Future Generation Computer Systems 67 (2017) 94–108

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

Profile-based application assignment for greener and more energy-efficient data centers

Meera Vasudevan [a], Yu-Chu Tian [a,b,*], Maolin Tang [a], Erhan Kozan [c]

[a] School of Electrical Engineering and Computer Science, Queensland University of Technology, GPO Box 2434, Brisbane QLD 4001, Australia
[b] College of Information Engineering, Taiyuan University of Technology, Taiyuan, Shanxi 030024, China
[c] School of Mathematical Sciences, Queensland University of Technology, GPO Box 2434, Brisbane QLD 4001, Australia

HIGHLIGHTS

- Integrating the concept of profiles into application assignment in data centers.
- Building various profiles from raw data of real-world data centers.
- Penalty-based optimization framework for profile-based application assignment.
- Penalty-based profile matching algorithm to solve the optimization problem.

ARTICLE INFO

Article history:
Received 15 September 2015
Received in revised form
28 May 2016
Accepted 29 June 2016
Available online 1 September 2016

Keywords:
Data center
Application assignment
Profile
Resource management
Energy efficiency

ABSTRACT

The cloud computing era has brought significant challenges in energy and operational costs of data centers. As a result, green initiatives with regard to energy-efficient management of data center infrastructure for cloud computing have become essential. Addressing a big class of widely deployed data centers with relatively consistent workload and applications, this paper presents a new profile-based application assignment approach for greener and more energy-efficient data centers. It builds realistic profiles from the raw data measured from data centers and then establishes a theoretical framework for profile-based application assignment. A penalty-based profile matching algorithm (PPMA) is further developed to obtain an assignment solution, which gives near-optimal allocations whilst satisfying energy-efficiency, resource utilization efficiency and application completion time constraints. Through experimental studies, the profiling approach is demonstrated to be feasible, scalable and energy-efficient when compared to the commonly used general and workload history based application management approaches.

1. Introduction

In today's economy, data centers and cloud computing are increasingly used everyday by the sky-rocketing number of Internet users. This is predictably escalating the energy and costs to power and maintain these systems at an alarming pace. Overall, data centers consume 1.1% to 1.5% of the world's total electricity consumption [1]. They are responsible for 14% of the Information and Communication Technology (ICT) carbon footprint according to the Smart2020 analysis [2]. More than 35% of the current data center operational expenses are accounted for by energy consumption. This figure is projected to double in a few years. According to a report by the Natural Resources Defence Council (NRDC), data centers consumed 91 billion kWh of electrical energy in 2013. This statistics is projected to increase by 53% by year 2020 [3].

With different purposes, various data centers contribute to the energy consumption and carbon footprint differently. Large-scale data centres are mainly used to host public clouds with dynamic workload. Typical hyper-scale large data centers are those from giant IT corporations like Microsoft, Google, Apple, Amazon, and Facebook. In comparison, medium- and small-scale data centers are typically run by business companies, universities and government agencies. They typically provide services via private clouds or clusters/grids with virtualized management. Therefore, they have relatively consistent workload. The NRDC

[*] Corresponding author at: School of Electrical Engineering and Computer Science, Queensland University of Technology, GPO Box 2434, Brisbane QLD 4001, Australia.
E-mail address: y.tian@qut.edu.au (Y.-C. Tian).

http://dx.doi.org/10.1016/j.future.2016.06.037
0167-739X/© 2016 Elsevier B.V. All rights reserved.

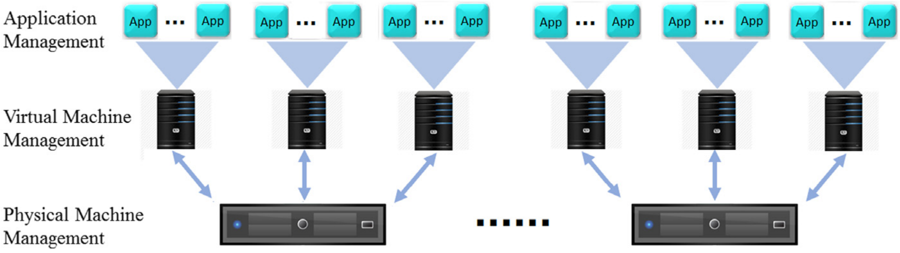

Fig. 1. Energy management architecture for data centers.

reports that there is a distinct gap in energy-efficient initiatives when comparing well-managed hyper-scale large data centers and the numerous less-efficient small- to medium-scale data centers. The hyper-scale large data centers only share 5% of the global data center energy usage, while the remaining 95% is made up of small- to medium-scale data centers [3]. Therefore, energy management for small- to medium-scale data centers with relatively consistent workload is globally more significant than that for hyper-scale large data centers with very dynamic workload. This paper targets the widely deployed small- to medium-scale data centers.

