Accepted Manuscript

LBBSRT: An efficient SDN load balancing scheme based on server response time

Hong Zhong, Yaming Fang, Jie Cui

PII: S0167-739X(16)30372-7
DOI: http://dx.doi.org/10.1016/j.future.2016.10.001
Reference: FUTURE 3173

To appear in: *Future Generation Computer Systems*

Received date: 22 May 2016
Revised date: 5 September 2016
Accepted date: 1 October 2016

Please cite this article as: H. Zhong, Y. Fang, J. Cui, LBBSRT: An efficient SDN load balancing scheme based on server response time, *Future Generation Computer Systems* (2016), http://dx.doi.org/10.1016/j.future.2016.10.001

LBBSRT: An Efficient SDN Load Balancing Scheme Based on Server Response Time

Hong Zhong, Yaming Fang, Jie Cui*
School of Computer Science and Technology, Anhui University, Hefei, 230039, China
[e-mail: cuijie@mail.ustc.edu.cn]
***Corresponding author: Jie Cui**

Abstract—The response time is the most important factor determining user experiences in the service provision model involving server clusters. However, traditional server cluster load balancing scheme are limited by the hardware conditions, and cannot completely exploit the server response times for load balancing. In order to effectively resolve the traditional load balancing schemes, we propose a load balancing scheme based on server response times by using the advantage of SDN flexibility, named LBBSRT. Using the real-time response time of each server measured by the controller for load balancing, we process user requests by obtaining an evenly balanced server loads. Simulation experiments show that our scheme exhibits a better load balancing effect and process requests with a minimum average server response times. In addition, our scheme is easy to implement, and exhibits good scalability and low cost characteristics.

***Index Terms*—**SDN, OpenFlow, load balancing, server response time

1 INTRODUCTION

Achieving optimum load balancing is of significant importance whilst combating network overhead issues in any distributed processing architectures. Service availability is paramount in measuring end user satisfaction[1], which is heavily impacted by the level of achievable load balancing among the process clusters. In general, a well-balanced load in the network helps to optimize the utilization of the available resource by the ways of maximizing the throughput, minimizing the response time, and avoiding overloading resources in the network [2]. For the purposes of alleviating heavy-traffic network flux and reducing the risk of single server becoming the main overhead contributor, many datacentres adopt dedicated hardware resources to achieve load balancing whilst supporting a large number of users [3]. However, the increasing costs and technical complications in the deployment of such hardware systems often require human intervention to ensure consistent functioning of such strategies [4].

Software-Defined networking (SDN) is one of the notable forms of computer networking [5][6], facilitating a simple and conveniently maneuverable network flow control method requiring minimal investment costs whilst availing maximum benefits for a massive number of users. SDN controls the data transportation by deploying the network switches as a software implementation, whereby a flow table lookup operation will be carried out whenever a data flow arrives at the switches. Flow tables [7] ([Header: Counters: Actions]) are widely used in SDN. The headers and counters of the flow table are updated accordingly whenever actions relevant to flow changes are