Performance Evaluation 72 (2014) 16–36

Contents lists available at ScienceDirect

Performance Evaluation

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

Performance evaluation of the random replacement policy for networks of caches

Massimo Gallo [a], Bruno Kauffmann [b], Luca Muscariello [b,*], Alain Simonian [b], Christian Tanguy [b]

[a] Alcatel-Lucent Bell Labs, Centre de Villarceaux, Route de Villejust, 91620 Nozay, France
[b] Orange Labs, 38–40 rue du Gnral Leclerc, 92794 Issy-Les-Moulineaux, France

ARTICLE INFO

Article history:
Received 6 November 2012
Received in revised form 23 August 2013
Accepted 19 October 2013
Available online 7 November 2013

Keywords:
Caching
Networks of caches
Performance analysis

ABSTRACT

The overall performance of content distribution networks as well as recently proposed information-centric networks rely on both memory and bandwidth capacities. The hit ratio is the key performance indicator which captures the bandwidth/memory tradeoff for a given global performance.

This paper focuses on the estimation of the hit ratio in a network of caches that employ the Random replacement policy (RND). Assuming that requests are independent and identically distributed, general expressions of miss probabilities for a single RND cache are provided as well as exact results for specific popularity distributions (such results also hold for the FIFO replacement policy). Moreover, for any Zipf popularity distribution with exponent $\alpha > 1$, we obtain asymptotic equivalents for the miss probability in the case of large cache size.

We extend the analysis to networks of RND caches, when the topology is either a line or a homogeneous tree. In that case, approximations for miss probabilities across the network are derived by neglecting time correlations between miss events at any node; the obtained results are compared to the same network using the Least-Recently-Used discipline, already addressed in the literature. We further analyze the case of a mixed tandem cache network where the two nodes employ either Random or Least-Recently-Used policies. In all scenarios, asymptotic formulas and approximations are extensively compared to simulation results and shown to be very accurate. Finally, our results enable us to propose recommendations for cache replacement disciplines in a network dedicated to content distribution.

1. Introduction

Communication networks use an ever increasing amount of data storage to cache information in the aim of performance improvement. Data caching consists in temporarily storing pieces of data into a memory, so as to directly provide the data upon possible forthcoming requests. The performance gain stems from the Round-Trip-Time reduction and the increase in network capacity when the cache is located downstream a bandwidth bottleneck, *e.g.*, a communication link with limited bandwidth or a shared bus in a network of chips.

As a major application, the increasing amount of content delivered to Internet users has pushed the use of Web caching into communication models based on distributed caching such as Content Delivery Networks (CDNs) or Peer-To-Peer (P2P)

* Corresponding author. Tel.: +33 1 45 29 60 37; fax: +33 1 45 29 60 69.
E-mail addresses: massimo.gallo@alcatel-lucent.com (M. Gallo), bruno.kauffmann@orange.com (B. Kauffmann), luca.muscariello@orange.com (L. Muscariello), alain.simonian@orange.com (A. Simonian), christian.tanguy@orange.com (C. Tanguy).

0166-5316/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.peva.2013.10.004

Networks. Additionally, new information-centric network architectures [1–3] have been recently proposed, that include built-in network storage as a central feature of the underlying communication model. Content storage then becomes a primary resource in such networks, aiming at minimizing content delivery time under an ever increasing demand that cannot simply be satisfied by increasing link bandwidth. On the economic side, the use of network storage to bypass bandwidth bottlenecks appears cost effective as memory turns out to be cheaper than transmission capacity.

One of the fundamental operations of a cache is defined by its replacement policy which determines the object to be removed from the cache when the latter is full. Many replacement policies are based on content popularity, with significant cost for managing the sorted lists. This is the case, in particular, for the Least Frequently Used (LFU) policy and more sophisticated variants of it. On the contrary, Most Recently Used (MRU), Least Recently Used (LRU), First-In-First-Out (FIFO) and Random (RND) policies have the compelling feature to replace cached objects with constant delay. In-network storage, as envisaged in the new architectures mentioned above, may require packet-level caching at line rate; current routers running complex replacement policies might not, however, sustain such high rates [4]. In this framework, RND or FIFO policies can therefore be seen as presenting the least possible complexity; in fact, the latter require less memory access per packet than LRU or MRU, and it has been shown [4] that this advantage is critical for sustaining high-speed caching with current memory technology.

