Journal of Network and Computer Applications 71 (2016) 86–98

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

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

Review

Load balancing mechanisms and techniques in the cloud environments: Systematic literature review and future trends☆,☆☆

Alireza Sadeghi Milani, Nima Jafari Navimipour*

Department of Computer Engineering, Tabriz Branch, Islamic Azad University, Tabriz, Iran

ARTICLE INFO

Article history:
Received 12 February 2016
Received in revised form
2 May 2016
Accepted 4 June 2016
Available online 8 June 2016

Keywords:
Cloud computing
Load balancing
Task scheduling
Virtual machine

ABSTRACT

With cloud computing, new services in information technology (IT) emerge from the convergence of business and technology perspectives which furnish users access to IT resources anytime and anywhere using pay-per-use fashion. Therefore, it should supply eminent functioning gain to the user and simultaneously ought to be advantageous for the cloud service provider. To accomplish this goal, many challenges have to be faced, where load balancing is one of them. The optimal selection of a resource for a particular job does not mean that the selected resource persists optimized for the whole execution of the job. The resource overloading/under-loading must be avoided which could be gained by appropriate load balancing mechanisms. However, to the best of our knowledge, despite the importance of load balancing techniques and mechanisms, there is not any comprehensive and systematic review about studying and analyzing its important techniques. Hence, this paper presents a systematic literature review of the existing load balancing techniques proposed so far. Detailed classifications have also been included based on different parameters which are relying upon the analysis of the existing techniques. Also, the advantages and disadvantages associated with several load balancing algorithms have been discussed and the important challenges of these algorithms are addressed so that more efficient load balancing techniques can be developed in future.

Contents

☆Neither the entire paper nor any part of its content has been published or has been accepted for publication elsewhere.
☆☆It has not been submitted to any other journal.
* Corresponding author.
E-mail address: jafari@iaut.ac.ir (N.J. Navimipour).

http://dx.doi.org/10.1016/j.jnca.2016.06.003
1084-8045/© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

Nowadays, with the rapid extension of the IT-based systems (Navimipour and Soltani, 2016; Zareie and Jafari Navimipour, 2016), many distribution systems such as the social networks (Sharif et al., 2013; Mohammad Aghdam and Jafari Navimipour, 2016), grid computing Khanli and Razavi, 2008; Navimipour and Khanli, 2008; Navimipour et al., 2014; Souri and Navimipour, 2014), cloud computing (Navimipour, 2015a, 2015b; Navimipour and Milani, 2015a, 2015b; Asghari and Navimipour, 2016; Chiregi and Jafari Navimipour, 2016; Milani and Navimipour, 2016), peer-to-peer computing (Navimipour and Milani, 2014), wireless networks (Navimipour and Rahmani, 2009; Jafari Navimipour, 2011; Jafari Navimipour and Es-Hagi, 2011; Navimipour, 2011; Navimipour et al., 2012), Expert Cloud (Jafari Navimipour et al., 2014; Navimipour, 2015a, 2015b) and MapReduce (Navimipour and Khezr, 2015) facilitate the data transfer and resource sharing (Navimipour and Milani, 2015a, 2015b, Navimipour and Zareie, 2015). Among them, cloud computing as a new concept to represent the cooperation among multiple computers and services via a network provides many powerful on-demand services to the users. Many new research and studies have been proposed in order to realize the concept of cloud computing, however, the fundamental idea of it is derived primarily from distributed computing and grid computing (Cho et al., 2014). The ordinary applications have been delivered online by cloud computing providers, which could access by a web browser as the software and data are stored on servers. Users no longer need the knowledge of the technology of the cloud that supports their computing requirements. Cloud computing provides a new supplement, delivery, and consumption model for IT services based on the Internet. It generally involves the provision of dynamically scalable and frequently virtualized resources as a service over the Internet (Chiregi and Navimipour, 2016). It is a consequence of the ease-of-access to remote computing sites provided by the Internet (Li, 2012). There are several types of cloud services available such as Software as a Service (SaaS) (Buyya et al., 2010; Liu et al., 2011; Abrishami and Naghibzadeh, 2012; Celesti et al., 2012; Jose Moura, 2015), Platform as a Service (PaaS) (Marston et al., 2011; Voorsluys et al., 2011; Beloglazov et al., 2012; Jose Moura, 2015), Infrastructure as a Service (IaaS) (Buyya et al., 2010; Chang, 2015; Chou, 2015; Jose Moura, 2015) and Expert as a Service (EaaS) (Navin et al., 2014; Ashouraie et al., 2015; Jafari Navimipour, 2015; Navimipour et al. 2015a; Navimipour et al. 2015b). Furthermore, to the best suit of requirements of an application, the elasticity, flexibility and scalability have been offered to acquire or release resources varying configuration (Banerjee et al., 2015).

A workload or task abundantly overloads a resource in a computing environment. Therefore, the workload ought to be migrated to another resource. Consequently, three common operations need to be performed including load balancing, which checks the load on resources, resource discovery, which discovers another suitable resource and workload migration, which transfers the workload to the selected resource. These operations are taken over by three separate units, commonly called load balancing, resource discovery and process migration units, respectively. Evidently, a better performance could be obtained in the case of reduction of these operations (Arab and Sharifi, 2014). Load balancing algorithms offer possibilities for increasing the performance of large-scale computing systems and applications since they are designed to redistribute the workloads over the components of the computing system in a way that guarantees to minimize response time, maximizing resource utilization and throughput, and avoiding the overload possibility (Daraghmi and Yuan, 2015). To make better utilize of resources, an efficient load balancing solution needs to potentially reduce the resource over-provisioning (Nakai et al., 2014). There are several models and techniques that offer efficient scheduling and load balancing such as statics and dynamics. Static mechanisms require prior knowledge of the environment and applications requirements. However, since the applications start executing, these models have no way to adapt to changes in the environment or requirements. In contrast, in dynamic mechanisms the load balancer monitors the environment and application requirements during run-time and attempts to make adjustments to redistribute the tasks and adjust the load as necessary (Mohamed et al., 2013a, 2013b).

Nevertheless, to the best of our knowledge, despite the importance of load balancing mechanisms in cloud environments, there is not any detailed and comprehensive systematic review of these mechanisms. Therefore, the purpose of this paper is to survey existing techniques, compares the differences between mentioned mechanisms, describes several popular load balancing mechanisms and outlines the types of challenges that could be addressed. We divided most of the introduced load balancing algorithms into two main categories, static and dynamic. To the best of our knowledge, this survey represents the first attempt to systematically examine load balancing with a specific focus on cloud computing. Briefly, the contributions of this paper are as follows:

- providing an overview of existing challenges in a range of problem domains associated with cloud computing that can be addressed using load balancing
- providing a systematic overview of the existing techniques for load balancing, and the manner in which these have been applied to cloud computing
- exploring the future challenges for cloud computing and the role that load balancing can play
- outlining the key areas where future research can improve the use of load balancing techniques in cloud computing

The rest of this paper is structured as follows. Section 2 discusses some related work. The research methodology is provided in Section 3. Section 4 discusses load balancing mechanisms in cloud computing and categorizes them, also presents the taxonomy and comparison of selected mechanisms. Section 5 maps out some validity threats. Section 6 discusses open issues. Finally, Section 7 concludes this paper.

2. Related work

Many types of research have been done in the field of cloud computing and general challenges including scheduling, resource provisioning and load balancing and etc. However, there is a little comprehensive research about cloud load balancing has been done yet. In this section, we refer to some papers that there are in the field of load balancing in cloud computing.

One of the significant surveys of the load balancing and job