Runtime Prediction of Queued Behaviour

Nurzhan Duzbayev and Iman Poernomo

King's College London
Strand, London, UK, WC2R2LS
{nurzhan.duzbayev, iman.poernomo}@kcl.ac.uk

Abstract. Service-based software architectures are often modeled with queues and queuing networks. Such models are useful for performance evaluation and design. They can also assist in runtime maintenance and administration, but, in this context, it is often far more valuable to be able to forecast how QoS characteristics are likely to evolve in the near future. This is particularly important in cases where systems can be adapted to counter QoS constraint violations: in such systems, given predictions of likely future QoS characteristics, pre-emptive adaptation strategies can be implemented.

This paper outlines an approach to runtime prediction of QoS characteristics of queued systems. Predictions are computed by applying ARIMA forecasting techniques to basic properties of a queued model, and then using the model to predict complex QoS characteristics. We outline how our methods integrate into our implementation framework for monitoring and pre-emptive adaptation of web service based systems.

1 Introduction

Many service-based software architectures can be modeled with queues and queuing networks. Such models can are useful for the performance evaluation and design of a good software architecture. They can also assist in effective maintenance of an implemented system during runtime, even if original, model-based predictions of relevant characteristics are not met. For example, during the design of a system, we can use a queued model to compute the average number of requests in a queue by estimating the average rate of requests serviced and the average number of incoming requests per time unit. Then, when the system is implemented, if either of these rates deviates significantly from their estimation, we can still use the queuing theory calculation to determine the actual average number of requests in the queue. In this way, it is possible to calculate actual values of important quality of service characteristics at runtime by application of the queuing model.

However, for the purposes of maintenance and system administration, it is often far more valuable to be able to forecast how QoS characteristics are likely to evolve in the near future. This is particularly important for systems that can be reconfigured and adapted to counter QoS constraint violations. In such systems, given predictions of likely future QoS characteristics, pre-emptive adaptation strategies can be implemented.

C. Hofmeister et al. (Eds.): QoSA 2006, LNCS 4214, pp. 78–94, 2006.

For example, consider a client web service using one of two functionally equivalent queuing server web services. By utilizing Universal Description, Discovery and Integration (UDDI) at runtime [16], it is possible to redirect the client's requests from one server to another. For the case of a single request, it would be preferable to redirect calls to the server that has the shortest queue length. In the case where a large number of calls make up a transaction that must be sent to the same server, then it may be preferable to redirect calls to the web service with the shortest *average* queue length (and so, the shortest average time for serving the transaction). The same situation holds in the case where an expensive UDDI lookup is required to search for equivalent web services: redirection should be done infrequently and the best overall server should be chosen. It is possible that one server might have the shortest current or average queue length, but, in a few minutes, the server will possess the longest queue, due to a steep increase in popularity. Performance of such an architecture could be further improved if adaptation was not based on current or average queue length, but on a *predicted* future queue length.

This paper outlines an approach to the prediction of QoS characteristics of queued systems. Predictions are computed from applying ARIMA forecasting techniques to basic properties of a queued model, and then using the model to predict complex QoS characteristics. Predictions are made with a confidence interval expressing the error associated with the measurement and prediction processes. We outline how our methods integrate into the MPA system for monitoring and pre-emptive adaptation of web service based systems, currently being developed by the Predictable Assembly Laboratory at King's.[1]

The paper proceeds as follows:

- Section 2 summarizes relevant notions from queuing theory.
- Our prediction and error analysis techniques are presented in section 3.
- An illustrate example is provided in section 4.
- Our implementation, the MPA system, is described in section 5, focusing on how prediction relates to monitoring and adaptation of web services using the Microsoft Windows Management Instrumentation framework and UDDI.
- Conclusions and related work are discussed in the final section.

2 Queued Communication Models

Queuing theory enables the mathematical analysis of queued communication between clients and a server (or set of servers) (see, for example, [7,17]). Such communication is commonplace in large-scale distributed systems, where the use of loosely coupled messaging permits messages to be sent asynchronously from multiple sources to the same component at the same time. Performance evaluation of such systems is essential, particularly when dealing with systems assembled through web services, as HTTP-based SOAP communication is susceptible to rapid performance deterioration.

[1] `http://palab.dcs.kcl.ac.uk`