Computers and Electrical Engineering 60 (2017) 30–44

Contents lists available at ScienceDirect

Computers and Electrical Engineering

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

A novel performance constrained power management framework for cloud computing using an adaptive node scaling approach☆

S. Suresh[a,*], S. Sakthivel[b]

[a] Department of Computer Science and Engineering, P.A. College of Engineering and Technology, Pollachi, Tamilnadu, India
[b] Department of Computer Science and Engineering, Sona College of Technology, Salem, Tamilnadu, India

ARTICLE INFO

Article history:
Received 23 January 2016
Revised 20 April 2017
Accepted 20 April 2017
Available online 12 May 2017

Keywords:
Cloud computing
Server virtualization
Adaptive algorithms
Power management
Load balancing
System modeling

ABSTRACT

Cloud computing is an on-demand IT resource delivery technology that is aided by server virtualization and load balancing. Power and performance management to improve operational efficiency and increase compaction are important considerations from a cloud service economic point of view. The objective of the present study was to draw new insights from existing approaches and techniques to design an innovative self-adapting mechanism to address the mismatch between server's energy-efficiency characteristics and the behavior of server-class workloads, which solves the power versus performance trade-off problem at cloud data centers. The proposed system was simulated and evaluated for highly variable cloud workloads. The results suggest that the proposed system functions reliably for cloud workloads and ensures an optimal server workload distribution (i.e., determines the allocations of the VM server), minimizing the average power consumption of the servers and ensuring that the average task response time does not exceed given performance limitations.

1. Introduction

Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable IT resources that can be rapidly provisioned and released with minimal management effort [1,2]. Server virtualization and load balancing are the underlying technologies that drive cloud computing for providing IT resources as a service. Server virtualization is a technology that reduces power consumption by improving server resource utilization. Thus, it has great potential for reducing energy and hardware costs via server consolidation. Load balancing is an optimization technique that distributes service requests to resources evenly across all of the available nodes in the entire cloud to avoid situations where certain nodes are heavily loaded while others are idle or underutilized. Thus, load balancing improves host utilization, thereby reducing energy consumption. Server virtualization and load balancing present an efficient way to run a cloud data center from a power and performance management point of view [3].

Power management is an important economic consideration because effective power management improves operational efficiency and increases compaction. As the CPU is the main hardware component, its energy consumption model helps

☆ Reviews processed and recommended for publication to the Editor-in-Chief by associate editor Dr. M. S. Kumar.
* Corresponding author.
E-mail addresses: ssuresh.siv.72@gmail.com (S. Suresh), sakvel75@gmail.com (S. Sakthivel).

http://dx.doi.org/10.1016/j.compeleceng.2017.04.018
0045-7906/

reveal the dynamic characteristics of drawn server power to develop suitable strategies for improving energy efficiency. In recent years, researchers have proposed several techniques, such as Dynamic Voltage and Frequency Scaling (DVFS), power state transitioning and server consolidation (switching server / node resources into the low-power mode or turned-off state), workload management or task scheduling and thermal-aware power management for managing the power consumption at cloud data centers [4]. Unfortunately, many types of server CPUs do not have as many levels of voltage and frequency as the CPUs of embedded devices; therefore, the power savings achieved by adjusting frequency and voltage vary significantly across CPU types. Hence, solely adjusting frequency and voltage cannot solve the power conservation problem [5]. Consequently, saving energy by switching idle servers on (scale-up servers) or off (scale-out servers) using an adaptive node scaling approach is important. Adaptive node scaling is a closed-loop dynamic power minimization technique that reduces power consumption by switching superfluous Virtual Machine Server (VMServer) computing nodes on or off the based on the actual incoming workload, i.e., the VMServer computing nodes are continuously adjusted by switching unneeded VMServers on or off during the run time of the VMServer cluster.

Numerous load balancing algorithms for conventional distributed systems have been reviewed and proposed over the past several years [6]. However, these algorithms are potentially unsuitable for cloud environments due to the unique characteristics of the cloud. Certain proposed load balancing algorithms take these cloud characteristics into account but do not follow system status changes. Others [7,8] set a fixed balance threshold for controlling the load situation of the entire cloud system. Specific heuristic algorithms [9–11] consider server setup time control [9], multi-core CPU architecture [10] and / or workload forecasting [11] parameters according to changes in the environment or type of job. However, insufficient research has been conducted on power management load balancing, and the aforementioned algorithms might be unsuitable for the highly varying workload in a dynamic cloud environment.

1.1. Motivation and objectives

The cloud introduces many challenges, for example, the effective power and performance management of cloud Virtual Machine (VM) servers for guaranteeing Quality of Service (QoS) satisfaction and minimizing Service Level Agreement (SLA) violations despite highly varying application workloads. By scaling capacity to match current demand, operators can reduce power consumption, decrease rental costs and get additional work done by repurposing unneeded VM servers for other tasks [12].

Virtualized server architecture is still far from being power aware SLA constrained performance proportional in that a significant amount of power and performance is lost when the server is virtualized. Thus, improvements to data center server consolidation are needed for reducing power and improving performance. The many challenges associated with this approach include (i) performance trade-offs due to dynamic resource management for unexpected loads, (ii) load unpredictability and (iii) short idle times and the energy cost of switching VM servers to lower power modes. Addressing the mismatch between a server's energy efficiency characteristics and the behavior of server-class workloads, the present study aimed to develop an energy proportional server consolidation algorithm that conserves server utilization energy. In a sense, the goal of our research was to maintain the availability of VM servers to satisfy the desired SLA while reducing the total power consumed by the cloud. Thus, the central theme of this study was to explore how virtualization allows for application agnostic solutions when dealing with power and performance management challenges.

To solve the above problem, we took advantage of adaptive algorithms to analyze the energy consumption of a cloud computing system. The main objectives and contributions of this paper include: (i) to explore, analyze and classify previous research in energy-efficient computing to gain a systematic understanding of existing techniques and approaches, (ii) to propose a framework by proactively adapting (a) the frequency of algorithm invocation and (b) the upper and lower utilization threshold as a function of the workload patterns of the applications during the lifetime of the VM, (iii) to demonstrate that the proposed framework significantly reduces VM setup time and effectively utilizes CPU resources, (iv) to demonstrate that the proposed system can effectively manage power for varying cloud loads without violating performance goals with respect to load balancing algorithms, (v) to validate the system via the CSIM simulation toolkit on a 12 to 7 server test bed with a highly varying cloud workload, (vi) to analyze the impact of performance control on the power model and confirm the control accuracy and system stability of the performance controller even in the face of model variation and (vii) to prevent the power usage of the virtualized server clusters from scaling up and down under varying loads.

The remainder of the paper is organized as follows: Section 2 reviews the state-of-the-art research in power and performance management at cloud computing data centers and highlights the distinction of our work, Section 3 presents a solution for the performance constrained power management problem with a novel adaptive performance constrained power management load distribution algorithm, Section 4 presents the modeling, design and analysis of the proposed system algorithm controller and provides the implementation details of each component in the control loops, Section 5 highlights our simulation experiments and presents our statistical results and study observations and Section 6 provides a conclusion and potential areas of future study.

2. Related work

Server consolidation with load balancing improves host utilization, thereby reducing power consumption. Several strategies have been proposed for optimizing and managing the energy in cloud computing systems, including energy-efficient