Performance Evaluation 92 (2015) 51–71

Contents lists available at ScienceDirect

Performance Evaluation

journal homepage: www.elsevier.com/locate/peva

Evaluating approaches to resource demand estimation

Simon Spinner [a,*], Giuliano Casale [b], Fabian Brosig [a], Samuel Kounev [a]

[a] University of Würzburg, Am Hubland, Würzburg, Germany
[b] Imperial College London, Department of Computing, SW7 2AZ, UK

ARTICLE INFO

Article history:
Received 27 August 2014
Received in revised form 2 March 2015
Accepted 16 July 2015
Available online 26 July 2015

Keywords:
Resource demand estimation
Workload characterization
Quantitative performance analysis
Performance modeling

ABSTRACT

Resource demands are a key parameter of stochastic performance models that needs to be determined when performing a quantitative performance analysis of a system. However, the direct measurement of resource demands is not feasible in most realistic systems. Therefore, statistical approaches that estimate resource demands based on coarse-grained monitoring data (e.g., CPU utilization, and response times) have been proposed in the literature. These approaches have different assumptions and characteristics that need to be considered when estimating resource demands. This paper surveys the state-of-the-art in resource demand estimation and proposes a classification scheme for estimation approaches. Furthermore, it contains an experimental evaluation comparing the impact of different factors (monitoring window size, number of workload classes, load level, collinearity, and model mismatch) on the estimation accuracy of seven different approaches. The classification scheme and the experimental comparison helps performance engineers to select an approach to resource demand estimation that fulfills the requirements of a given analysis scenario.

1. Introduction

Performance models can be used to answer performance-related questions for a software system during system design, capacity planning and sizing, or system operation. There are different performance modeling formalisms, e.g. stochastic performance models (Queueing Networks (QN) [1], Queueing Petri Nets (QPN) [2]), or architecture-level performance models (e.g., Palladio Component Model (PCM) [3]. The performance models can be analyzed using analytic methods or simulation to predict the performance of a system. However, the creation of model instances for a given system can be a complex and time-consuming task. During model creation, various model parameters need to be quantified. This usually requires experimentation with the system under study to obtain the measurement data required for model parameterization. It is of paramount importance to find representative parameter values in order to ensure accurate and reliable performance predictions.

A key parameter of stochastic performance models are *resource demands* (a.k.a. service demands). A resource demand is the average time a unit of work (e.g., request or transaction) spends obtaining service from a resource (e.g., CPU or hard disk) in a system over all visits excluding any waiting times [4,5]. The resource demand for processing a request is influenced by different factors, for example, the application logic specifies the sequence of instructions to process a request, and the hardware platform determines how fast individual instructions are executed. The definition of a resource demand implies that the value of a resource demand is platform-specific (i.e., only valid for a specific combination of application, operating system, hardware platform, etc.).

* Corresponding author.
E-mail addresses: simon.spinner@uni-wuerzburg.de (S. Spinner), g.casale@imperial.ac.uk (G. Casale), fabian.brosig@uni-wuerzburg.de (F. Brosig), samuel.kounev@uni-wuerzburg.de (S. Kounev).

http://dx.doi.org/10.1016/j.peva.2015.07.005
0166-5316/

In order to quantify resource demands, a dynamic analysis of the system of interest is required. Resource 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. Many applications serve different requests with one or more operating system processes (e.g., HTTP web servers). Standard profiling tools for performance debugging [6,7] can be used to obtain execution times of individual application functions when processing an individual request. However, the resulting execution times are not broken down to the processing times at individual resources and profiling tools typically introduce high overheads significantly influencing the performance of a system. Furthermore, advanced instrumentation techniques have been proposed in the literature to measure resource demands on the operating system layer [8], or the application layer [9–11]. These techniques build upon specific capabilities of the underlying platform and are not generally applicable.

This survey focuses on statistical approaches to resource demand estimation. The advantage of resource 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 (e.g., in data centers). Therefore, approaches to resource demand estimation are also 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. When selecting an appropriate approach to resource demand estimation, one has to consider different characteristics of the estimation approach, such as the expected input parameters, its accuracy and its robustness to measurement anomalies. Depending on the constraints of the application context, only a subset of the estimation approaches may be applicable.

The target audience of this paper are performance engineers who want to apply resource demand estimation techniques to build a performance model of a system as well as researchers working on improved estimation approaches. This paper makes the following contributions: (i) a survey of the state-of-the-art in resource demand estimation, (ii) a classification scheme for approaches to resource demand estimation, and (iii) an experimental comparison of a subset of the estimation approaches.

The remainder of the paper is organized as follows. Section 2 summarizes the state-of-the-art and introduces the different approaches to resource demand estimation. Section 3 describes the classification scheme including a categorization of existing estimation approaches. Section 4 presents the experimental comparison of the estimation approaches and discusses the results. Section 5 concludes the paper.

2. Approaches to resource demand estimation

In this section, we survey the state-of-the-art in resource demand estimation and introduce the different approaches that have been proposed in the literature.

2.1. Methodology

In order to obtain the estimation approaches listed in Table 2, we started the literature search by reading the titles and abstract of articles in the proceedings of 12 established conferences and workshops in the performance engineering community in the last 10 years. Relevant articles were analyzed further regarding references to other articles on resource demand estimation. Based on the found articles found we compiled a list of keywords to use for a broader search in common scientific search engines (scholar.google.com, portal.acm.org and citeseerx.ist.psu.edu) The keywords used for search were *resource demand (estimation)*, including synonyms *service demand*, *service time*, *service requirement*. Furthermore, we also considered the more general terms *workload characterization*, *parameter estimation* and *model calibration*. The list of articles resulting from this search was then filtered based on the titles and abstracts. After filtering, we got the list of 37 papers on resource demand estimation shown in Table 2.

2.2. Notation and assumptions

In the following, we use a consistent notation for the description of the different approaches to resource demand estimation. We denote resources with the index $i = 1 \dots I$ and workload classes with the index $c = 1 \dots C$. The variables used in the description are listed in Table 1. We assume the *Flow Equilibrium Assumption* [12] to hold, i.e., that over a sufficiently long period of time the number of completions is approximately equal to the number of arrivals. As a result, the arrival rate λ_c is assumed to be equal to the throughput X_c. Furthermore, we use the term resource demand as a synonym for service demand and for simplicity of exposition we assume $V_{i,c} = 1$, i.e., no distinction is made between service demand and service time.

2.3. Description of approaches

In this section, we describe the different approaches to resource demand estimation that exist in the literature. Table 2 gives an overview of all approaches.