The Journal of Systems and Software 111 (2016) 105–118

Contents lists available at ScienceDirect

The Journal of Systems and Software

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

Multi-criteria analysis of measures in benchmarking: Dependability benchmarking as a case study

Jesús Friginal [a,b,*], Miquel Martínez [c], David de Andrés [c], Juan-Carlos Ruiz [c]

[a] SCASSI CIBERSEGURIDAD, Capitán Haya 38-4, 28020 Madrid, Spain
[b] CNRS, LAAS, 7 avenue du colonel Roche, Toulouse 31400, France
[c] STF-ITACA Universitat Politècnica de València, Campus de Vera s/n, 46022, Spain

ARTICLE INFO

Article history:
Received 24 September 2014
Revised 30 August 2015
Accepted 31 August 2015
Available online 24 September 2015

Keywords:
Multiple-Criteria Decision Making (MCDM)
Dependability benchmarking
Quality models

ABSTRACT

Benchmarks enable the comparison of computer-based systems attending to a variable set of criteria, such as dependability, security, performance, cost and/or power consumption. It is not despite its difficulty, but rather its mathematical accuracy that multi-criteria analysis of results remains today a subjective process rarely addressed in an explicit way in existing benchmarks. It is thus not surprising that industrial benchmarks only rely on the use of a reduced set of easy-to-understand measures, specially when considering complex systems. This is a way to keep the process of result interpretation straightforward, unambiguous and accurate. However, it limits at the same time the richness and depth of the analysis process. As a result, the academia prefers to characterize complex systems with a wider set of measures. Marrying the requirements of industry and academia in a single proposal remains a challenge today. This paper addresses this question by reducing the uncertainty of the analysis process using quality (score-based) models. At measure definition time, these models make explicit (i) which are the requirements imposed to each type of measure, that may vary from one context of use to another, and (ii) which is the type, and intensity, of the relation between considered measures. At measure analysis time, they provide a consistent, straightforward and unambiguous method to interpret resulting measures. The methodology and its practical use are illustrated through three different case studies from the dependability benchmarking domain, a domain where various different criteria, including both performance and dependability, are typically considered during analysis of benchmark results.. Although the proposed approach is limited to dependability benchmarks in this document, its usefulness for any type of benchmark seems quite evident attending to the general formulation of the provided solution.

1. Introduction

Benchmarks are well-known tools to compare and select distributed systems mainly attending to their performance, cost and power consumption. Standardization bodies, such as the Transaction Processing Performance Council (TPC, 2013), currently propose a set of representative (since widely accepted by the community) benchmarks for distributed systems. In the last decade, some initiatives have addressed the challenging goal of including the evaluation of dependability and security properties in conventional benchmarks. Resulting benchmarks are typically called dependability benchmarks.

Like in conventional benchmarks, controllability and repeatability of experiments and interpretation of results are essential in dependability benchmarks (DBench, 2003; Almeida et al., 2010; Ceccarelli, 2012). To date, most of the efforts done in the community around this topic have been oriented towards providing controllability and repeatability of experiments. These efforts can be understood given the need to obtain the same (or at least statistically similar or comparable) experimental measures when the same experimental setup is considered.

However, and without taking importance away from this point, controllability and repeatability also affects other stages of the benchmarking process, such as the analysis of results. The reader should understand that dependability benchmarks introduce the need of performing a more complex analysis of target systems, considering their behavior in the presence of faults and attacks, and characterizing such behavior through a larger set of measures, including dependability and security specific ones. This evidence becomes a challenge when considering the evaluation of complex systems formed by a large and heterogeneous set of sub-systems and components. This is a challenge not only for the amount of measures to consider, but also for their variety of origin and typology.

* Corresponding author at: SCASSI CIBERSEGURIDAD, Capitán Haya 38-4, 28020, Madrid, Spain.
E-mail addresses: jesus.friginal@scassi.com (J. Friginal), mimarra2@disca.upv.es (M. Martínez), ddandres@disca.upv.es (D. de Andrés), jcruizg@disca.upv.es (J.-C. Ruiz).

http://dx.doi.org/10.1016/j.jss.2015.08.052
0164-1212/

To date, the analysis of results from dependability benchmarks has been an aspect strongly relying on the human factor. Evaluators subjectively interpret measures following considerations that are usually omitted in the finally generated reports. In consequence, repeating the same analysis of measures and obtaining the same conclusions, even when results are the same, becomes sometimes a complex task.

The underlying problem is that most proposals limit their purpose to the delivery of benchmark measures. In deed, the consideration of a representative set of measures has been traditionally enough to justify their selection for benchmarking purposes (Vieira and Madeira, 2003). Then, the analysis of such measures, and consequently the related comparison of alternatives, is typically considered outside the purpose of the specification of most benchmarks, including dependability benchmarks. This can be something acceptable in the context of conventional benchmarks but it is unaffordable in the case of dependability benchmarks, since any aspect leading to a wrong alternative selection may have a negative impact on the safety or security of the system, with the subsequent losses, of reputation, money or lives.

