Information Sciences 199 (2012) 58–77

Contents lists available at SciVerse ScienceDirect

Information Sciences

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

Adaptive admission control algorithm in a QoS-aware Web system

Katja Gilly [a,*], Carlos Juiz [b], Nigel Thomas [c], Ramon Puigjaner [b]

[a] *Departamento de Física y Arquitectura de Computadores, Miguel Hernández University, Elche, Spain*
[b] *Departament de Ciències Matemàtiques i Informàtica, University of Balearic Islands, Palma de Mallorca, Spain*
[c] *School of Computing Science, Newcastle University, Newcastle upon Tyne, UK*

ARTICLE INFO

Article history:
Received 10 October 2009
Received in revised form 7 February 2012
Accepted 13 February 2012
Available online 28 February 2012

Keywords:
Admission control
Resource allocation
Load balancing
Overloaded Web server
Network monitoring
Performance evaluation

ABSTRACT

Internet traffic tends to show significant growth of demand at certain times of the day, or in response to special events. The consequence of these traffic peaks is that Web systems that are responding to user demands are congested due to their inability to serve a large volume of requests. The case for admission control in these situations is even stronger when Quality of Service (QoS) is considered as a primary objective in the Web system. In this work, we address two issues: on one hand, we consider and compare five throughput predictors to be used in a Web system in order to track its performance and, on the other hand, we propose a QoS-aware admission control and load balancing algorithm that prevents the Web system from sudden overload. The admission control algorithm is based on a resource allocation scheme that includes a throughput predictor. In order to obtain a low overhead, the monitoring of traffic arriving at the Web system is performed following an adaptive time slot scheduling based on the burstiness factor that we defined in previous work. Results show the benefits of our adaptive time slot scheduling compared to a fixed time scheduling. A discussion of the results of the five throughput predictors and the admission control algorithm is provided. We also compare the performance of our algorithm with Intelligent Queue-based Request Dispatcher (IQRD). The algorithm is designed to be included in a Web system composed by a set of Web servers distributed locally, which can also form part of a wider geographically distributed load balancing architecture.

1. Introduction

For Web services to satisfy the demands of end users, they must continue to offer a good level of performance, even in the fact of erratic and unpredictable demands. It has been widely observed that Internet traffic is self similar and that sudden bursts of packets can reach a point in a network infrastructure that offers Web services. This affects the performance of the system if it is not able to process that increase in the demand. High variance in incoming traffic and service time distributions can collapse the system in few seconds; therefore it is necessary to control these features by an adaptive algorithm. We propose an adaptive admission control algorithm that prevents the system from a sudden overload by predicting the throughput of the Web servers. Five different throughput predictors are considered in this work.

The problem of allocating resources to a Web System that considers QoS is also addressed in this work. We have considered a Web system that is composed of a cluster of Web servers, as shown in Fig. 1. The algorithm includes adaptive resource

Abbreviations: App/DB, application/database; AS, autonomous system; CPU, central processing unit; HTTP, hypertext transfer protocol; IQRD, intelligent queue-based request dispatcher; LMS, least mean square; NLMS, normalised least mean square; RR, round robin; SLA, service level agreement; TCP, transmission control protocol; OS, operative system; QoS, quality of service; URI, uniform resource identifier.

* Corresponding author. Tel.: +34 96665 8565; fax: +34 96665 8814.
E-mail addresses: katya@umh.es (K. Gilly), cjuiz@uib.es (C. Juiz), Nigel.Thomas@ncl.ac.uk (N. Thomas), putxi@uib.es (R. Puigjaner).

0020-0255/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2012.02.018

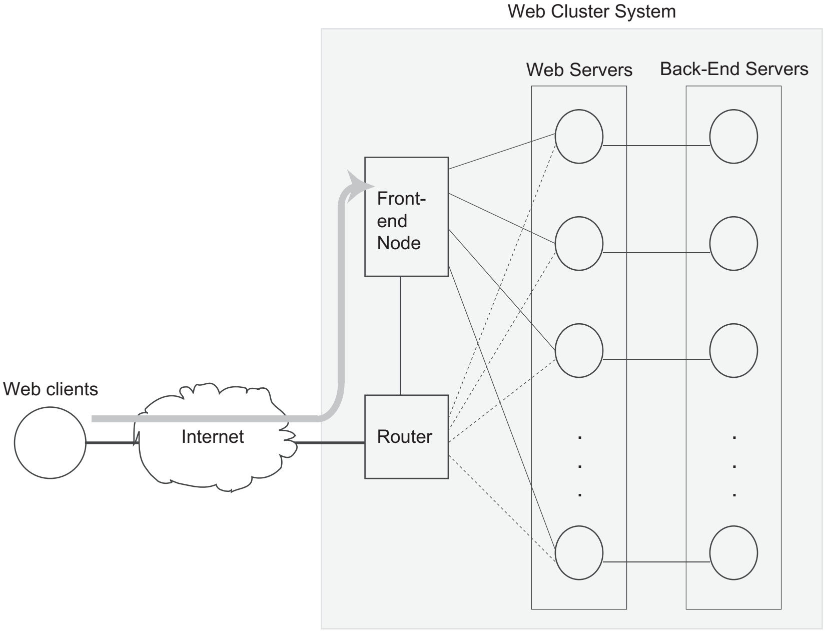

Fig. 1. The Web architecture is made up of several mirrored Web servers and their corresponding database servers. The model architecture is one-way, which means that the incoming HTTP requests go through the front-end node but their HTTP responses use a different way to prevent a system bottleneck in this node.

allocation for each server, considering the values of certain monitored performance metrics that allow the algorithm to learn the current state of the Web system. Predictions of throughput and utilisation are computed based on these metrics and, in case future congestion in the Web system is forecast, the admission control part of the algorithm is initiated. Hence, overload situations can be addressed by this algorithm to guarantee satisfactory performance of the system by controlling the utilisation level of the servers within the cluster.

An important contribution of our work is the adaptive overhead our solution introduces in the system. It is critical to avoid checking the system continuously, e.g. at each incoming request, because this produces an enormous overhead in the front end of the system. It is also risky to check the system during fixed intervals because a sudden increase may not be detected until the system is already overloaded. We have analysed the most important related work in order to learn how other authors handle or control overhead, and we have observed that very few works propose methods to reduce the overhead.

The following sections of this paper are organised as follows:

- Section 2 describes related studies on admission control and includes a table that sums up the characteristics of the proposals we are principally interested in.
- The steps we take in order to obtain a low overhead are detailed in Section 3.
- Section 4 introduces an overview of the algorithm that includes the system architecture details, the QoS and the metrics considered by the algorithm.
- The throughput predictors we propose to be included in the algorithm are described in Section 5.
- The resource allocation strategy and load balancing policy are in Section 6.
- In Section 7, we have included a description about the workload used in the simulations and the load balancing policy applied.
- Simulation results of the comparison of the different throughput predictors are included in Section 8.
- In order to compare our adaptive time slot scheduling to a fixed time slot scheduling, we have run a set of simulations whose results are included in Section 9.
- We present some results from an implementation of Intelligent Queue-based Request Dispatcher, the algorithm proposed by Sharifian et al. in [41], and comment on the performance differences detected in comparison to our algorithm in Section 10.
- Finally, we discuss some concluding points and the open problems.

2. Related work

The problem of the admission control in a Web server has been widely addressed in literature. In this Section, we introduce the most significant Web admission control related algorithms and organise them in Table 1, where we include a summary of the characteristics we consider the most important from the cited admission control proposals.