Web Services QoS: External SLAs and Internal Policies Or: How do we deliver what we promise?

Heiko Ludwig

IBM T.J. Watson Research Center

hludwig@us.ibm.com

Abstract

Whether offered within an organization or as a part of a paid service across organizational boundaries, quality of service (QoS) aspects of services are important in a service-oriented computing environment. While managing QoS in distributed systems is not a novel problem, a number of additional issues arise in the context of a service-oriented computing environment.

In the past years, we developed multiple means of describing, advertising and signing up to Web and Grid services at defined QoS levels. This includes HP's Web Services Management Language (WSML) and framework, IBM's Web Service Level Agreement (WSLA) language, the Web Services Offer Language (WSOL) as well as approaches based on WS-Policy. These efforts enable us to describe quality metrics of services, such as response time, and the associated service level objectives flexibly and in a way that is meaningful for the business needs of a service client.

However, it is non-trivial to derive what it takes to provide the quality of service that is offered or that has been agreed upon. In many cases, we rely on experience to decide, for example, the size of a cluster for a particular workload. In addition, managing a service at different QoS levels on the same infrastructure is not easy. Although a number of performance management technologies have been developed, such as workload managers and network dispatchers to control response times of individual systems and clusters and various availability management approaches, it is not straightforward to configure, for example, workload managers to satisfy response time goals for a set of different SLAs. All of those issues are challenging for all distributed systems.

In the field of service-oriented computing and Web services, loosely coupled distributed systems, we face a number of additional issues:

- Clients want to define QoS parameters from their perspective. While this is not so much an issue in traditional distributed systems, where the network and other system environment components are under control of the same organization, it has to be taken into account by providers and users of Web services.
- Services are often to be included in composite Web services defined using, for example, BPEL. In this case, we have to understand how the individual QoS properties of one element of a composite service contribute to the overall QoS. This is particularly interesting in the case of stochastic QoS: In 90% of the cases, the average response time will be less than 2 seconds.
- Finally, composite services can be offered as web services, thereby creating recursive service relationships. Stochastic QoS properties aggregated in multiple steps of aggregation get very broad and meaningless very quickly with an increasing number of aggregation steps. How can we find meaningful QoS properties for those aggregates and what are the limits of aggregation from a QoS point of view?

The keynote will discuss some of these issues, look at how ongoing efforts in academia and industry contribute to solving them and point at a list of open future research topics.