2015 International Conference on Parallel Architecture and Compilation

DVFS-Aware Consolidation for Energy-Efficient Clouds

Patricia Arroba, José M. Moya
Integrated Systems Lab. - CSS
Universidad Politécnica de Madrid
Madrid, Spain
e-mail: parroba, josem@die.upm.es

José L. Ayala
DACYA
Complutense University of Madrid
Madrid, Spain
e-mail: jayala@fdi.ucm.es

Rajkumar Buyya
CLOUDS Laboratory
The University of Melbourne
Melbourne, Australia
e-mail: rbuyya@unimelb.edu.au

***Abstract*—Nowadays, data centers consume about 2% of the worldwide energy production, originating more than 43 million tons of CO2 per year. Cloud providers need to implement an energy-efficient management of physical resources in order to meet the growing demand for their services and ensure minimal costs. From the application-framework viewpoint, Cloud workloads present additional restrictions as 24/7 availability, and SLA constraints among others. Also, workload variation impacts on the performance of two of the main strategies for energy-efficiency in Cloud data centers: Dynamic Voltage and Frequency Scaling (DVFS) and Consolidation. Our work proposes two contributions: 1) a DVFS policy that takes into account the trade-offs between energy consumption and performance degradation; 2) a novel consolidation algorithm that is aware of the frequency that would be necessary when allocating a Cloud workload in order to maintain QoS. Our results demonstrate that including DVFS awareness in workload management provides substantial energy savings of up to 39.14% for scenarios under dynamic workload conditions.**

Cloud Computing; Dynamic Voltage and Frequency Scaling; Dynamic Consolidation; Energy Efficiency

I. INTRODUCTION

Computational demand on data centers is increasing due to the growing popularity of Cloud applications. However, data centers are becoming unsustainable in terms of power consumption and growing energy costs so they must be placed on a more scalable curve. Recently, there has been a growing interest in developing techniques to provide power management in Clouds. Dynamic Voltage and Frequency Scaling (DVFS) helps to reduce the consumption of underutilized resources dynamically, while consolidation strategies decrease significantly the static consumption by reducing the number of active servers, thus increasing their utilization. However, DVFS is traditionally applied locally, regardless the consolidation techniques. Understanding the relationship between power, DVFS and consolidation is crucial to enable new energy-efficient strategies that combine these effective techniques. To this purpose, the dependency of power on some traditionally ignored factors like frequency or static consumption, which are increasingly influencing the consumption patterns of these infrastructures, must now be considered. Furthermore, as Cloud services are provided under strict Service Level Agreement (SLA) conditions, power consumption in data centers may be minimized, taking into account a trade-off between DVFS and performance, without violating the SLA requirements whenever it is feasible. Also, Cloud workloads vary significantly over time, difficulting the optimal allocation of resources that requires a tradeoff between consolidation and performance. Therefore, the implementation of consolidation policies that are aware of both DVFS and energy consumption while considering QoS has the potential to optimize the sustainability of Cloud data centers.

II. PROPOSED SOLUTION

In this work we aim to find an energy optimization strategy for Cloud data centers that combines DVFS and consolidation techniques. Our policy is not only aware of the utilization of the incoming workload to be assigned, but also is conscious of the impact of its allocation in terms of frequency. One of the main challenges when designing data center optimizations is to implement fast algorithms that can be evaluated during runtime. For this reason, our research is focused on the design of an optimization algorithm that is simple in terms of computational requirements, in which both decision making and its execution in a real infrastructure are fast. The proposed algorithm is based on a bin packing problem [1] where servers are represented as bins with variable sizes due to the frequency scaling. To design our optimization technique, we first characterize performance and power contributions in terms of those architectural parameters most influenced by DVFS and consolidation [2]. The obtained power model offers an accuracy of about 4.46% and allows a better understanding of how energy varies depending on frequency and utilization simultaneously. In our proposed optimization, VMs are consolidated in those hosts that have a high utilization but, on the contrary, have a low increase in frequency due to the utilization increment. This approach minimizes the number of bins used by this combinatorial NPhard problem while taking full advantage of the range of CPU utilization available for each frequency.

