Automated Parameterization of Performance Models from Measurements

Giuliano Casale
Imperial College London
London, UK
g.casale@imperial.ac.uk

Simon Spinner
University of Würzburg
Würzburg, Germany
simon.spinner@uni-wuerzburg.de

Weikun Wang
Imperial College London
London, UK
weikun.wang11@imperial.ac.uk

ABSTRACT

Estimating parameters of performance models from empirical measurements is a critical task, which often has a major influence on the predictive accuracy of a model. This tutorial presents the problem of parameter estimation in queueing systems and queueing networks. The focus is on reliable estimation of the *arrival rates* of the requests and of the *service demands* they place at the servers. The tutorial covers common estimation techniques such as regression methods, maximum-likelihood estimation, and moment-matching, discussing their sensitivity with respect to data and model characteristics. The tutorial also demonstrates the automated estimation of model parameters using new open source tools.

Keywords

Demand Estimation; Arrival Processes

1. INTRODUCTION

The emergence of cloud computing and DevOps poses an increasing demand for tools to automatically instantiate performance models and to reduce the effort of practitioners in using them. However, it is challenging to translate performance measurements into concrete parameters of performance models. We here focus on queueing systems and networks, where the critical input parameters are service demands, which represent the cumulative amount of time a request is processed at a server before completion, and the arrival rates of requests to a queue from the external world.

Service demands, in particular, are hard to obtain as these are not explicitly tracked by log files, and deep monitoring instrumentation typically poses unacceptably large overheads, especially at high resolutions. Considering that real-world application requests can complete in a few milliseconds, individual monitoring may become too expensive to perform in a production system. Inference methods can help tackling this problem by extracting demands from partial measurements. To solve this problem, maximum likelihood based approaches has been developed for demand estimation problem. Maximum likelihood estimation is an optimization method that aims at finding the value of the parameters of a probabilistic model, such that the likelihood of obtaining a given set of samples is maximal. Response time and queue length data is required for this procedure.

Compared to service demands, arrival rates of requests to servers are often simple to measure, but it is difficult to estimate a representative model that captures the statistical characteristics of the arrivals, such as their time-varying patterns. Recent research has increasingly explored the fitting of arrival processes using Markovian Arrival Processes (MAPs) and Marked MAPs, which can be used in conjunction with matrix-analytic methods to predict queueing performance at servers.

In this tutorial, we overview the literature on the fitting of arrival rates and the estimation of service demands for queueing model parameterization. The tutorial covers estimation techniques and high-level paradigms for demand estimation (e.g., methods based on utilization, response-time, or queue-length) together with their sensitivity with respect to data and model characteristics. The tutorial also demonstrates the automated estimation of model parameters using open source tools, such as LibRede and FG for service demands, and M3A for arrival rates.

2. PARAMETER ESTIMATION

2.1 Service Demand Estimation

In order to quantify service demands, a dynamic analysis of the system of interest is required. Service demands are difficult to measure directly with state-of-the-art monitoring tools. Modern operating systems can only provide resource usage statistics on a per-process level. However, the mapping between operating system processes and application requests is non-trivial, because many applications serve different requests with one or more operating system processes (e.g., HTTP web servers).

The advantage of service demand estimation compared to direct measurement techniques is their general applicability and low overheads. These estimation approaches rely on coarse-grained measurements from the system (e.g., CPU utilization, and end-to-end response times), which can be easily and cheaply monitored with state-of-the-art tools without the need for fine-grained code instrumentation. These measurements are routinely collected for many applications,

ICPE'16, March 12-18, 2016, Delft, Netherlands
© 2016 ACM. ISBN 978-1-4503-4080-9/16/03…$15.00
DOI: http://dx.doi.org/10.1145/2851553.2851565

which makes the service demand methods applicable on systems serving production workloads. Over the years, a number of approaches to resource demand estimation have been proposed using different statistical estimation techniques (e.g., linear regression, Kalman filter, etc.) and based on different laws from queueing theory. We have surveyed and evaluated different approaches to resource demand estimation in our previous work [2]. Our evaluation shows, that one has to consider different characteristics of the estimation approach, such as the expected input parameters, its accuracy and its robustness to measurement anomalies when selecting an appropriate approach.

