Service Composition. In *Proceedings of the 4th IEEE International Conference on Web Services*, pages 72–79.

Boudec, J.-Y. L. and Thiran, P. (2001). *Network Calculus. Number 2050 in Lecture Notes in Computer Science*. Springer-Verlag.

Canfora, G., Penta, M. D., Esposito, R., and Villani, M. L. (2005). An approach for QoS-aware service composition based on genetic algorithms. In *Proceedings of Genetic and Evolutionary Computation Conference*, pages 1069–1075.

Cardoso, J., Sheth, A., Miller, J., Arnold, J., and Kochut, K. (2004). Modeling Quality of Service for workflows and web service processes. *Web Semantics: Science, Services and Agents on the World Wide Web Journal*, 1:281–308.

Eckert, J., Pandit, K., Repp, N., Berbner, R., and Steinmetz, R. (2007). Worst-Case Performance Analysis of Web Service Workflows. In *9th International Conference on Information Integration and Web-based Application & Services (IIWAS 2007)*.

Eckert, J., Schulte, S., Niemann, M., Repp, N., and Steinmetz:, R. (2008a). Worst-Case Workflow Performance Optimization. In *3rd International Conference on Internet and Web Applications and Services (ICIW'08)*.

Eckert, J., Schulte, S., Repp, N., Berbner, R., and Steinmetz:, R. (2008b). Queuing-based Capacity Planning Approach for Web Service Workflows Using Optimization Algorithms. In *IEEE International Conference on Digital Ecosystems and Technologies (IEEE DEST 2008)*.

Harverkort, B. R. (1998). *Performance of Computer Communication Systems: A Model-Based Approach*. John Wiley & Sons Inc., New York.

Naseri, M. and Towhidi, A. (2007). QoS-Aware Automatic Composition of Web Services using AI Planners. In *Proceedings Second International Conference on Internet and Web Applications and Services*.

Orleans, L. F. and Furtado, P. N. (2007). Optimization for QoS on Web-Service-Based Systems with Task Deadlines. In *Third International Conference ion Autonomous Systems (ICAS 2007)*.

Peng, D., Yuan, Y., Yue, K., Wang, X., and Zhou, A. (2004). Capacity Planning for Composite Web Services Using Queuing Network-Based Models. *LNCS (2004) 3129*, pages 439–448.