III. RESULTS

We have performed an extensive evaluation on CloudSim [3] that represents accurately the modeling of virtualized data centers. Our application framework consists of real Cloud traces from the global research network PlanetLab [4]. Hosts are modeled as Fujitsu RX300-S6 servers based on Intel Xeon E5620 @2.4GHz, virtualized by the QEMU-KVM hypervisor. VM instances correspond to existing types of Amazon EC2. To evaluate the performance

1089-795X/15 $31.00 © 2015 IEEE
DOI 10.1109/PACT.2015.59

of our frequency-aware optimization we compare our work with two different approaches where the frequency is not considered during consolidation: the Baseline and the DVFS-only scenarios, where the DVFS is switched off and on respectively. The three provided scenarios, including our frequency-aware optimization, Freq-Aware, are tested for 15 different tests, each of them representing a specific combination of overloading detection and VM selection algorithms in the consolidation process.

Our algorithm speeds up consolidation and the elastic scale out of the IT infrastructure, presenting a global utilization increase of up to 23.46% by reducing the number of active hosts by 44.91% (see Figure 1). Our strategy reduces the number of VM migrations by 22.17% and by 19.61% when compared with Baseline and DVFS-only. This behavior impacts on the energy usage of the data center, where the consumption of both Baseline and DVFS-only grows at a higher rate during dynamic workload variations than for Freq-Aware scenario. We achieve competitive energy savings of 37.86% and 35.72% in average respectively (see Figure 2), maintaining QoS, even improving slightly SLA violations around 0.01%.

IV. Novelty and Contributions

The key contribution of our work is a novel frequency-aware consolidation algorithm that reduces the energy consumption of the data center while maintaining its QoS. The algorithm is light, scalable and offers an elastic scale-out under varying demand of resources, making it suitable for quickly adaptation to workload fluctuations in the data center. We have performed an extensive evaluation using real Cloud traces and an accurate power model based on data gathered from real servers. Our results demonstrate that including DVFS awareness in workload management provides substantial energy savings of up to 51.62% for scenarios under dynamic workload conditions.

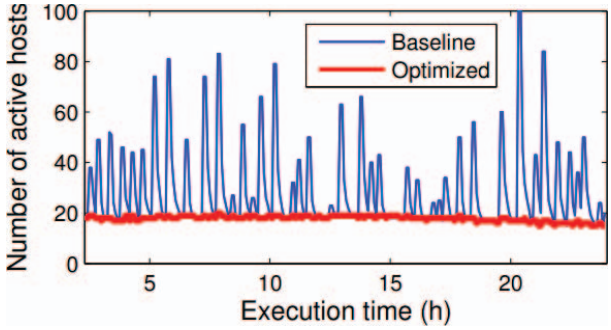

Figure 1. Number of active hosts during runtime

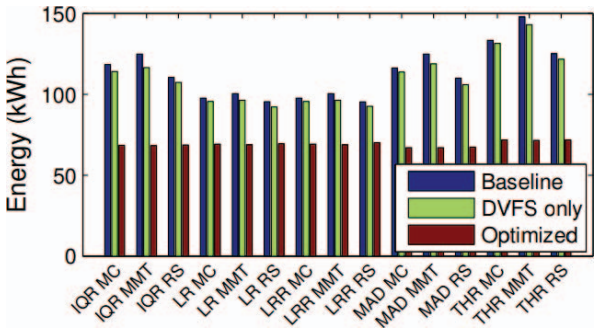

Figure 2. Average energy consumption comparison per test

Acknowledgment

This research has been partly supported by a collaboration fellowship from the European Network of Excellence on High Performance and Embedded Architecture and Compilation (HiPEAC), the Spanish Ministry of Economy and Competitiveness, under research grants TEC2012-33892, IPT-2012-1041-430000, RTC-2014-2717-3 and by a project grant from the Australian Research Council (ARC).

References

[1] A. Beloglazov and R. Buyya, "Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers," Concurr. Comput. : Pract. Exper., vol. 24, no. 13, Sep. 2012, pp. 1397–1420.

[2] P. Arroba, J. L. Risco-Mart´ın, M. Zapater, J. M. Moya, J. L. Ayala, and K. Olcoz, "Server power modeling for runtime energy optimization of cloud computing facilities," Energy Procedia, vol. 62, no. 0, 2014, pp. 401 – 410.

[3] R. N. Calheiros, R. Ranjan, A. Beloglazov, C. A. F. De Rose, and R. Buyya, "Cloudsim: A toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms," Softw. Pract. Exper., vol. 41, no. 1, Jan. 2011, pp. 23–50.

[4] Park and V. S. Pai, "Comon: A mostly-scalable monitoring system for planetlab," SIGOPS Oper. Syst. Rev., vol. 40, no. 1, Jan. 2016, pp. 65–74.

.