Computers and Electrical Engineering 40 (2014) 1621–1633

Contents lists available at ScienceDirect

Computers and Electrical Engineering

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

Service level agreement based energy-efficient resource management in cloud data centers ☆

Yongqiang Gao [a], Haibing Guan [a,*], Zhengwei Qi [a], Tao Song [a], Fei Huan [a], Liang Liu [b]

[a] Shanghai Key Laboratory of Scalable Computing and Systems, Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[b] IBM Research – China, Beijing 100193, China

ARTICLE INFO

Article history:
Available online 25 November 2013

ABSTRACT

As cloud computing has become a popular computing paradigm, many companies have begun to build increasing numbers of energy hungry data centers for hosting cloud computing applications. Thus, energy consumption is increasingly becoming a critical issue in cloud data centers. In this paper, we propose a dynamic resource management scheme which takes advantage of both dynamic voltage/frequency scaling and server consolidation to achieve energy efficiency and desired service level agreements in cloud data centers. The novelty of the proposed scheme is to integrate timing analysis, queuing theory, integer programming, and control theory techniques. Our experimental results indicate that, compared to a statically provisioned data center that runs at the maximum processor speed without utilizing the sleep state, the proposed resource management scheme can achieve up to 50.3% energy savings while satisfying response-time-based service level agreements with rapidly changing dynamic workloads.

1. Introduction

Cloud computing [1] has recently received considerable attention in both academic community and industrial community as a promising approach for sharing resources that include infrastructures, software, applications, and business processes. As a direct result of cloud computing's increasing popularity, cloud computing service providers such as Amazon, Google, IBM and Microsoft have begun to establish increasing numbers of energy hungry data centers for satisfying the growing customers resource (e.g. computational and storage resources) demands. Cloud computing model provides flexibility and reduction in management costs. Cloud users can access the required services from anywhere in the world on an on-demand and pay-per-use basis. From the point of view of the cloud operator, the key issue is to maximize their profit margins by minimizing the operational costs of the data center. Power management has become an increasingly prominent concern in cloud data centers since operating costs of a cloud data center are dominated by their power consumption. As reported in [2], it is expected that by 2014, infrastructure and energy (I&E) costs would contribute about 75% of the total operation cost, while information technology (IT) expenses, i.e., the equipment itself, would contribute a significantly smaller 25% of the cost (compared with 20% I&E and 80% IT in the early 1990s). What's more, it is worth noting that the cost to manage a server will exceed the cost to buy one [3]. The rated power consumption of a typical server is estimated to have increased tenfold over the last ten years [4].

☆ Reviews processed and recommended for publication to Editor-in-Chief by Associate Editor Dr. Jian Li.
* Corresponding author. Tel./fax: +86 21 3420 7150.
E-mail addresses: gaoyongqiang@sjtu.edu.cn (Y. Gao), hbguan@sjtu.edu.cn (H. Guan), qizhwei@sjtu.edu.cn (Z. Qi), songt333@gmail.com (T. Song), huanfei@sjtu.edu.cn (F. Huan), liuliang@cn.ibm.com (L. Liu).

0045-7906/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compeleceng.2013.11.001

A number of key technologies have been developed over the last years to reduce the power consumption of cloud data center. Virtualization and consolidation are two crucial methods to achieve energy saving. Virtualization is a technology that partition a physical server into a number of small, virtual servers. Modern cloud computing systems exploit virtualization to achieve significant gains in terms of flexibility and power consumption. With the rapid development of virtualization technology, such as VMware, Xen, KVM and OpenVZ, server consolidation can be achieved by consolidating applications running on a large number of low utilization servers to a smaller number of highly utilized servers, each of which is encapsulated in virtual machines (VMs) separately. Server consolidation using virtualization is an effective approach to achieve better energy efficiency of cloud data center. The reason is that at times of low load, VMs are consolidated on a limited subset of the available physical resources, so that the remaining (idle) computing nodes can be switched to low power consumption modes or turned off. Dynamic voltage/frequency scaling (DVFS) is also a commonly-used power-management technique. The key idea behind DVFS techniques is to dynamically scale the frequency and voltage of the microprocessor at runtime according to processing needs and thereby reducing the energy dissipation. DVFS techniques has proven to be a highly effective method of reducing data center power consumption [5]. Many commercial microprocessors such as Intel's XScale and Transmeta's Cruso are now equipped with DVFS capabilities.