2.2 Arrival Process Fitting

Arrival processes of requests to queues introduce unique challenges in the estimation. Compared to service demands, which are infeasible or impractical to measure directly, inter-arrival times of requests can often be collected, either at the granularity of individual requests or in terms of aggregated arrival counts in a time period. Very often, the arrival patterns of requests exhibit temporal dependence, periodicities and non-exponential distributions that require special modelling techniques for their fitting. Markovian arrival processes (MAPs) fit an arrival process into a continuous-time Markov chain (CTMC) with marked and unmarked transitions. Upon activation of a marked transition in the chain, an arrival event occurs; conversely an unmarked transition has no associated arrival. The fitting problem is to determine the number of states, and the transition rates of marked and unmarked transitions so that the behavior of the MAP resembles as closely as possible the empirical trace. Upong addressing this problem, matrix-geometric techniques accept MAPs as descriptions of arrival processes to queueing systems and return predictions of response times and other performance metrics of interests. MAPs include as special cases renewal processes with hypo-exponential and hyper-exponential inter-arrival times and ON/OFF processes with exponential holding times, which are important to describe non-Poissonian traces. Recently, we have investigated Marked MAPs (MMAPs), which further extend the MAP model to capture a trace of multi-class arrivals [1]. Similarly to MAPs, matrix-geometric techniques enable the use of MMAPs to describe multi-class arrival processes in queueing systems. MMAPs introduce novel fitting concepts such as backward and foward moments.

3. TOOLS

The parameterization of queueing models in terms of service demands and arrival processes requires software tools to automate the process. The tutorial contemplates two tools for service demands estimation, LibReDE [4] and FG [4], and a tool for arrival process fitting, called M3A, which implements the methods in [1, 5].

3.1 LibReDE

LibReDE is a library of ready-to-use implementations of state-of-the-art approaches for resource demand estimation that can be used for online and offline analysis [3]. It is the first publicly available tool for this task and aims at supporting performance engineers during performance model construction. Currently it supports different estimation approaches based on approximation, linear regression, Kalman filter and non-linear constrained optimization techniques. LibReDE automatically checks the preconditions of different estimation approaches according to the input measurements. One or several approaches are then executed and the accuracy of the results is evaluated using cross-validation. Thus it is possible to dynamically choose the estimation approach that provides the best results for a given set of measurement data. Furthermore, LibReDE also supports the automatic derivation of a workload description (i.e., resources and workload classes) from an architecture-level performance model (e.g., Descartes Modeling Language (DML)).

3.2 Filling-the-Gap

Filling-the-Gap (FG) is a tool for continuous performance model parametrization to support Quality-of-Service (QoS) analysis [4]. It implements a set of statistical estimation algorithms to parameterize performance models from runtime monitoring data. Multiple algorithms are included, allowing for alternative ways to obtain estimates for different metrics, but with an emphasis on service demand estimation. FG tool supports advanced algorithms to estimate parameters based on response times and queue-length data, which makes the tool useful in particular for applications running in virtualized environments where utilization readings are not always available. Besides, FG also provides feedback to users by generating reports on the application performance.

3.3 M3A

M3A is a set of MATLAB functions designed for computing the statistical descriptors of MMAPs and fitting marked traces with MMAPs [1]. M3A implements a novel method to match inter-arrival time moments in a multi-class trace using a class of acyclic processes. The fitting methodology is based on moment matching and allows the fitting of traces with geometrically decaying autocorrelation functions.

Acknowledgment

Giuliano Casale is partially supported by the UK Engineering and Physical Sciences Research Council grant EP/M009211/1 (OptiMAM). Simon Spinner is partially supported by the German Research Foundation (DFG) under grant No. KO 3445/11-1. Weikun Wang is supported by the Horizon 2020 project DICE (grant agreement no. 644869).

4. REFERENCES

[1] Andrea Sansottera, Giuliano Casale, and Paolo Cremonesi. Fitting second-order acyclic marked Markovian Arrival Processes. In *Proc. of DSN*, pages 1–12. IEEE, 2013.

[2] Simon Spinner, Giuliano Casale, Fabian Brosig, and Samuel Kounev. Evaluating Approaches to Resource Demand Estimation. *Perf. Eval.*, 92:51 – 71, October 2015.

[3] Simon Spinner, Giuliano Casale, Xiaoyun Zhu, and Samuel Kounev. LibReDe: A library for resource demand estimation. In *Proc. of ICPE*, pages 227–228. ACM, 2014.

[4] Weikun Wang, Juan F Pérez, and Giuliano Casale. Filling the gap: a tool to automate parameter estimation for software performance models. In *Prod. of QUDOS*, pages 31–32. ACM, 2015.

[5] Giuliano Casale, Andrea Sansottera, and Paolo Cremonesi. Compact Markov-Modulated Models for Multiclass Trace Fitting. *Under submission.*