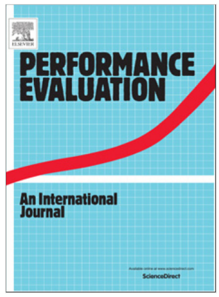
Accepted Manuscript



Transient provisioning and performance evaluation for cloud computing platforms: A capacity value approach

Brendan Patch, Thomas Taimre

PII: S0166-5316(17)30175-X
DOI: https://doi.org/10.1016/j.peva.2017.10.001
Reference: PEVA 1937

To appear in: *Performance Evaluation*

Received date : 10 May 2017
Revised date : 12 October 2017
Accepted date : 19 October 2017

Transient Provisioning and Performance Evaluation for Cloud Computing Platforms: A Capacity Value Approach

Brendan Patch*†‡ and Thomas Taimre†

October 12, 2017

Abstract

User demand on the computational resources of cloud computing platforms varies over time. These variations in demand can be predictable or unpredictable, resulting in 'bursty' fluctuations in demand. Furthermore, demand can arrive in batches, and users whose demands are not met can be impatient. We demonstrate how to compute the expected revenue loss over a finite time horizon in the presence of all these model characteristics through the use of matrix analytic methods. We then illustrate how to use this knowledge to make frequent short term provisioning decisions — transient provisioning. It is seen that taking each of the characteristics of fluctuating user demand (predictable, unpredictable, batchy) into account can result in a substantial reduction of losses. Moreover, our transient provisioning framework allows for a wide variety of system behaviors to be modeled and gives simple expressions for expected revenue loss which are straightforward to evaluate numerically.

Keywords: cloud computing, performance analysis, transient analysis, queueing theory, capacity value, matrix analytic methods.

1 Introduction

Highly complex systems are becoming an integral contributor to the productivity of many industries. The introduction of these systems is accompanied by an increase in the demand for computational resources. Distributed cloud computing platforms have emerged as the leading method for provision of these resources to end users. In addition to the environments available online (e.g. Amazon EC2, Microsoft Azure, Google AppEngine, GoGrid), many private organizations and universities now have computer clusters that allow users to distribute computing tasks across many nodes. This substantially reduces the need for each user to have expensive individually held computing resources that become idle when not needed or which are impractical to store at the users' geographical location.

Distributed computing constitutes a substantial portion of the energy consumption in modern computer and communication networks [28]. As such, well designed provisioning policies, which match the availability of resources with the demand for resources remains an active area of research [28]. An obvious avenue to reducing the energy use of a distributed cloud platform is to switch compute nodes off, or place them into a power saving mode, when they are not needed. For example, in [34] it is estimated that perfectly provisioning capacity to match demand in a production compute cluster at Google would result in a 17–22% reduction in energy use.

*Corresponding author. Email: b.patch@uq.edu.au
†The School of Mathematics and Physics, The University of Queensland, Qld 4169, Australia.
‡Korteweg-de Vries Institute for Mathematics, University of Amsterdam, 1090 GE Amsterdam, The Netherlands.