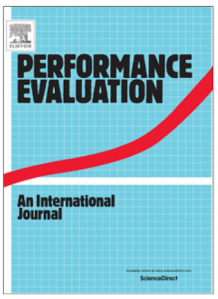
# Accepted Manuscript




Characterizing machines lifecycle in Google data centers

Stefano Sebastio, Kishor S. Trivedi, Javier Alonso

PII: S0166-5316(18)30004-X
DOI: https://doi.org/10.1016/j.peva.2018.08.001
Reference: PEVA 1965

To appear in: *Performance Evaluation*

Received date : 5 January 2018
Revised date : 7 June 2018
Accepted date : 19 August 2018

Please cite this article as:, Characterizing machines lifecycle in Google data centers, *Performance Evaluation* (2018), https://doi.org/10.1016/j.peva.2018.08.001

# Characterizing Machines Lifecycle in Google Data Centers

Stefano Sebastio[a,*], Kishor S. Trivedi[b], Javier Alonso[c]

[a]*London Institute for Mathematical Sciences (LIMS), 22 South Audley St. Mayfair, London, W1K 2NY, UK*
[b]*Department of Electrical and Computer Engineering, Duke University, Durham, 27708, USA*
[c]*Amazon.com, Seattle, USA*

**Abstract**

Due to the increasing need for computational power, the market has shifted towards big centralized data centers. Understanding the nature of the dynamics of these data centers from machine and job/task perspective is critical to design efficient data center management policies like optimal resource/power utilization, capacity planning and optimal (reactive and proactive) maintenance scheduling. Whereas jobs/tasks dynamics have received a lot of attention, the study of the dynamics of the underlying machines supporting the jobs/tasks execution has received much less attention, even when these dynamics would substantially affect the performance of the jobs/tasks execution. Given the limited data available from large computing installations, only a few previous studies have inspected data centers and only concerning failures and their root causes. In this paper, we study the 2011 Google data center traces from the machine dynamics perspective. First, we characterize the machine events and their underlying distributions in order to have a better understanding of the entire machine lifecycle. Second, we propose a data-driven model to enable the estimate of the expected number of available machines at any instant of time. The model is parameterized and validated using the empirical data collected by Google during a one month period.

*Keywords:* statistical analysis, distributed architectures, cloud computing, system reliability, large-scale systems, empirical studies

## 1. Introduction

The growing demand for storage and computational resources led the service providers to adopt cloud computing technologies. Cloud-based services are executed in large and centralized data centers spread around the world offering high quality performance and availability to the end users. In turn, due to the increasing popularity and complexity of cloud-based services, the resource demand in the data centers is increasing as well, and thus, new resources (i.e., machines/servers) are continually being added. Larger data centers executing heterogeneous workloads represent challenges in terms of management in order to maximize the data center efficiency. Moreover, the pressure to maximize the data center utilization while reducing its cost are posing management challenges like resource/power utilization optimization, capacity planning or optimal maintenance scheduling.

Understanding the dynamics of the data centers from the machine and job/task perspectives is therefore critical to design more efficient data centers and more effective data center management policies aimed at facing the above mentioned challenges. Nonetheless, difficulties in collecting or finding publicly available log artifacts have affected and limited potential studies and improvements on data centers [1, 2, 3]. Traditionally, it is possible to discern four machine analysis approaches [4]: (i) measurements of logs collected from a deployed system; (ii) test systems undergoing stress tests; (iii) trace logs on system failures; (iv) analytical models.

*Measuring and collecting logs from a production system* [5, 6] allows one to analyze the actual behavior of the system that, depending on the deployed environment (e.g., in terms of imposed workload), reflects all the characteristics and phenomena of a real environment. The drawback to this approach is that measurement insights can be specific

*Corresponding author
*Email address:* stefano.sebastio@alumni.imtlucca.it (Stefano Sebastio)