Performance Evaluation 73 (2014) 18–40

Contents lists available at ScienceDirect

Performance Evaluation

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

Indirect estimation of service demands in the presence of structural changes

Paolo Cremonesi, Andrea Sansottera *

Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133, Milan, Italy

ARTICLE INFO

Article history:
Available online 13 August 2013

Keywords:
Service demand
Estimation
Regression
Robust statistics

ABSTRACT

According to the utilization law, throughput and utilization are linearly related and their measurements can be used for the indirect estimation of service demands. In practice, however, hardware and software modifications as well as non-modeled loads due to periodic maintenance activities make the estimation process difficult and often impossible without manual intervention to analyze the data. Due to configuration changes, real world datasets show that workload and utilization measurements tend to group themselves into multiple linear clusters. To estimate the service demands of the underlying performance models, the different configurations have to be identified. In this paper, we present an algorithm that, exploiting the timestamps associated with each throughput and utilization observation, identifies the different configurations of the system and estimates the corresponding service demands. Our proposal is based on robust estimation and inference techniques and is therefore suitable to analyze contaminated datasets. Moreover, not only sudden and occasional changes of the system, but also recurring patterns in the system's behavior, due for instance to scheduled maintenance tasks, are detected. An efficient implementation of the algorithm has been made publicly available and, in this paper, its performance is assessed on synthetic as well as on experimental data.

1. Introduction

In order to satisfy quality of service guarantees, capacity planning must be based on accurate performance models. In this context, models based on queueing networks are widely adopted, since they allow us to predict the end-to-end delays experienced by the users under different workload intensities. Queueing network models, however, require a number of parameters. In particular, the service demand has to be provided for every combination of service station and workload class. Unfortunately, direct measurements of service demands are rarely available in real systems and obtaining them might require invasive techniques such as benchmarking, load testing, profiling, or source code instrumentation. The application of these techniques to the performance modeling of large scale data centers, with hundreds of systems and applications, is prohibitively expensive.

An alternative approach relies on the automatic estimation of service demands from other metrics that are more easily obtained, such as the number of transactions processed and the average utilization of the servers. In particular, according to the utilization law, the utilization of a load-independent service station is linearly related to the throughput of the different workload classes. The coefficients in this linear relationship are the service demands and, in principle, can be estimated using well-known regression techniques, as discussed in [1–3].

This approach relies on the assumption that the system configuration does not change during the observation period, as shown in Fig. 1(a). In practice, however, a number of outliers can affect the data and the system (either the service stations

* Corresponding author. Tel.: +39 02 2399 9677.
E-mail addresses: paolo.cremonesi@polimi.it (P. Cremonesi), sansottera@elet.polimi.it (A. Sansottera).

0166-5316/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.peva.2013.08.001

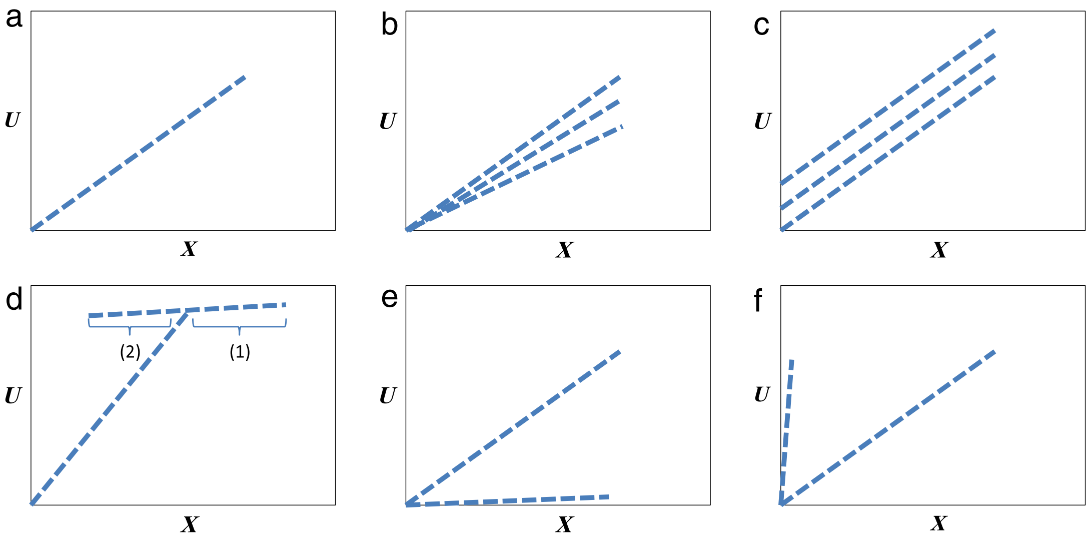

Fig. 1. Examples on how structural changes affect utilization.

or the workloads) might change during the observation period [4,5]:

- Software might be updated to a new version, reducing or increasing the service demands, as shown in Fig. 1(b). Moreover, high-end servers with hot-plugging and fault tolerance features might be upgraded or subject to maintenance with no disruption of the system's activities. For instance, processors and memory modules can be added or removed without the need to stop the system. These modifications to the hardware configuration might impact the service demands.
- Certain background tasks, such as scheduled backups, can introduce recurring workloads that are not measured. These "hidden" workloads modify the utilization which is periodically shifted toward higher values, as shown in Fig. 1(c).
- When the incoming request rate exceeds the maximum throughput of a system, some requests may be buffered for later processing in batch mode, as shown in Fig. 1(d). In particular, online requests arriving when the system is close to its saturation point—as in region (1)—are processed later, when the load on the system is lower—as in region (2). This behavior occurs when the throughput is measured as the number of arrivals over the observation period, which is not accurate when the system is close to the saturation point.
- Failures in the software used to monitor performance counters may lead to underestimate utilization— Fig. 1(e)—or throughput— Fig. 1(f).

Real examples are provided in Section 6. Changes in service demands are even more frequent in virtualized data centers, since virtual machine parameters can be easily modified affecting the share of resources available to an application (e.g., number of virtual cores), without the monitoring tools being aware of these modifications.

In the literature, there are two families of approaches used to estimate service demands when there are changes in the system properties.

- The *cluster-wise* regression methods completely neglect the timestamps and make no assumption on the ordering of the data. In fact, the goal of cluster-wise regression is to discover multiple linear models within a set of data points, estimating the parameters of the linear models and simultaneously determining the point-to-cluster membership [6].
- The *change-point* regression methods, given a set of ordered and linearly related samples, finds the points (or breaks) at which the parameters of the linear model change [7].

However, both cluster-wise and change-point models have difficulties in detecting periodically and short (i.e., spanning just one or a few observations) alterations of the model parameters, which we refer to as *recurring patterns*. Moreover, cluster-wise regression is NP-hard [8] and cannot be applied to large scale datasets.

In this work, we propose a novel approach to the estimation of service demands under structural changes, leveraging the timestamps directly in the clustering process to improve clustering quality and provide a time-based description of each cluster that can be used for diagnostic purposes or forecasting. The *Time-based Linear Clustering (TLC)* algorithm tackles service demand estimation as a combined change-point and pattern detection problem and, being based on robust estimation and inference techniques, is able to analyze heavily contaminated datasets. Unlike change-point regression models, our approach does not only identify sudden and occasional changes of the system, but also recurring patterns in the system's behavior (e.g., due to scheduled maintenance tasks). The algorithm works in two steps. In the first step, the observation period is segmented into multiple epochs, corresponding to different parameter values. In the second step, outliers are processed to identify recurring structural changes, which periodically manifest by altering one or a few observations.