In this paper, we address performance issues of caching networks running the RND replacement policy; our analysis also holds for the FIFO policy whose performance is known to be equivalent to that of RND. We mainly focus on the analytical characterization of the miss probabilities under the Independent Reference Model (IRM) assumptions when the number of available objects is infinite. Exploiting the Markovian properties of the cache occupancy and its associated product-form distribution, we first express the miss rate as a ratio of normalizing constants. This enables us to provide exact formulae for the miss rate in case of either geometric or specific Zipf content popularity distributions. On the basis of Large Deviations results for discrete probability distributions, Proposition 3.9 then asserts our main result: when the popularity distribution follows a general power-law with decay exponent $\alpha > 1$, the miss probability is asymptotic to $A\rho_\alpha/C^{\alpha-1}$ for large cache size C, where constants A and ρ_α depend on α only. In Proposition 3.10, we extend that result to miss probabilities conditioned by the object popularity rank.

A second major contribution of the paper is given by Proposition 5.1, where we evaluate the performance of networks of caches under the RND policy, for both linear and homogeneous tree networks and asymptotically Zipf popularity distributions. An approximate closed formula for the miss probability across the network is provided and compared to corresponding estimates for LRU cache networks. The analysis is also extended to the mixed tandem cache network where one cache employs LRU and the other uses RND.

The specific focus on Zipf distributions or, more generally, power-law distributions is motivated by numerous studies on Internet object popularity, starting from the late 90s experiments on World Wide Web documents [5,6] to the content stored in enterprises media servers [7,8] and recent studies on Internet media content [9,10]. While other content popularity distributions might be considered, we do not provide here a complete review of the literature on Internet content popularity characterization; the above references confirm the relevance of Zipf distributions for studying caching performance.

The remaining part of the paper is organized as follows. Section 2 presents related work on the analytic performance evaluation of caching systems. Section 3 analyzes the RND cache replacement policy and its comparison to LRU for a single cache; these analytic results are compared to exact numerical and simulation results in Section 4. Section 5 reports the approximate analysis of the network of RND caches for two topologies, namely the line and the tree. Numerical and simulation results for the network case are reported in Section 6. Section 7 further evaluates the tandem cache system where one cache implements LRU and the other RND. Section 8 concludes the paper.

2. Related work

There is a significant body of work on caching systems and their associated replacement policies; we here only report the literature focusing on the analytic characterization of the performance of such systems.

The replacement policy most often analyzed is LRU whose performance is evaluated considering the move-to-front rule, consisting in putting the latest requested object in front of a list; a miss event for a LRU cache with finite size takes place when the position (also referred to as search cost) of an object in the list is larger than that size. Under the Independent Reference Model, [11] calculates the expected search cost and its variance for finite lists. An explicit formula is given in [12,13] for the probability distribution of that cost; such a formula is, however, impractical for numerical evaluation in case of large object population and large cache size. Integral representations obtained in [14,15] using the Laplace transform of the search cost function reduce the problem to numerical integration.

An asymptotic analysis of LRU miss probabilities for Zipf and Weibull popularity distributions is derived in [16] and provides simple closed formulas. Extensions to correlated requests are obtained in [17,18], showing that short-term correlation does not impact the asymptotic results derived in [16]; the case of variable object sizes is also considered in [19]. The average miss probability for a LRU cache when requests are issued according to a general, possibly non-stationary, stochastic point process is obtained in [20].

The analytic evaluation of the RND policy has first been initiated in [21] for a single cache where the miss probability is given a general expression for any popularity distribution. To the best of our knowledge, its application to specific popularity distributions has, however, not yet been envisaged together with its numerical tractability for large object population and