International Journal of Grid and High Performance Computing
Volume 10 • Issue 2 • April-June 2018

An Approach for Detection of Overloaded Host to Consolidate Workload in Cloud Datacenter

Nimisha Patel, Rai University, Ahmedabad, India & Sankalchand Patel College of Engineering, Visnagar, India

Hiren Patel, LDRP Institute of Technology and Research, Gandhinagar, India

ABSTRACT

This article describes the process of workload consolidation through detection of overloaded hosts in Cloud datacenter which leads to saving in energy consumption. Cloud computing is a novice paradigm where virtual resources are provisioned on pay-as-you-go basis. Upon receiving users' job requirement, it is mapped onto virtual resources running on hosts in datacenter. To achieve workload consolidation, it is required to detect the overloaded hosts. Overloaded host detection is carried out for balancing workload, creating a list of overloaded hosts which will be useful while placing VMs (by not putting a VM on already overloaded host) to reduce Service Level Agreement (SLA) violation and while checking the underloaded host, the overloaded hosts are omitted to reduce computational cost. Most common mechanism to detect overloaded hosts is to calculate upper threshold values based on hosts' utilization statically or dynamically. Most researchers recommend dynamic calculation of threshold values. In this research, the authors propose to use moving range (MR) method of variables control charts to calculate upper threshold. The experimentation results show that MR performs better in terms of reduction in SLA violation, minimization in VM migration.

KEYWORDS

Cloud Computing, Energy Efficiency, Overloaded Host, Service Level Agreement, Workload Consolidation

INTRODUCTION

Using the strong backbone of grid, cluster and utility computing, Cloud has gain lot of attraction and adoption in just less than a decade. Because of the pay-as-you-go approach and avoiding the hassle of large IT infrastructure investment and maintenance, the usage of Cloud has been flooded like anything, from individual users to corporate clients and in the domain of e-Commerce, Reality sector, Academic and what not. Mell and Grance (2011) defines Cloud as a "model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction." The NIST further lists five essential characteristics of Cloud computing viz. (1) on-demand self-service, (2) broad network access, (3) resource pooling, (4) rapid elasticity or expansion, and (5) measured service. It also lists

DOI: 10.4018/IJGHPC.2018040105

three service models viz. (a) Software as a Service (SaaS), (b) Platform as a Service (PaaS) and (c) Infrastructure as a Service (IaaS), and four deployment models viz. (1) private, (2) community, (3) public and (4) hybrid. These Cloud services and deployment models assist programmers having ground-breaking thoughts without huge capital investment in computing infrastructures to deploy their products in the real market.

The Cloud data centers usually comprise of a great number of well-configured and interconnected computing resources (Luo, Li, & Chen, 2014) which consume a significant amount of electricity for their functioning. Increase usage of Cloud computing has lead to augmentation in this electrical energy consumption by the huge amount of servers in a large number of data centers. According to estimation (Christian & Belady, 2007), infrastructure and energy costs would contribute about ¾th whereas IT would contribute just ¼th to the overall cost of operating a data center in next three years. Data show that the average energy consumption of a data center is comparable to that consumed by 25,000 domestic usages (Kaplan, Forrest, & Kindler, 2008). This has attracted consideration of research community in recent years. Out of many different mechanisms to address the issue, workload consolidation and task scheduling have been recognized as few of the popular techniques. Server consolidation works on the principle of minimizing active servers in a data center without compromising the performance of tasks. Sleep/Wakeup has been identified as one of the top classifications by (Brienza et al., 2016) in which some of the servers are switched off when not in use to save energy and are awakened whenever necessary because it has been seen that even idle servers consume about 70% of their peak power (Fan, Weber, & Barroso, 2007). In a nutshell, proper distribution of existing tasks among available servers may result into minimizing the active servers without negotiation Service Level Agreements (SLA) with Cloud users.

In this research, we wish to address the scenario by identifying overloaded host from the data center. For doing so, we study various existing statistical methods used by researchers and propose to use a novel method, Moving Range (MR), which has not been used by the researchers so far. Using MR method, both upper and lower threshold values for host utilization can be computed, which help us in detecting overloaded or underloaded host, respectively. Detection of overloaded hosts would result into reduction in SLA violation and proper balance of existing workload. We further wish to reduce the overall energy consumption of data center by effective load balancing across the available host.

Rest of the paper is organized as follows. In next section, we describes related work pertaining to detection of overloaded host followed by discussion on various statistical methods used by researchers to compute threshold values along with our proposal. In subsequent part of this paper, we illustrate the experimentation scenario and simulation results. At the end, we conclude our research followed by list of references used in the paper.

RELATED WORK

Beloglazov and Buyya (2010, May) propose an idea of setting up upper and lower utilization thresholds to identify over and underloaded servers. If the host utilization exceeds the upper threshold, authors propose migrating some VMs from the host to reduce SLA violation (SLAV). On the contrary, if the utilization falls below the lower threshold, all VMs are to be migrated from the host and the host is to be switched off to save energy consumption. However, no specific technique or method has been proposed for calculating upper and lower threshold. Beloglazov and Buyya (2010) propose a technique based on statistical analysis of the VM utilization history for auto-adjustment of the utilization thresholds. Authors propose equation to calculate lower threshold while taking into consideration, factors such as host utilization, number of hosts, standard deviation of utilization and probability intervals. However, the same authors, (Beloglazov & Buyya, 2012) propose a simple approach of comparing relative host utilization for detection of underloaded host. This approach searches for the host with minimum utilization in a data center and tries to place all the VMs from this host to other host keeping them not overloaded. If this is successfully accomplished, the original host is switched to power saving mode

Push-based Prefetching in Remote Memory Sharing System

Rui Chu, Nong Xiao and Xicheng Lu (2011). *Cloud, Grid and High Performance Computing: Emerging Applications (pp. 269-283).*

www.igi-global.com/chapter/push-based-prefetching-remote-memory/54934?camid=4v1a

Supercomputers: A Philosophical Perspective

Jeremy Horne (2015). *Research and Applications in Global Supercomputing (pp. 378-410).*

www.igi-global.com/chapter/supercomputers/124353?camid=4v1a