Analysis & Minimization of the Effect of Delay on Load Balancing for Efficient Web Server Queueing Model

Harikesh Singh, Department of Computer Science & Engineering, Jaypee University of Engineering & Technology, Guna, India

Shishir Kumar, Department of Computer Science & Engineering, Jaypee University of Engineering & Technology, Guna, India

ABSTRACT

Load balancing applications introduce delays due to load relocation among various web servers and depend upon the design of balancing algorithms and resources required to share in the large and wide applications. The performance of web servers depends upon the efficient sharing of the resources and it can be evaluated by the overall task completion time of the tasks based on the load balancing algorithm. Each load balancing algorithm introduces delay in the task allocation among the web servers, but still improved the performance of web servers dynamically. As a result, the queue-length of web server and average waiting time of tasks decreases with load balancing instants based on zero, deterministic, and random types of delay. In this paper, the effects of delay due to load balancing have been analyzed based on the factors: average queue-length and average waiting time of tasks. In the proposed Ratio Factor Based Delay Model (RFBDM), the above factors are minimized and improved the functioning of the web server system based on the average task completion time of each web server node. Based on the ratio of average task completion time, the average queue-length and average waiting time of the tasks allocated to the web server have been analyzed and simulated with Monte-Carlo simulation. The results of simulation have shown that the effects of delays in terms of average queue-length and average waiting time using proposed model have minimized in comparison to existing delay models of the web servers.

Keywords: Average Queue-Length, Average Waiting Time, Delay, Load Balancing, Queueing Analysis

1. INTRODUCTION

The performance of web applications has affected due to several transactions of data, such as e-commerce transactions. Each access requires around 500 to 1400 ms for establishing the connection and downloading the web page. This transaction time increases drastically as any multimedia web image has accessed, so the response time of a simple web transactions may be required 3-6 s. Some classic online transactions may have taken 2-3 min and such

DOI: 10.4018/ijsda.2014100101

web performances are undesirable in the current communication system. These communication delays are happening under the different setup of web servers and affect the load balancing mechanism of the web servers (Bhargava, 2001).

A distributed computing systems used widely to improve the performance and resource sharing of web applications. Some hysterical task arrivals overloaded the server nodes even other nodes are idle. An approach of adaptive load sharing for queue control becomes useful to achieve optimal or near-optimal efficiency and performance. Several adaptive load sharing algorithms for heterogeneous distributed computing systems has performed for the delay analysis while transferring the tasks from one node to another, and validated (Kabalan et al., 2002).

In a web server system, the effects of the delay has analyzed by the linear model suggested by Abdallah et al. (2003). The load balancing process used the processing time while transferring the tasks from one node to another. The processing time in the network and bandwidth are the factors which are commonly used for taking the benefits of uniform load distribution among the nodes to decrease the overall processing time (Abdallah et al., 2003). There are several types of delay introduced by the network based on the factors such as availability of the network and processing time of the software, etc.

The process of load balancing on the web servers enhances its service capability and introduces the delays throughout the process of load migration from heavily loaded servers to lightly loaded servers. These delays may be categorized as deterministic or random depending upon the network settings. Randomness in delay is one of the significant issues in the load balancing system. The exact values of the load balancing factors cannot be easily calculated because of the random delay, as a result the performance of the load balancing approaches designed for dedicated communication links and systems gets affected (Birdwell et al., 2004; Chaisson et al., 2005; Hayat et al., 2004).

A dynamic time delay model for load balancing was proposed in distributed heterogeneous system by Hayat et al. (2004). But how to compute the load-transfer delay and the gain coefficient were not given, so it was impossible to quantify analyze the performance of a distributed system. Jie et al. (2004) have designed an improved dynamic load balancing model with load-transfer delay for Local Area Network (LAN), and the gain coefficient defined by a number of simulation test to meet the requirements of LAN. The simulation results of Jie et al. (2004) saved the resources and also realized a reliable and efficient operation of the system and the improved algorithm can be used to the distributed system in the LAN.

Several numbers of computations can be performed on multiple web server nodes using the dynamic load balancing model as a nonlinear time delay system (Dhakal, 2003; Dhakal et al., 2005; Dhakal et al., 2007). If the model is dependable, conserved and supportive then the load balancing takes place in such a way so that it can neither generate nor misplace the tasks. Each load balancing system generates the random delays in a cluster of web server nodes and predicts various effects of random delays on the performance of load balancing approaches (Dhakal et al., 2005; Tang et al., 2004).

In this paper, an effective mathematical model, Ratio Factor Based Delay Model (RFBDM) for heterogeneous environments has been proposed and implemented by improvement of the models proposed by Birdwell et al. (2004) and Hayat et al. (2004). Through this model the average queue-length and average waiting time of load balancing tasks have been analyzed considering the effects of delays commenced by transferring the loads among web servers. The proposed model covers the development of the random queue-length of each web server's node in presence of delay for an effective load balancing approach of web servers. The proposed mathematical model has been simulated using Monte-Carlo simulation for average queue-length and average waiting time for the tasks assigned on the web servers and simulation results have been compared with