International Journal of Performability Engineering, Vol. 13, No. 2, March 2017, pp. 240-242

Self-Optimization in Cloud Computing Considering Reliability and Energy

PENG SUN, DEMIAO WU*, SHENGJI YU and YANPING XIANG

School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, P.R. China

(Received on September 5, 2016; Revised on November 24, 2016 and February 14, 2017)

Abstract: In the virtual data-center, how to map virtual machines (VMs) to physical machines (PMs) is becoming a hot issue. However, most of existing VM scheduling schemes have not fully considered the reliability and dynamical workloads of VMs. This paper presents a novel bionic autonomic nervous system (BANS) based approach for cloud resource management. This approach supports self-optimization that provides a dynamic and autonomic way to adapt to dynamical workloads and VM resource requirements. For the VM allocation in the self-optimization, this paper presents a reliability-performance-energy correlation model that can model, analyze and evaluate reliability, performance and power consumption simultaneously.

Keywords: *Cloud Resource Management, Reliability, Correlation Model, Self-optimization*

1. Introduction

One of the major service of Cloud Provider (CP) is IaaS, which offers resources to Service Providers (SPs) on a pay-per-usage basis. In turn, each SP delivers the services to end users by using the offered resources over the Internet [1]. When the SP orders a group of virtual machines (VMs), how to map these VMs to physical machines (PMs) with meeting the service level agreements (SLAs) is the prime problem for CP. Although it has been studied so far, most of them devote to improve cloud service performance, reduce energy consumption or balance the performance-energy and ignore the reliability aspect of PM failure caused by individual hardware failures. Beyond that, studies have found that cloud service's workloads and VM resource requirements are highly dynamic [2]. Thus, traditional models and approaches could not be well adapted for large-scale cloud resource management.

In view of this, we design a novel bionic autonomic nervous system (BANS) based cloud resource management system and it supports self-optimization, which provides an autonomic behavior to adapt to variable workloads and VM resource requirements that significant reduce power consumption, reliability and performance degeneration by the collaboration of BANS components. In order to maximize the CP's net profit by striking a balance between reliability, performance and energy, this paper also presents a reliability-performance-energy correlation model that can model, analyze and evaluate reliability, performance and power consumption simultaneously for the VM allocation problem in BANS based self-optimization.

*Corresponding author's email: demiaowu@163.com

2. Model Overview

This section presents an overview of the BANS based cloud resource manager model. As is showed in Figure 1, the model is composed of the following five components:

Cyber Axon (CA) is similar to axon of biology. It is the data collector plug-in for monitoring specified resource. One CA maps one plug-in, which is responsible for one type resource's monitoring. These CAs in the PM are used for monitoring the PM's resource utilization, while the VM has its own CAs to monitoring VM's resource utilization.

Cyber Neuron (CN) refers to the VM. It emulates the biological neurons and dedicates to the data analysis/prediction on the monitor data collected by CAs in the VM. By the analysis of historical monitor data, it can catch the time-varying resource intensity in VM.

Cyber Peripheral Nervous (CPN) is analogous to the peripheral nerve of biology. It is capable of local autonomy in PM. It analyzes/predicts the time-varying resource intensity and assesses the resource utilization state of PM.

Cyber Central Nervous (CCN) is analogous to central nervous of biology. It breaks the whole datacenter into smaller clusters. By collecting and further analyzing the data from CNs and CPNs, CCN guides dynamic resource allocation scheduling to achieve self-optimization. Section 3 describes it in details.

Cyber Brain (CB) is similar to the brain used for global management. It is responsible for receiving SP's requests which order a group of VMs and selecting an appropriate CCN to provide these VMs. Each request carries some requirements information such as the parameters of VMs, reliability, performance metrics etc.

Compared with the centralized approach, the BANS based approach avoids the single point of failure by dividing whole data-center into several server clusters managed by CCN.

