Performance Evaluation of the Random Replacement Policy for Networks of Caches

Massimo Gallo, Bruno Kauffmann, Luca Muscariello, Alain Simonian, Christian Tanguy
Orange Labs, France Telecom
firstname.lastname@orange.com

Categories and Subject Descriptors

C.2.1 [**Computer-Communication Networks**]: Network Architecture and Design; C.4 [**Performance of systems**]: Performance attributes

Keywords

Content Distribution Networks, Information-Centric Networking, Cache Replacement Policies, Asymptotic Analysis

1. INTRODUCTION

Caching is a key component for Content Distribution Networks and new Information-Centric Network architectures. In this paper, we address performance issues of caching networks running the RND replacement policy. We first prove that when the popularity distribution follows a general power-law with decay exponent $\alpha > 1$, the miss probability is asymptotic to $O(C^{1-\alpha})$ for large cache size C. We further evaluate network of caches under RND policy for homogeneous tree networks and extend the analysis to tandem cache networks where caches employ either LRU or RND policies. This paper is an extended abstract of the full version [2], where detailed results and proofs are provided.

2. SINGLE CACHE

Consider a cache memory of finite size which is offered requests for objects. When a request for an object cannot be satisfied (this will be called a *miss* event), the requested object is fetched from the repository server and cached locally at the expense of replacing some other object in the cache. The object replacement policy is assumed to follow the RND discipline, *i.e.*, whenever a miss occurs, the object to be replaced is chosen at random, uniformly among the objects present in the cache. Given the total number N of objects, the probability that object r, $1 \le r \le N$, is requested is defined by q_r, $1 \le r \le N$. Let $C \le N$ be the cache capacity and denote by $M(C)$ the stationary miss probability; finally, define $M_r(C)$ as the per-object miss probability, given that the requested object is precisely $r \in \{1, ..., N\}$. A general combinatorial expression of $M(C)$ has been given in ([3], Theorem 4) for any popularity distribution. We here provide asymptotics for probabilities $M(C)$ and $M_r(C)$ for large cache size C and with a Zipf popularity distribution $q_r = O(r^{-\alpha})$ for large r and exponent $\alpha > 1$.

PROPOSITION 2.1. *For a Zipf popularity distribution with exponent $\alpha > 1$, the miss probability $M(C)$ is asymptotic to*

$$M(C) \sim \rho_\alpha C q_C \tag{1}$$

for large C, with prefactor $\rho_\alpha = \left(\frac{\pi/\alpha}{\sin(\pi/\alpha)}\right)^\alpha$. Furthermore, given the object rank r, the per-object miss probability $M_r(C)$ is estimated by

$$M_r(C) \sim \frac{\rho_\alpha r^\alpha}{C^\alpha + \rho_\alpha r^\alpha}. \tag{2}$$

Fig. 1 depicts $M_r(C)$ as a function of the object rank r for both RND and LRU policies with fixed $C = 25$ and $\alpha = 1.7$. Numerical results confirm the asymptotic accuracy of estimate (2) for RND and the corresponding one for LRU policy [4] when compared to simulation. Besides, RND and LRU performance are very close for large enough r.

3. IN-NETWORK CACHE MODEL

We generalize the single-cache model and consider a *homogeneous tree* network, where all leaves are located at a common depth of the root, and the cache size of any node at a given level ℓ is equal to C_ℓ. Requests for content are routed towards the root of the network, until they experience a hit at some cache. Any request experiencing a miss at some cache of level k, $1 \le k \le \ell$, and an object hit at level $\ell + 1$ is copied backwards to all downstream caches on the request path. A request miss corresponds to a miss event at all levels. We assume that **(H)** *any cache considered in iso-*

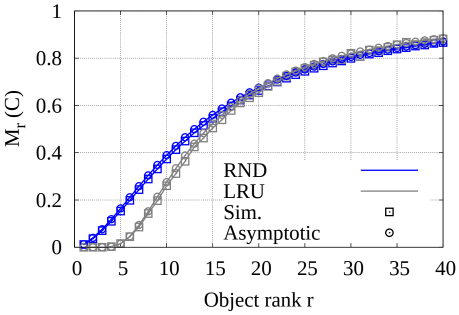

Figure 1: Asymptotic of $M_r(C)$ for a single cache with $C = 25$, $\alpha = 1.7$ for RND and LRU policies.

SIGMETRICS'12, June 11–15, 2012, London, England, UK.
ACM 978-1-4503-1097-0/12/06.

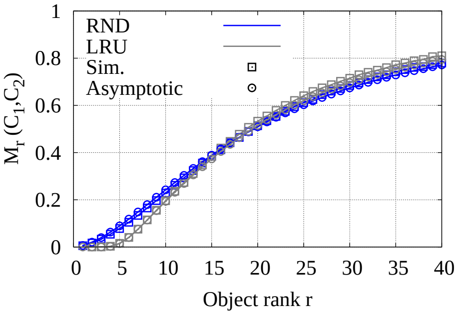

Figure 2: Asymptotics of $M_r(C_1, C_2)$ for RND and LRU policies compared to simulation for a two-level tree network with $C_1 = C_2 = 25$, $\alpha = 1.7$.

lation behaves as a single cache with IRM input produced by consecutive missed requests originated by its children nodes.

