LRU Cache under Stationary Requests

Bo Jiang
College of Information and Computer Sciences
University of Massachusetts
Amherst MA 01003, USA
bjiang@cs.umass.edu

Philippe Nain
Inria Research Centre: Grenoble - Rhône-Alpes
69364 Lyon Cedex 07, France
philippe.nain@inria.fr

Don Towsley
College of Information and Computer Sciences
University of Massachusetts
Amherst MA 01003, USA
towsley@cs.umass.edu

ABSTRACT

In this paper we extend an approximation first proposed by Fagin [4] for the LRU cache under the independence reference model to systems where requests for different contents form independent stationary and ergodic processes. We show that this approximation becomes exact as the number of contents goes to infinity while maintaining the fraction of the contents that can populate the cache to be constant. Last, we provide results on the rate of convergence.

Keywords

Cache, LRU, Characteristic time approximation, Stationary request processes

1. INTRODUCTION

Caches form a key component of many computer networks and systems. Moreover they are becoming increasingly more important with the current development of new content-centric network architectures. A variety of cache replacement algorithms has been introduced and analyzed over the last few decades, most based on the least recently used algorithm (LRU). Considerable work has focused on analyzing these policies. However even the simple LRU policy defies exact analysis that leads to computationally tractable results. This has led instead to the development of accurate approximations, [4, 3, 2].

The most useful approximation for LRU was introduced by Fagin in [4] for the independent reference model (IRM). Briefly, Fagin introduced the concept of a characteristic time (our terminology) and showed asymptotically that the performance of LRU converges to that of a TTL cache with a timer set to the characteristic time. This work disappeared and several papers [2, 5] reintroduced this approximation. More recently, [6] extended the characteristic time (CCT) approximation to a setting where requests for distinct contents are independent and described as renewal processes. The accuracy of this approximation is supported by simulations but a theoretical basis is lacking. Providing a theoretical justification of this extended CCT approximation is the focus of this paper.

This research was sponsored by the U.S. ARL and the U.K. MoD under Agreement Number W911NF-16-3-0001 and by the NSF under grant NSF CNS-1413998.

The main contribution of this paper is to extend Fagin's results for LRU under IRM assumptions to the more general setting where requests for different content are independent of each other but requests to each content are described by a stationary and ergodic process. Based on these results, we develop a CCT approximation for the performance of an LRU cache. Furthermore, we develop bounds on the error introduced by the CCT approximation along with a convergence rate of the approximation to the asymptotic limit as the cache size and number of contents increase to infinity.

The rest of the paper is organized as follows. Section 2 presents our model of an LRU cache under a general request model. Section 3 presents the extension to Fagin's result to the case where requests for contents are described by independent stationary and ergodic processes. Section 4 investigates the rate of convergence of the LRU cache hit probability to the TTL cache hit probability. Last, concluding statements are provided in Section 5.

2. MODEL

We consider a cache of size C serving n unit size contents labelled $i = 1, \ldots, n$ where $C \in (0, n)$. Requests for the contents are described by n independent stationary and ergodic point processes $N_i := \{t_i(k), k \in \mathbb{Z}\}$, where $-\infty \leq \cdots < t_i(-1) < t_i(0) \leq 0 < t_i(1) < \cdots \leq \infty$ represent the successive request times to content $i = 1, \ldots, n$ having probability measure $\mathbb{P}$. Let $0 < \lambda_i < \infty$ denote the intensity of request process N_i, i.e., the long term average request rate for content i (see e.g. [1, Sections 1.1 and 1.6] for an introduction to stationary and ergodic point processes).

Let $\mathbb{P}_i^0$ be the Palm probability associated with the point process N_i (see e.g. [1, Eq. (1.2.1)]). In particular, $\mathbb{P}_i^0[t_i(0) = 0] = 1$. In other words, under $\mathbb{P}_i^0$ content i is requested at time $t = 0$. Define $G_i(x) = \mathbb{P}_i^0[t_i(1) \leq x]$, the cdf of the duration between two successive requests to content i under $\mathbb{P}_i^0$. It is known that $\mathbb{E}_i^0[t_i(1)] = 1/\lambda_i$ [1, Exercice 1.2.1], with $\mathbb{E}_i^0$ the expectation operator associated with $\mathbb{P}_i^0$.

Last, we define $\mathbb{P}^0$, the Palm probability associated with the point process $\{t(k), k \in \mathbb{Z}\}$, $-\infty \leq \cdots < t(-1) < t(0) \leq 0 < t(1) < \cdots \leq \infty$, resulting from the superposition of the n independent point processes $N_1, \ldots, N_n$. Under $\mathbb{P}^0$ a content is requested at $t = 0$ (since $\mathbb{P}^0[t(0) = 0] = 1$). Let $X_0 \in \{1, \ldots, n\}$ denote this content. We denote by $\mathbb{E}^0$ the expectation operator associated with $\mathbb{P}^0$. It is known that (see e.g. [1, Section 1.4.2])

$$\mathbb{P}^0[X_0 = i] = \frac{\lambda_i}{\Lambda^{(n)}} := p_i^{(n)}, \tag{1}$$

with $\Lambda^{(n)} := \sum_{i=1}^{n} \lambda_i$.