The necessity for green and energy-efficient measures to reduce carbon footprint and the exorbitant energy costs has become very real and emerging. Energy and cost distribution studies, e.g., Le et al. [4], have confirmed that deploying green initiatives at data centers reduces the carbon footprint by 35% at only a 3% cost increase. However, energy-aware measures with simultaneous maximum performance efficiency and minimum energy consumption [5] are not easy to achieve. In most cases, deploying an energy-efficient solution inevitably degrades the performance efficiency of the data centers.

To tackle this challenging issue, our preliminary work [6] introduced the concept of profiling for application assignment to Virtual Machines (VMs). It formulated the application assignment as a linear optimization with utilization of fully synthetic application and VM profiles. It also developed a simple profile matching algorithm to solve the optimization problem. The aim of the preliminary work was to introduce the profiling concept as a feasible and scalable application assignment method.

Extending our preliminary work significantly for improved solutions, this paper aims to develop a new profile-based application assignment framework for greener and more energy-efficient data centers. The new framework uses realistic profiles and also fulfils energy, resource and performance constraints or requirements. In comparison with our preliminary work [6], distinct contributions of this paper include the following four aspects:

- Physical Machine (PM) profiles: In addition to application and VM profiles, PM profiles are integrated into the profile-based application assignment, enabling derivation of actual energy savings of the servers from the application assignment;
- Profile building: Different from synthetic profiles, realistic application, VM and PM profiles are built from raw data of a real-world data center through systematic methods, allowing more realistic application assignment based on profiles;
- Optimization framework: a penalty-based linear optimization framework is formulated for profile-based application assignment with consideration of memory constrains in addition to CPU resources; and
- Solution algorithm: Refined from a simple profile matching algorithm, a penalty-based profile matching algorithm (PPMA) that uses some heuristics is presented to solve the new penalty-based optimization problem with considerations of memory, CPU and performance constraints.

Moreover, new and comprehensive case studies are carried out in this paper to demonstrate the effectiveness of the Profiling approach. The experimental results are compared with those from the commonly used general approach and workload history based application management strategy.

The energy management of a virtualized data center can be implemented at three layers: application, VM and PM layer, as shown in Fig. 1. The application management at the top layer assigns applications to VMs. The VM management layer is responsible for VM placement to PMs, VM sizing and VM migration. The PM management layer at the bottom layer is in charge of ON/OFF operations of PMs, sleep cycles, cooling and DVFS. While each of the three layers contributes to the overall data center energy savings, this paper limits its scope to the application management layer. Applications requested by cloud consumers or data center users are assigned to VMs, thereby allowing access to data center resources such as CPU and memory. The application assignment strategies typically consider application runtime, server workload, resource requirements or availability, energy consumption and performance efficiency. Thus, one of the key objectives of our research is to create such an energy-efficient application management strategy whilst maintaining the data center performance efficiency. Our investigation into the application assignment to VMs complements current research on the problem of VM placement to PMs.

Among various data centers, a big class of widely deployed data centers with nearly consistent workload and applications is investigated in this paper. These data centers are generally managed by universities, government agencies, and small corporate businesses. According to our investigations into a real-world data center, such data centers typically have a well-defined workload characterized by an almost constant number of VMs. The number of VMs hosted in PMs is typically reviewed every three to six months during which no adjustment is made. From the raw data collected from the real data center and using a workload model, this paper generates load synthetically for profile-based application assignment to VMs.

The paper is organized as follows. Section 2 reviews related work and motivates the research. Section 3 discusses the concept of profiles and the methodology of building profiles. A profile-based energy-efficient framework is presented in Section 4 with framework formulation and algorithmic solution. Experimental studies are conducted in Section 5. Finally, Section 6 concludes the paper.

2. Related work and motivations

Energy-efficiency of data centers has been a focus of many studies from various perspectives. In the work [7], energy consumption was minimized for fattree data center networks. The issue of colocation demand response was investigated by Ren and Islam [8]. Yoon et al. [9] presented techniques of efficient data mapping and buffering for multilevel cell phase-change memories. The work by Kumar et al. [10] studied cloud data management through workload-aware data placement and replica strategies.