Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computers & Operations Research 32 (2005) 2129–2145
computers & operations research
www.elsevier.com/locate/dsw

Determination of the parameters in the dynamic weighted Round-Robin method for network load balancing

Der-Chiang Li[a,*], Chihsen Wu[a], Fengming M. Chang[b]

[a] *Department of Industrial and Information Management, National Cheng Kung University, Tainan, Taiwan*
[b] *Department of Industrial Engineering and Management, Tungfang Institute of Technology, Kaohsiung, Taiwan*

Abstract

The purpose of this paper is to determine the values of the parameters in a new method (Dynamic Weighted Round-Robin, DWRR) developed for solving Internet traffic jam problems. Using the traditional Round-Robin (RR) method as a base, DWRR was developed to efficiently control loads in a multiple-link network. Unlike least-load algorithm, DWRR does not need to trace system loads continually, but achieves a far better load balancing than RR does. Mathematical functions are developed for predicting the optimal time interval of detection of line loads in this method, while the concept of variance in statistics is used as the criterion for evaluating the load balance level. A couple of related coefficients have also been determined by analyzing the simulation data. A centralized gateway with a multi-link-load-balancer is modeled for explaining the proposed algorithm. In addition, both theoretical and practical approaches are provided in this paper, along with performance comparisons between them. The results obtained from the computational experiments show that DWRR achieves a superior network loads balancing.

Keywords: Dynamic weighted Round-Robin; Load balancing; Round-Robin; Network; Internet

1. Introduction

Applications for the Internet are growing explosively, causing enormous of network loads. This high growth rate, however, leads to a decrease in Quality of Service (QoS) and performances of networks, both in network line bandwidth allocation and in server responding time. The solution currently widely used for this kind of Internet traffic problem is to build up a multiple-Internet-link or multiple-server distributed system to distribute the loads. Fig. 1 illustrates a multiple-Internet-link structure with three lines connecting outward to the Internet. Since a system with balanced loads

* Corresponding author. Tel.: +886-6-2757575; fax: +886-6-2341171.
E-mail addresses: lidc@mail.ncku.edu.tw (D.-C. Li), wucs@mail.ncku.edu.tw (C. Wu), bit@anet.net.tw (F.M. Chang).

0305-0548/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cor.2004.02.004

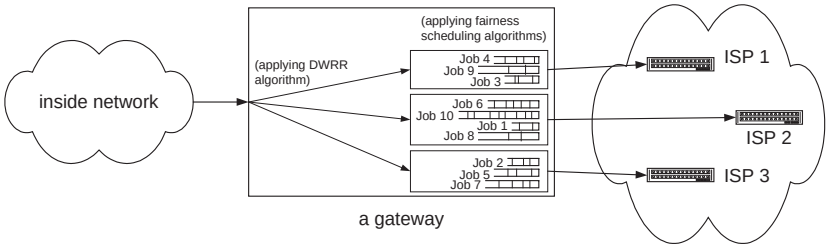

Fig. 1. A gateway for load balancing function.

among its multiple links or servers yields better system performance, this study focuses on finding a load balancing method to further improve its performance efficiency.

Quite a few load balancing techniques for the Internet have already been proposed [1,2]. Some of them solved the problem by adding extra hardware to the system, such as Round-Robin DNS [2], others solved the problem by applying software to the controller. The former achieve approximate balancing but increases extra cost incurred through connecting additional systems, and are used in multiple web site servers. The latter are popularly applied in network devices using control algorithms directly and generally obtain fairly good results. Traditional control algorithms for load balancing include Random, Round-Robin (RR) [3], Weighted Round-Robin (WRR) [4], Least Load, Least Connections, and Fastest Response algorithms [5]. For the job of assigning loads, Random algorithm selects links or servers randomly, regardless of the load on each line. This can leave the system unsteady. The performance of Random algorithm for load balancing is worse than that of RR [6]. Round-Robin algorithm is the simplest one, in which links merely serve the system in turn. Logically, in a system with n links, if the current outgoing traffic goes through $link_i$, $1 \leqslant i \leqslant n$, then the next service line is $link_{(i+1) \bmod n}$. RR was first used in network problems by Nagle [3] and compared with other algorithms in a lot of other research [6–10]. The benefit of RR lies in the omission of tracing the links, for the tracing work is known to add extra load to the system, therefore detracting from the performance. WRR has the same scheme as RR does, except that a fractional weight is given to each link according to the link's performance. Least Load algorithm detects the load of each server or link and allocates the new load to the least loaded one. Least Connections algorithm tracks the number of connected users on each server or link and assigns new traffic to the one with the fewest connections. Fast Response algorithm keeps track of the response time of each server and balances the loads by choosing the one with the fastest response one. Fastest Response algorithm can only be used on servers. Both Least Load and Least Connections algorithms continually detect, measure, and rank each link's utilization as the basis for line selection. For systems using these two algorithms, the load balancing devices are costly and can become a bottleneck for the whole system. It's suitable to apply RR to a system only when the servers have fixed performance levels, since they have no sense of line congestion.

There are variants of RR-related methods for flow controlling, but concentrate on fairness scheduling for lines: Deficit Round-Robin algorithm (DRR) [9] is proposed on the basis of RR. It uses a quantum concept to control the speed of packets traveling among flows. To do that, it continually traces the states of flows for measuring deficits in the lines. Bit-by-bit round-robin (BRR) [11] tries to perform Round-Robin algorithm bit by bit, but it's almost impossible to execute in a high speed