Optimal Deployment of Workflows in Distributed Environments

Stein Desmet

Supervisor(s): Filip De Turck, Bruno Volckaert

I. Introduction

Over the last few years Web Services and Service Oriented Architecture (SOA) in general have rapidly gained in popularity, due to promises of increased flexibility and reusability. Services expose functionality over the network, using well-defined standards to establish service description, service discovery and service invocation. Service Oriented Architectures subsequently combine these services into higher-level services with new functionality. In this way, business processes can be easily implemented and automated using existing functionality. The automated business processes are often referred to as *workflows*.

However, an important but often unaddressed aspect in Service Oriented Architectures is the optimal allocation of the services to the available resources (Figure 1). Optimal resource usage is not only vital to scalability and eliminating bottlenecks, but also to the Quality-of-Service (QoS) that can be provided. Smart resource selection may result in improved reliability or faster execution times.

This is certainly true for media related environments, such as for example broadcasting companies [1]. Media services like video processing or video analyzing are computationally very intensive, often extremely long-running and exchange considerable amounts of data, making intelligent service and resource allocation of vital importance to implementing business processes. With the current shift from Standard-Definition towards High-Definition video material, this becomes even more stringent, as file sizes and execution times are easily quadrupled.

S. Desmet is with the Department of Information Technology, Ghent University (UGent), Gent, Belgium. E-mail: Stein.Desmet@intec.ugent.be .

II. Static approach

Determining this optimal selection of services is obviously non-trivial. This problem is related to both the mapping problem and the scheduling problem. Services need to be mapped to resources, and scheduling may be necessary in order to meet deadlines or to assure mutually exclusive use of a particular resource. Both the scheduling and the mapping problem are known to be NP-hard.

Techniques such as Integer Linear Programming (ILP) [2] exist, which are able to find the best solution, but these are not usable in practice, due to excessive running times. In other words, it is necessary to develop heuristics such as Line Scheduling that are able to find a a good solution in a reasonable amount of time. Note that techniques such as ILP are still very usefull, as they can be used to gauge how close a solution from a heuristic approaches the best solution.

Media related SOAs add to the complexity of the problem. Services are long running, but deadlines for workflows are often extremely important. For example, a broadcasting company will definitely want to air its shows on time. Services exchange large amounts of data, but this may happen by prefetching or by streaming the data. Both cases pose other requirements and demand a different approach.

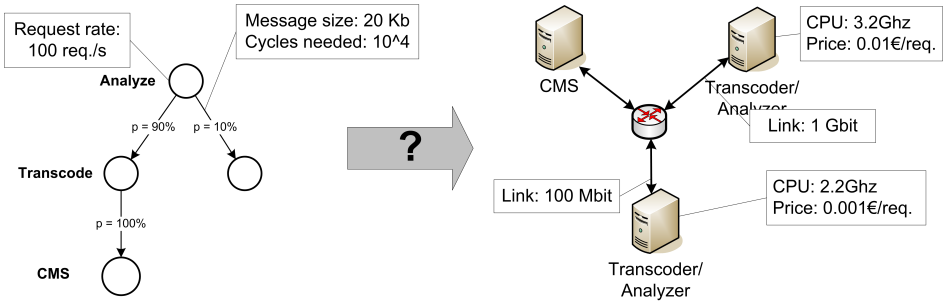

Figure 1: The problem of allocating of workflows to resources.

Data storage and replication is often important as well. These are but a few examples to illustrate the complexity of media SOAs.

III. Dynamic approach

Naturally, a given service allocation won't remain optimal forever. A company is not a static environment. It is quite possible and even likely that typical invocation patterns of services change over time, changes in the network occur, new QoS demands need to be met or new workflows are added. The altered situation obviously requires a new allocation, as the current solution will no longer be optimal. Ideally, finding this new assignment is done as fast as possible. A poor service allocation is detrimental to the company, as are long downtimes due to implementing the new allocation. Static algorithms can take time to find a new solution, and often cannot start from an existing situation It is therefore useful to develop dynamic optimisation algorithms that can handle continuously changing search spaces, so that optimal resource usage and QoS are guaranteed at all times.

IV. Dimensioning

When designing heuristics to find an optimal service allocation, it is also usefull to adapt them in order to allow performing dimensioning analyses. A dimensioning analysis can show the impact of removing or adding a resource, or can show where potential problems such as bottlencks are located. Such an analysis can also be used to determine the maximum load an infrastructure can handle.

V. Conclusions

Web Services and Service Oriented Architectures are becoming increasingly popular. Finding an optimal allocation of services and workflows to the available resources can yield substantial benefits to the host of the workflows. These benefits include improvements in the Quality-of-Service and more intelligent resource usage.

Several static optimization heuristics have been implemented for two different use cases [3]. Both use cases are situated in a media related context, in an intra- and intercompany context. Extensive Integer Linear Programming algorithms have been implemented as well. Preliminary results clearly show an improvement in resource usage, scalability and Quality-of-Service over random or ad-hoc service allocations.

References

[1] D. Van der Weken, S. Van Assche, D. Clabaut, S. Desmet, and B. Volckaert, "Automating workflows with service oriented media applications," *Services, IEEE Congress on*, vol. 0, pp. 507–514, 2009.

[2] H. Karloff, *Linear programming*, Birkhauser Boston Inc., Cambridge, MA, USA, 1991.

[3] S. Desmet, W. De Cock, S. Van Assche, D. Van der Weken, B. Muylaert, D. Clabaut, and K. Demarest, "Geisha work package 5 – service oriented architecture for media production," Tech. Rep., IBBT, 2009.