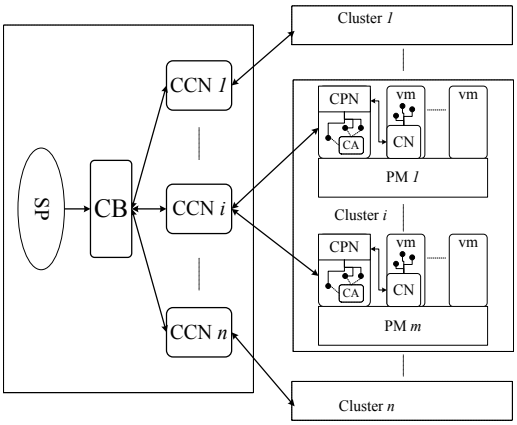

Figure 1: The BANS Based Cloud Resource Management Model

.

3. BANS Based Self-Optimization

Due to the different time-varying demands on different resources for VMs [3], the BANS based self-optimization approach is proposed. As shown in Figure 2, it is an autonomic way that implements dynamic resource allocation scheduling by using live migration to adapt to VM workload variability and dynamical different resources demand. It aims to find the under-load PMs, and then consolidate VMs onto as few PMs as possible for reducing the energy and wasting of resources, or find overload PMs, and then select VMs to migrate out for reducing reliability and performance degeneration. Note that this paper not considers the influence of migration on the cloud services' SLAs. Since with rapid development of virtualization and container (eg. Docker) techniques, the migration of VMs is very fast that almost have no influence on the cloud service.

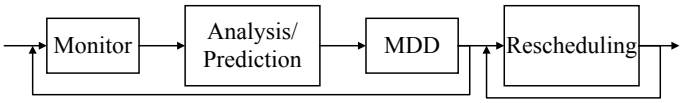

Figure 2: BANS Based Self-Optimization

3.1 Monitor

CPN and CN monitor the different resources by CAs (data collector plug-in) periodically and store these data for future analysis.

3.2 Analysis/Prediction

In CPN, the analysis/prediction process involves two phases: 1) Analyze historical monitor data and predict future usage of each resource in PM. 2) Evaluate each resource utilization state respectively. Considering that different types of resources have different thresholds [4], we give three different state determination range, i.e. underload, minor and overload. In CN, the analysis/prediction process only need to analyze historical monitor data and predict future usage of each resource in VM. We use the approaches in [3, 4] to analyze historical monitoring data and predict future usage of the resource.

3.3 Multivariate Decision Diagram (MDD)

In CPN, after determining the utilization state of each resource respectively, MDD is employed to quickly assess the resource utilization states of PM. Figure 3 shows a MDD of PM state using a special kind of MDD called a ternary decision diagram. It is a directed acyclic graph (DAG) with up to 3 sink nodes labelled by a distinct logic value 0, 1, 2 which correspond to PM entire resource utilization state underload, minor, overload respectively. Each non-sink node is labelled with a ternary-valued variable which represents one resource's utilization state, and it has three outgoing edges called "0-edge", "1-edge", and "2-edge" (from left to right), which represent the three resource utilization levels-underload, minor and overload respectively. In Fig 3, non-sink node P, M, and B represent the PM's utilization state of CPU, memory, and bandwidth respectively. For example, if the resources utilization states obtained from Analysis/prediction module are CPU-underload, memory-minor and bandwidth-minor, then the path from P via 0-edge to M, and then from M via 1-edge to B, and finally via 1-edge to PM state 1 (minor). If the PM state outputted from MDD is 0-underload or 2-overload, the CPN will send the PM state back to CCN.