Cloud providers face the challenge of two contradicting goals, namely reducing energy consumption of the cloud infrastructure while ensuring compliance with service level agreements (SLAs) negotiated between customers and cloud providers. SLAs generally specify performance-related quality of service (QoS) properties, such as response time and availability, that must be maintained by a cloud provider during the services runtime. On one hand, the reduction of energy consumption helps to reduce the total cost of ownership (TCO) and increases the return on investment (ROI) of cloud infrastructures. However, on the other hand, a violation of SLAs may lead to reduced customer satisfaction as well as penalty payments. Thus, an important issue in resource management for cloud environments is how to correctly provision resource, such that service level agreement (SLA) requirements are met while minimizing power consumption.

This paper presents a dynamic resource management scheme that is able to automatically manage physical resources of a cloud infrastructure in such a way to maximize the profit of the cloud provider by minimizing SLA violations while reducing the energy consumed by the physical infrastructure. Our scheme utilizes both DVFS and server consolidation to minimize power consumption for cloud data centers while providing application-level performance guarantees. We integrate timing analysis, queuing theory, integer programming and control theory to achieve the optimal dynamic configuration of the cloud data center, that is, which servers must be active and their respective CPU frequencies. We experimentally validate it in a prototype data center consisting of 12 servers with a multi-tier application benchmark. Our experimental results indicate that, compared to a statically provisioned data center that runs at the maximum CPU speed without utilizing the sleep state, the proposed system can achieve up to 50.3% energy savings while satisfying response-time-based SLAs with rapidly changing dynamic workloads.

The rest of the paper is organized as follows. Section 2 describes related works. Section 3 describes the system model. Section 4 describes the heuristic algorithm we propose. Section 5 gives analysis of our resource actuator. Section 6 describes the experiment results. We conclude in Section 7.

2. Related work

In recent years, extensive efforts have been put into the research of performance modeling for virtualized systems. Kalyvianaki et al. [6] presented a new resource management scheme that integrates the Kalman filter into feedback controllers for dynamically allocating the CPU resources of multi-tier virtualized servers. Rao et al. [7] proposed a response time-based fuzzy control approach for resource allocation in virtualized environments. Wang et al. [8] developed a distributed control framework that dynamically optimizes the CPU capacity provided to each VM in response to incoming workload intensity such that desired response times are satisfied. In contrast, we combine queueing prediction and feedback control to dynamically adjust the CPU resources allocated to each VM in order to maintain the desired response time for multi-tier applications.

There is a huge body of work on reducing power consumption in data centers. Krioukov et al. [9] presented a power-aware cluster manager, called *NapSAC*, which minimizes the server idle power waste by keeping the minimum number of servers awake for matching demand. Elnozahy et al. [10] proposed five distinct policies for applying various combinations of intra-node (dynamic voltage scaling) and inter-node (node vary-on/vary-off) power management mechanisms to reduce the power consumption of a server cluster. Singh et al. [11] proposed a new energy optimization methodology using the Potluck Problem concept to model the optimization of resource allocation in large and complex IT systems. These studies are different from our work because they focus solely on reducing power consumption of the infrastructure and thus cannot provide guarantees for performance targets of hosted applications.

A significant number of papers have appeared in the literature examining dynamic performance- and power-aware resource management in cloud infrastructures. In [12], the authors proposed a two-levels control architecture that provides explicit guarantees on both power and application-level performance for virtualized clusters. Guazzone et al. [13] proposed a resource management framework to automatically adjust physical and virtual resources of a cloud infrastructure in order to simultaneously achieve suitable QoS levels and to reduce the energy consumed by the physical infrastructure. More recently, Beloglazov et al. [14] developed several novel heuristics for dynamic consolidation of VMs to reduce energy