

MARKOV PROCESS ANALYSIS FOR WEB CACHING WITH MULTIPLE POPULARITY LEVELS

Wenying Feng

Computer Science and Mathematics
Trent University, Peterborough, Ontario K9J 7B8, Canada

Abstract. In this paper, a Markov process approach to the study of Web caching systems that implement multiple popularity levels is presented. Equilibrium analysis is applied to derive formulas of calculating the mean and the upper and lower bounds of cache hit rates. Data collected from the simulations are compared with the theoretical results.

Keywords. Web cache, Markov process, birth-death model, QoS, LRU, equilibrium states.

AMS subject classification: 60K20, 68M20.

1 Introduction

Today, the World Wide Web plays a significant role in the day-to-day life of modern society. In order to satisfy the high demands of these contemporary clients (users), it has been necessary to improve the performance of Web servers through applications of caching techniques. Consequently, various cache replacement algorithms have been proposed, and applied, to serve this universal purpose of meeting the growing expectations of the modern day clients [2].

In a system that is integrated with a single server and a cache, information is stored at the server as *pages*. Requests for pages from the clients arrive at the cache. If the requested page is found in the cache, it is simply returned to the client and the cache is counted to have a *hit*. In the case of a *miss*, the page is retrieved from the server and sent to the client, a copy of it is stored locally in the cache. Cache replacement algorithms are used to decide which *victim page* should be overwritten, or *kicked out*.

Most commonly used cache replacement algorithms include the LRU (Least Recently Used), LFU (Least Frequently Used), random, size-based and function-based policies [1] [10]. In [9], Web pages are classified as QoS (Quality of Service) classes. Caching schemes aimed at increasing the hit rate for pages in the higher level classes are discussed. Later, in [3], two specified criteria for QoS classes, popularity and priority, are individually examined. QoS caching schemes are also studied in [7] and recently in [11]. As the most common methods, computer simulations and experiments have been utilized in various investigations [3] [7] [9] [11]. Although much information can be obtained from simulation programs, the random