For any cdf F with support in $[0, \infty)$, let

$$\hat{F}(x) = \frac{1}{\mathbb{E}F} \int_0^x \bar{F}(y)dy, \tag{2}$$

where $\bar{F} = 1 - F$ is the ccdf, and $\mathbb{E}F = \int_0^\infty \bar{F}(y)dy$ is the mean. It is well-known that (see e.g. [1, Section 1.3.4])

$$\mathbb{P}[-t_i(0) \leq x] = \mathbb{P}[t_i(1) \leq x] = \lambda_i \int_0^x \bar{G}_i(y)dy = \hat{G}_i(x) \tag{3}$$

for each i. We assume that

$$G_i(x) = G(\lambda_i x), \tag{4}$$

for some cdf G with mean 1. Note that (4) holds if $G_i(\cdot)$ is the exponential distribution. It follows from (2) and (4) that

$$\hat{G}_i(x) = \hat{G}(\lambda_i x). \tag{5}$$

We also assume that there exists a continuously differentiable cdf F with support in $[0, 1]$ such that for $i = 1, 2, \ldots, n$,

$$p_i^{(n)} = F\left(\frac{i}{n}\right) - F\left(\frac{i-1}{n}\right) = \frac{1}{n} F'(\xi_i^{(n)}), \tag{6}$$

where $\xi_i^{(n)} \in \left(\frac{i-1}{n}, \frac{i}{n}\right)$. The existence of $\xi_i^{(n)}$ is guaranteed by the mean-value theorem. We assume that $F'(x) > 0$ a.e. on $[0, 1]$. We allow $F'(0)$ to be infinite, to allow Zipf's law in particular.

Let $Y_i(t) = 1$ if content i was requested during the interval $[-t, 0)$ and $Y_i(t) = 0$ otherwise. With this notation, $Y(t) := \sum_{i=1}^{n} Y_i(t)$ is the number of distinct contents requested during $[-t, 0)$. Let $[-\tau, 0)$ be the smallest past interval such that there have been C distinct contents referenced in that interval, i.e.,

$$\tau = \inf\{t : Y(t) \geq C\}.$$

Note that if we reverse the arrow of time, we obtain in steady-state statistically the same request processes, and τ is a stopping time for the process $Y(t)$. The stationary hit probability of an LRU cache is then given by

$$H^{\text{LRU}} = \mathbb{P}^0[Y_{X_0}(\tau) = 1]. \tag{7}$$

If the cache is a TTL cache with timer T, the stationary hit probability is

$$H^{\text{TTL}}(T) = \mathbb{P}^0[Y_{X_0}(T) = 1]. \tag{8}$$

Note that $\mathbb{P}_i^0[Y_i(\tau) = 1]$ and $\mathbb{P}_i^0[Y_i(T) = 1]$ are the stationary hit probabilities of content i in an LRU cache and in a TTL cache with timer T, respectively. Define

$$\beta^\star(\nu) = \int_0^1 \hat{G}(\nu F'(x))dx, \;\; h^\star(\nu) = \int_0^1 F'(x)G(\nu F'(x))dx. \tag{9}$$

The next section shows that, as n becomes large, an LRU cache behaves as a TTL cache with a timer value that we identify.

3. ASYMPTOTIC BEHAVIOR

Throughout $T_n(\nu) = n\nu/\Lambda^{(n)}$ and $\beta_0 \in (0, 1)$.

PROPOSITION 1. *Assume that $C \sim n\beta_0$. Then,*

$$\lim_{n\to\infty} H^{\text{LRU}} = \lim_{n\to\infty} H^{\text{TTL}}(T_n(\nu_0)) = h^\star(\nu_0),$$

where ν_0 is the unique solution in $(0, \infty)$ of $\beta^\star(\nu) = \beta_0$. If G is continuous then, as $n \to \infty$,

$$\max_{1\leq i\leq n} \left|\mathbb{P}_i^0[Y_i(\tau) = 1] - \mathbb{P}_i^0[Y_i(T_n(\nu_0)) = 1]\right| \to 0. \tag{10}$$

This result states that the hit probability of a LRU cache converges to that of a TTL cache with timer $T \sim \nu_0 n/\Lambda^{(n)}$ as the cache size and number of contents increase to infinity. It was first proved rigorously by Fagin [4] in the IRM setting. Proposition 1 provides a rigorous extension of Fagin's result to independent and stationary request processes. It is easy to extend this proposition to several content popularity cdfs. The latter setting is useful when, for instance, several service providers share a common LRU cache and that contents associated with different providers exhibit different popularity probability distributions - see [7] for details.