On the one hand, benchmark measures must be contextualized during the analysis process. Without contextualizing their meaning throughout factors such as the environment, the type of system targeted, or the evaluation performer, same results may have different interpretations depending on the evaluation consumer's subjectivity. On the other hand, it must be clearly specified in the analysis process which are the relations considered among measures, and the intensity of such relations. Otherwise, it may be very difficult to guess which have been all the assumptions adopted by someone analyzing a set of benchmark measures. In other words, it may be difficult to verify the conclusions issued from the analysis of a set of benchmark measures.

It is worth mentioning that even if all this effort is done, the analysis and interpretation of results remains an error-prone process requiring a very deep dependability expertise, in the case of dependability benchmarks. This situation increases the *uncertainty* of evaluation analyses and thus negatively affects the credibility of the conclusions obtained. This ambiguous interpretation of concepts is commonly known as *semantic heterogeneity* (Anaby-Tavor et al., 2003).

This challenge could be addressed through a process of *semantic reconciliation* (Anaby-Tavor et al., 2003). Such process involves covering the existing gap between the explicit result of the evaluation, that is, the conclusions distilled from the analysis of measures, and the implicit real intention of evaluators, which concerns the interpretation procedure to obtain such conclusions. This fact increases the sensitivity of analyses, potentially revealing surprising insights about the system under evaluation. This approach is specially useful when there is no obvious optimal (or unanimous) solution due to the large number of criteria that need to be taken into account, or when decisions often require the fulfillment of conflicting objectives (e.g., design or choice of systems maximizing their dependability and performance). It has also the potential for improving the work of system evaluators by leading them to unequivocal and more objective conclusions. Unfortunately, to date, *semantic reconciliation* remains a non-addressed issue in the domain of distributed systems dependability benchmarking.

Contributions of this research are two-fold. First, providing a multi-criteria analysis methodology to ease the multiple interpretations that the measures issued from benchmarks may have depending on the criteria followed by evaluators. The goal is to make explicit the subjective interpretation rules that evaluators typically apply implicitly when determining to what extent measures satisfy evaluation requirements. Doing this in a systematic and repeatable way is essential when different evaluators need to make a fair comparison of their results. This is why the proposal relies on a set of rigorous mathematical basis enabling the quantification of the uncertainty underlying analysis conclusions. Second, defining a suitable methodology to align the two opposing viewpoints (i) the viewpoint of those evaluators that prefer having all the possible measures as field data for enabling deep result analysis and promote data sharing among community members (Kanoun et al., 2005) (e.g., people from academia), and (ii) the point of view of those others adopting a more pragmatical viewpoint that ask for an small set of meaningful and representative scores to enable the fastest possible characterization, comparison and ranking of evaluated systems (European New Car Assessment Programme (EuroNCAP), 2013) (e.g., people from industry). To cope with this goal the approach rely on the notion of quality model, adopted from ISO/IEC 25000 standards (International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC), 2010), to formulate not only rigorous but also usable and flexible interpretation rules.

Before closing this introduction, it is important to say that the integration of a multi-criteria analysis methodology in very simple benchmarks may be useless, specially where few, or only one, measure or measure type is under consideration. The use of the methodology proposed in this paper makes sense in benchmarking contexts where the analysis process asks for the simultaneous consideration (aggregation and/or comparison) of different measures of different type. The higher the number of measures or the heterogeneity of such measures the higher the usefulness of the proposal. Since this is what happens in dependability benchmarks, for the sake of exemplification, the present proposal limits its purpose to this type of benchmarks, and this despite its obvious potential for any other type of benchmarks.

The rest of the paper is structured as follows. Section 2 introduces a brief background about dependability benchmarking and multi-criteria analysis. Section 3 presents our multi-criteria analysis methodology. Section 4 shows the feasibility of our approach through three different case studies and finally, Section 5 concludes the paper.

2. Background

Computer benchmarks are standard tools that enable the evaluation and comparison of different systems, components, and tools according to specific characteristics (Gray, 1992). Benchmarks have been widely used to compare the performance of systems, e.g. transactional systems (TPC, 2013) or embedded systems (EEMBC, 2014). From a high-level viewpoint, the specification of a conventional benchmark encompasses with the definition of the following components:

- The *system under benchmarking* and the *benchmark target*, which specify the context of use of the system under evaluation and the model of the considered target;
- The *measures* that will be employed to characterize and compare existing alternatives;
- The *execution profile* required to parameterize and exercise both the system under benchmarking and the benchmark target during experimentation. This is typically a *workload*;
- The *experimental procedure* specifying how to run the selected execution profile on the considered target and how to trace the resulting activity;
- The process to follow in order to transform resulting traces (experimental measurements) into expected benchmark measures.

The main benefit of conventional benchmarks is that, once the set of proposed measures are widely accepted by a community, systems produced by such community can be compared in a quite straightforward and unambiguous way. The key issue here is that most of the considered measures are homogeneous. In deed, this type of benchmarks simply characterize evaluated systems in terms of either their performance or their cost. As a result, comparisons among systems