Information Sciences 319 (2015) 152–170

Contents lists available at ScienceDirect

Information Sciences

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

Learning a goal-oriented model for energy efficient adaptive applications in data centers

Monica Vitali [a,*], Barbara Pernici [a], Una-May O'Reilly [b]

[a] Politecnico di Milano, piazza Leonardo da Vinci 32, 20133 Milan, Italy
[b] Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139-4307, United States

ARTICLE INFO

Article history:
Received 30 June 2014
Received in revised form 13 January 2015
Accepted 28 January 2015
Available online 7 February 2015

Keywords:
Energy efficiency
Data center
Goal-oriented adaptation

ABSTRACT

This work has been motivated by the growing demand of energy coming from the IT sector. We propose a goal-oriented approach where the state of the system is assessed using a set of indicators. These indicators are evaluated against thresholds that are used as goals of our system. We propose a self-adaptive context-aware framework, where we learn both the relations existing between the indicators and the effect of the available actions over the indicators state. The system is also able to respond to changes in the environment, keeping these relations updated to the current situation. Results have shown that the proposed methodology is able to create a network of relations between indicators and to propose an effective set of repair actions to contrast suboptimal states of the data center. The proposed framework is an important tool for assisting the system administrator in the management of a data center oriented towards Energy Efficiency (EE), showing him the connections occurring between the sometimes contrasting goals of the system and suggesting the most likely successful repair action(s) to improve the system state, both in terms of EE and QoS.

1. Introduction

The general interest about environmental issues is continuously growing. One of the main contributors to pollution and energy consumption is Information Technology (IT). The impact of technology and its growing rate is evident at the everyday scale, and even more significant at a larger scale.

This work has been motivated by this recent attention towards Green IT and Energy Efficiency (EE) [4,36,5] from different perspectives (usage of resources, applications, and networks), and by the growing demand of energy coming from IT. This demand is comparable with the one of the major countries and it is expected to keep growing in the next years [11]. In this work we focus on efficiency considering the data center level, with a special attention to the contribution of Service Oriented Architecture (SOS) to energy consumption and performance. In SOAs, applications hosted in the servers are services that can be represented by processes. Each service is decomposed in a set of activities connected in a workflow. Recently, virtualization and cloud computing have changed the way resources are handled in a data center, allowing a more dynamic redistribution of computational resources among the several applications hosted by the data center or the cloud. As a consequence, the waste of resources is (potentially) reduced.

* Corresponding author. Tel.: +39 0223993424.
E-mail addresses: monica.vitali@polimi.it (M. Vitali), barbara.pernici@polimi.it (B. Pernici), unamay@csail.mit.edu (U.-M. O'Reilly).

http://dx.doi.org/10.1016/j.ins.2015.01.023
0020-0255/

We face the problem of EE by proposing a comprehensive approach. The approach includes the definition of the goals for the organization, the selection of metrics for the assessment of the system state, and the definition of techniques for the improvement of the system. In order to do so, we propose the use of a goal-oriented approach.

The context in which we are going to implement our goal-oriented approach for EE is a data center composed of several servers, each one with different features in terms of amount and kind of components. We focus on the application perspective for improving efficiency. For this reason we are not considering issues related to cooling or other facilities, but we look in details at applications and their IT infrastructure. We consider that the data center uses virtualization for hosting its applications and we consider that each single activity runs on a dedicated Virtual Machine (VM). A monitoring system is available for monitoring the system behavior with sensors that can work at several levels: data center level, server level, Virtual Machine level, and application level. Different kinds of monitoring systems can be employed (e.g. Zabbix[1] and Nagios[2]), as they do not have any influence on the approach.

The purpose of this work is to design a model which allows the described environment to improve its behavior in terms of EE while respecting constraints related to the Quality of Service (QoS) required by the users. The tradeoff between these two perspectives is a challenging topic nowadays and has been faced in many research papers [13,17,19]. However, most of the approaches underestimate the complexity of the environment they describe. A data center, especially modern data centers employing techniques such as virtualization and consolidation, is a dynamic environment where the number of applications and their configuration and deployment are continuously and rapidly changing. Also, external factors can impact on the performance and efficiency of the deployed applications, due to servers overload or to service unavailability. In such an environment, adaptation techniques requiring human intervention can be applied but at the cost of a lack of reactivity in the system. An automatic approach is needed to maximize the performance of the data center and to optimize its efficiency in terms of energy consumption by applying a set of adaptation techniques. In the literature, several techniques for adaptation are described and their effectiveness is proved. Starting from this knowledge we propose a methodology for modeling adaptation in data centers while automatically reacting to its dynamic modifications.

The main contributions of this work can be summarized as follows:

- **Definition and learning of a goal-oriented model for efficiency management.** An adaptation approach is proposed to learn a goal-oriented model for efficiency management. The aim of the model is to provide an adaptive environment able to automatically react to suboptimal situations by enacting adaptation strategies that impact directly or indirectly over efficiency and quality. The model is not derived from the knowledge of an expert, but it is build starting from available data using machine learning techniques. It is also able to automatically respond to modifications in a dynamic environment.
- **Representation and learning of indicators relations.** We express the relations occurring among the various metrics used for monitoring the system state. These relations are very important for predicting the future behavior of the system and for conducting what-if analysis.
- **Adaptive strategy selection.** Effects of actions over EE and QoS are automatically learned providing an adaptive and dynamic environment.

In order to realize the described model, we used a goal-driven approach where two layers are defined: a goal layer and a treatment layer. Goals correspond to the desired values for a set of metrics collected through the monitoring system, such as the energy consumption of the data center or of the physical machine and the QoS provided measured in terms of response time or throughput. Relations among metrics are represented using a Bayesian Network (BN) and automatically learned through the employment of machine learning techniques, such as a modification of the Max–Min Hill Climbing Algorithm (MMHC) [32] for learning links and the Maximum a Priori Estimation (MAP) for learning parameters [30]. Results have shown that the proposed methodology is able to learn the whole model from a set of monitored data. Given the model, it is possible to decide which is the best repair action in order to bring the system in a new configuration as near as possible to the optimal one, given a specific context for the system. Repair strategies are contained in the treatment layer and examples are migration, server switch off/on, and VM resources reconfiguration, or combinations of them. The effect of actions over the system are learned using the Multi Armed Bandit Selection (MABS) paradigm [7] and properly adapting the Adaptive Operator Selection (AOS) algorithm [18]. Once learned, the model can be used to prescribe the best adaptation action given a context.

The rest of this work is organized as follows. In Section 2 we analyze other approaches using adaptive models to manage EE and/or QoS of a data center. Then we propose our approach in Section 3. In Section 4 and Section 5 we analyze in details the two layers of the proposed model and show how it can be used to improve the system state in Section 6. Finally we discuss some experimental results (Section 7) and future work (Section 8).

[1] http://www.zabbix.com/.
[2] http://www.nagios.org/.