Autonomic Resource Management System for Virtualized Environments

Govinda Mohini G. Bezerra, Otto Carlos M. B. Duarte
Grupo de Teleinformática e Automação – PEE/COPPE – DEL/POLI
Universidade Federal do Rio de Janeiro – Rio de Janeiro – RJ – Brasil

***Abstract*—Virtualization is a growing technology that can substantially increase flexibility and manageability in data centers. This paper proposes an autonomic resource management system for virtualized environments. The tool builds and analyses usage profiles of physical and virtual machines. Based on the profiles, the algorithm can detect resource scarcity and distribute the load among the monitored physical machines, reducing idle resources and avoiding Service Level Agreement (SLA) violations.**

I. Introduction

In typical data centers, the server average utilization is about 30% [1]. The low utilization of the servers is a major concern for data center managers because energy costs are expressive and low machine utilization implies that a substantial amount of the energy is wasted by idle systems.

Virtualization has been widely employed in data centers because it is technology that facilitates the management of resources, increases IT infrastructure flexibility and it can significantly reduce idle resources. One of the major challenges in virtualized environments is to properly manage the computational resources allocation to increase resources utilization rate without violating the service level agreements (SLAs). The difficulty lies in ensuring sufficient resources allocation to properly run the applications on the virtual machine, despite the dynamic nature of workloads.

II. Proposed System

The proposed system aims to monitor and manage physical and virtual machines in order to allow high resources utilization rates and to ensure the fulfillment of the SLAs. The key idea of the proposal is to create usage profiles of different resources, such as processor, memory and network, for all physical and virtual machines of a cluster and, based on these profiles, to detect overloads and distribute workload through virtual machine migration. The live migration feature allows moving a running virtual machine between two physical machines without stopping the applications.

The proposed system uses a black box approach that collects usage data by simply observing each virtual machine (VM) from the outside, without any knowledge of the application resident within each VM. This approach was chosen because it does not directly interfere in the execution of the virtual machine and ensures privacy, which is highly desirable in virtualized environments, since, in many scenarios, the hardware owner is not necessarily the applications owner.

The system architecture consists of three main modules. The first module is the Profile Generator that is responsible for tracking the processor, network and memory usage of each virtual and physical machine. It interacts with the physical machines to retrieve usage samples of the monitored resources and uses this data to build usage profiles. The second module is the Overload Detector, which uses the generated profiles to detect resource scarcity and also sorts the physical machines by charge. The third module is the Migration Orchestrator that is trigged by the Overload Detector module when a physical machine is overcharged. The Orchestrator will act in order to reduce the load on the overcharged host, choosing a virtual machine to be migrated and a physical machine with sufficient resources to accommodate the chosen VM.

In order to validate the proposal, the system was tested in the FITS testbed [2] and we also implemented a control panel that shows the usage profile of all monitored machines in real time. We designed a test scenario that allows testing the system in different cases of overload and the obtained results show that the developed system is able to satisfactorily meet the proposed objectives.

III. Conclusion and Future Work

The work proposes an autonomic resource management system that monitors resource usage, detects overloads and distributes workloads among physical machines in order to prevent applications performance degradation. As future work, we intend to integrate this tool with FITS testbed adding it to the testbed web interface, so it can assist and help to validate the proposals tested in the platform FITS.

Acknowledgments

The authors would like to thank fiNEP, FUNTTEL, CNPq, CAPES, FAPERJ and UOL for their financial support.

References

[1] D. Meisner, B. T. Gold, and T. F. Wenisch, "Powernap: eliminating server idle power." in *ASPLOS*, M. L. Soffa and M. J. Irwin, Eds. ACM, 2009, pp. 205–216. [Online]. Available: http://dblp.uni-trier.de/db/conf/asplos/asplos2009.html#MeisnerGW09

[2] P. H. Guimarães, L. Ferraz, J. V. Torres, D. Mattos, A. Murillo, M. A. Lopez, I. Alvarenga, C. Rodrigues, and O. C. M. B. Duarte, "Experimenting Content-Centric networks in the future internet testbed environment," in *IEEE International Conference on Communications 2013: IEEE ICC'13 - Workshop on Cloud Convergence: challenges for future infrastructures and services (WCC 2013) (ICC'13 - IEEE ICC'13 - Workshop WCC)*, Budapest, Hungary, Jun. 2013, pp. 1398–1402.