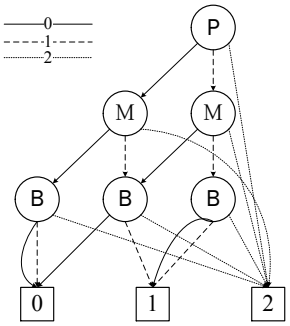

Figure 3: Resource Utilization State

3.4 Rescheduling

Finally, when CCN receives the feedback information from CPN and it will trigger rescheduling according to different cases. Case 1, if the feedback reveals the overload of the PM, rescheduling is to select some VMs and migrate them from overload PM and then select destination PMs to migrate to; Case 2, if the feedback shows the underload of the PM, rescheduling only tries to migrate all VMs from the underload PM and switch it to the sleep mode. The method presented in [3,4] is used to decide which VM should be migrated away. Then, CCN can generate the rescheduling list of VMs and execute rescheduling.

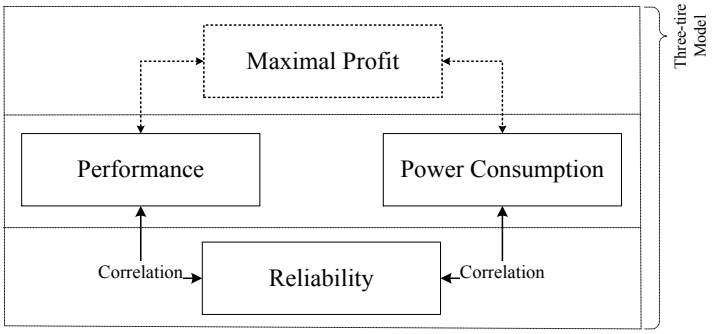

Figure 4: Three-Tier Reliability-Performance-Energy Correlation Model

The VMs rescheduling is the core process for the BANS based self-optimization that maximizes the CP's net profit. In order to analysis reliability, performance and energy simultaneously, a three-tire correlation model is proposed as shown in Figure 4. The evaluation of both performance and energy are based on analysis of hardware reliability in the first tire. The reliability-performance and reliability-energy are two important sub-models that indicate both performance and energy affected by reliability. Finally, the third tire profit model's goal is to develop an efficient reliability-, performance- and energy-aware VM allocation approach to maximize net profit by striking a balance between reliability, performance and energy. Formally, it can be formulated as follows:

$$\max \quad Netprofit = \left(U_{perf} - E_{power}\right) * t \tag{1}$$

Where, U_{perf} is the utility of performance and E_{power} is the spending of energy.

4. Conclusions

This paper presents a novel BANS based approach for cloud resource management and it supports self-optimization that can maximize the CP's net profit.

References

[1]. Zhang, Q., M.F. Zhani, M. Jabri, and R. Boutaba. *Venice: Reliable Virtual Data Center Embedding in Clouds*. In INFOCOM, 2014 Proceedings IEEE 2014 Apr 27:289-297.

[2]. Tighe, M., G. Keller, M. Bauer, and H. Lutfiyya. *A Distributed Approach to Dynamic VM Management*. International Conference on Network and Service Management, Oct 14, 2013;166-170.

[3]. Chen, L., H. Shen, and K. Sapra. (2014). *Rial: Resource Intensity Aware Load Balancing in Clouds*. INFOCOM, 2014, Proceedings IEEE Apr 27, 2014;1294-1302.

[4]. Xiao, Z., W. Song, and Q. Chen. *Dynamic Resource Allocation Using Virtual Machines for Cloud Computing Environment*. IEEE Transactions on Parallel and Distributed Systems, June 2013;24(6):1107-17.

Peng Sun is pursuing the Ph.D. degree at the Next Generation Internet and Data Processing Technology of National Local Joint Engineering Laboratory in University of Electronic Science and Technology of China (UESTC). His research interests include cloud computing, reliability modeling and optimization.

De-Miao Wu is a Master student at the National Local Joint Lab in UESTC. His research interests focus on cloud computing and reliability modeling.

Sheng-Ji Yu is pursuing the Ph.D. degree at the National Local Joint Lab in UESTC. His current research interests include cloud computing, cloud storage and operating system.

Yan-Ping Xiang received Ph.D. degree from National University of Singapore in 2003. Now, she is a Professor at School of Computer Science and Engineering in UESTC. Her current research interests include cloud computing and decision making.