Proposition 3.1. *For the homogeneous tree network, suppose that the request process at the leaves is IRM with Zipf popularity distribution with exponent $\alpha > 1$, and that assumption **(H)** holds for all caches.*

For any level $\ell \in \{1, \dots, K\}$ and cache sizes $C_1, \dots, C_\ell$, let $M_r(C_1, \dots, C_\ell)$ denote the "global" miss probability for request r over all caches of a route through levels $1, \dots, \ell$. Then for large C_1, ..., C_ℓ,

$$M_r(C_1, \dots, C_\ell) \sim \frac{\rho_\alpha r^\alpha}{\rho_\alpha r^\alpha + \sum_{1 \le j \le \ell} C_j^\alpha}. \tag{3}$$

Fig. 2 reports miss probability $M_r(C_1, C_2)$ at the second level, *i.e.*, the probability to query an object of rank r at the repository server. We note that objects with small rank are slightly more frequently requested at the server when using RND rather than LRU, but RND is more favorable than LRU for objects with higher rank. We also be observe that the miss probability at the second level is very similar either using RND or LRU, with a slight advantage to LRU. We notice that the approximations calculated in Section 3 for RND and in [1] for LRU are very accurate.

4. MIXTURE OF RND AND LRU

This section addresses the case of a two-level tree network, where the RND replacement algorithm is used at one level while the LRU algorithm is used at the other. As in Section **3**, these results also hold in the case of a homogeneous tree.

Proposition 4.1. *For the two-level tree network, suppose the request process at cache level 1 is IRM with Zipf popularity distribution with exponent $\alpha > 1$ and that assumption **(H)** for cache level 2 holds.*
I) *When level 1(resp. level 2) uses the RND (resp. LRU) replacement policy, the global miss probability at level 2 is given by*

$$M_r(C_1, C_2) \sim \frac{\rho_\alpha r^\alpha}{\rho_\alpha r^\alpha + C_1^\alpha} \exp\left(-\frac{\rho_\alpha C_2^\alpha}{\alpha \lambda_\alpha \left(\rho_\alpha r^\alpha + C_1^\alpha\right)}\right) \tag{4}$$

for large cache sizes C_1, C_2 and constants ρ_α, λ_α.
II) *When level 1 (resp. level 2) uses the LRU (resp. RND) replacement policy, the global miss probability on level 2 is given by*

$$M_r(C_1, C_2) \sim \frac{\rho_\alpha r^\alpha}{\rho_\alpha r^\alpha \exp\left(\dfrac{C_1^\alpha}{\alpha \lambda_\alpha r^\alpha}\right) + C_2^\alpha}. \tag{5}$$

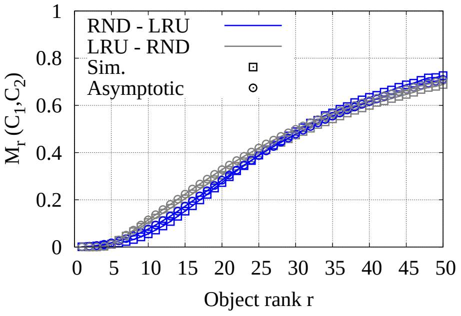

Figure 3: Asymptotic of $M_r(C_1, C_2)$ for a mixed tree cache networks with $C_1 = 25, C_2 = 50$, $\alpha = 1.7$.

In numerical experiments, we have simulated tree networks with 2 leaves with cache size C_1 and one root with cache size C_2. In Fig. 3, we observe that the total miss probability $M_r(C_1, C_2)$ in the two mixed tandem caches is similar, while the distribution of the objects across the two nodes varies considerably. Further numerical experiments suggest to prefer LRU at the first cache because it performs better in terms of miss probability, and to use RND at the second one in order to save significant processing time.

5. CONCLUSION

The performance of RND being reasonably close to that of LRU, RND is a good candidate for high-speed caching when the complexity of the replacement policy becomes critical. In the presence of a hierarchy of caches, caches at deep levels (i.e. access networks) typically serve a small number of requests per second which can be sustained by a cache running LRU, thus providing the best performance at the bottom level. Higher-level caches that serve many aggregated requests should use the RND policy, which yields similar performance while being less computationally expensive.

6. REFERENCES

[1] G. Carofiglio, M. Gallo, L. Muscariello, and D. Perino. Modeling data transfer in content-centric networking. In *Proc. of ITC23*, 2011.

[2] M. Gallo, B. Kauffmann, L. Muscariello, A. Simonian, and C. Tanguy. Performance evaluation of the random replacement policy for networks of caches (14 pages version). http://arxiv.org/abs/1202.4880, 2012.

[3] E. Gelenbe. A unified approach to the evaluation of a class of replacement algorithms. *IEEE Transactions on Computer*, 22(6):611–618, 1973.

[4] P. R. Jelenković and X. Kang. Characterizing the miss sequence of the LRU cache. *ACM SIGMETRICS Performance Evaluation Review*, 36:119–121, August 2008.