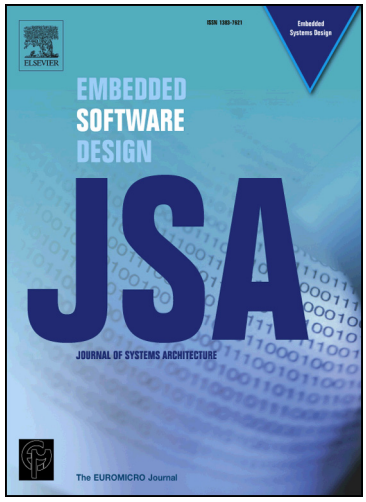

A Cloud Middleware for Assuring Performance and High Availability of Soft Real-time Applications

Kyoungho An, Shashank Shekhar, Faruk Caglar, Aniruddha Gokhale, Shivakumar Sastry

PII: S1383-7621(14)00025-3
DOI: http://dx.doi.org/10.1016/j.sysarc.2014.01.009
Reference: SYSARC 1227

To appear in: *Journal of Systems Architecture*

Please cite this article as: K. An, S. Shekhar, F. Caglar, A. Gokhale, S. Sastry, A Cloud Middleware for Assuring Performance and High Availability of Soft Real-time Applications, *Journal of Systems Architecture* (2014), doi: http://dx.doi.org/10.1016/j.sysarc.2014.01.009

A Cloud Middleware for Assuring Performance and High Availability of Soft Real-time Applications☆

Kyoungho An, Shashank Shekhar, Faruk Caglar, Aniruddha Gokhale[a], Shivakumar Sastry[b]

[a] *Institute for Software Integrated Systems (ISIS)*
Department of Electrical Engineering and Computer Science
Vanderbilt University, Nashville, TN 37235, USA
Email: {kyoungho.an, shashank.shekhar, faruk.caglar, a.gokhale}@vanderbilt.edu
[b] *Complex Engineered Systems Lab*
Department of Electrical and Computer Engineering
The University of Akron, Akron, OH 44325, USA
Email: ssastry@uakron.edu

Abstract

Applications are increasingly being deployed in the cloud due to benefits stemming from economy of scale, scalability, flexibility and utility-based pricing model. Although most cloud-based applications have hitherto been enterprise-style, there is an emerging need for hosting real-time streaming applications in the cloud that demand both high availability and low latency. Contemporary cloud computing research has seldom focused on solutions that provide both high availability and real-time assurance to these applications in a way that also optimizes resource consumption in data centers, which is a key consideration for cloud providers. This paper makes three contributions to address this dual challenge. First, it describes an architecture for a fault-tolerant framework that can be used to automatically deploy replicas of virtual machines in data centers in a way that optimizes resources while assuring availability and responsiveness. Second, it describes the design of a pluggable framework within the fault-tolerant architecture that enables plugging in different placement algorithms for VM replica deployment. Third, it

☆This work was supported in part by NSF awards CAREER/CNS 0845789 and SHF/CNS 0915976. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.