The proof of Proposition 1 relies on the three following lemmas whose proofs are found in [7].

LEMMA 1. *The equation $\beta^\star(\nu) = \beta_0$ has a unique solution $\nu = \nu_0$ in $(0, \infty)$.*

The next lemma focuses on the hit probability in a TTL cache with timer T as the number of contents $\to \infty$.

LEMMA 2. *For $\nu > 0$,*

$$\lim_{n\to\infty} H^{\text{TTL}}(T_n(\nu)) = h^\star(\nu).$$

LEMMA 3. *For $T > 0$,*

$$\begin{aligned} \mathbb{P}^0[Y_{X_0}(\tau) = 1, \tau \leq T] &\leq \mathbb{P}^0[Y_{X_0}(T) = 1, \tau \leq T] \\ \mathbb{P}^0[Y_{X_0}(\tau) = 1, \tau \geq T] &\geq \mathbb{P}^0[Y_{X_0}(T) = 1, \tau \geq T]. \end{aligned}$$

The next lemma shows that τ is concentrated around $T_n(\nu_0)$.

LEMMA 4. *Assume that $C \sim \beta_0 n$. For $\nu_1 < \nu_0 < \nu_2$,*

$$\lim_{n\to\infty} \mathbb{P}^0[\tau < T_n(\nu_1)] = \lim_{n\to\infty} \mathbb{P}^0[\tau > T_n(\nu_2)] = 0.$$

We are now in position to prove Proposition 1.

PROOF OF PROPOSITION 1. Let $\nu_1 < \nu_0 < \nu_2$. We have

$$\begin{aligned} H^{\text{LRU}} &= \mathbb{P}^0[Y_{X_0}(\tau) = 1] \geq \mathbb{P}^0[Y_{X_0}(\tau) = 1, \tau \geq T_n(\nu_1)] \\ &\geq \mathbb{P}^0[Y_{X_0}(T_n(\nu_1)) = 1, \tau \geq T_n(\nu_1)] \text{ by Lemma 3} \\ &\geq \mathbb{P}^0[Y_{X_0}(T_n(\nu_1) = 1] - \mathbb{P}^0[\tau < T_n(\nu_1)] \\ &= H^{\text{TTL}}(T_n(\nu_1)) - \mathbb{P}^0[\tau < T_n(\nu_1)]. \end{aligned}$$

With the help of Lemmas 2 and 4 we find

$$\liminf_{n\to\infty} H^{\text{LRU}} \geq \lim_{n\to\infty} H^{\text{TTL}}(T_n(\nu_1)) = h^\star(\nu_1).$$

Letting $\nu_1 \to \nu_0$ gives $\liminf_{n\to\infty} H^{\text{LRU}} \geq h^\star(\nu_0)$. For the other direction, note that

$$\begin{aligned} H^{\text{LRU}} &= \mathbb{P}^0[Y_{X_0}(\tau) = 1] \\ &\leq \mathbb{P}^0[Y_{X_0}(\tau) = 1, \tau \leq T_n(\nu_2)] + \mathbb{P}^0[\tau > T_n(\nu_2)] \\ &\leq \mathbb{P}^0[Y_{X_0}(T_n(\nu_2)) = 1, \tau \leq T_n(\nu_2)] + \mathbb{P}^0[\tau > T_n(\nu_2)] \qquad (11) \\ &\leq \mathbb{P}^0[Y_{X_0}(T_n(\nu_2)) = 1] + \mathbb{P}^0[\tau > T_n(\nu_2)] \\ &= H^{\text{TTL}}(T_n(\nu_2)) + \mathbb{P}^0[\tau > T_n(\nu_2)], \end{aligned}$$

where (11) follows from Lemma 3. With the help of Lemmas 2 and 4 we obtain

$$\limsup_{n\to\infty} H^{\text{LRU}} \leq \lim_{n\to\infty} H^{\text{TTL}}(T_n(\nu_2)) = h^\star(\nu_2).$$

Letting $\nu_2 \to \nu_0$ yields $\limsup_{n\to\infty} H^{\rm LRU} \le h^\star(\nu_0)$. Therefore $\lim_{n\to\infty} H^{\rm LRU} = h^\star(\nu_0)$. The proof of the uniform convergence result (10) can be found in [7]. □

4. RATE OF CONVERGENCE

In this section we investigate the rate of convergence of $H^{\rm LRU}$ to $H^{\rm TTL}(T_0)$, where the timer T_0 is defined below. $\Lambda^{(n)} = 1$. Define $C(T) = \mathbb{E}[Y(T)]$, the expected number of contents in a TTL cache with timer T.

LEMMA 5. *$C(T) = \sum_{i=1}^n \hat{G}_i(T)$ and $T \to C(T)$ is strictly increasing for all T such that $C(T) < n$.*

From $C(0) = 0$, $C(\infty) = n$ and the strict increasingness of $T \mapsto C(T)$, we know that $C(T) = C \in (0, n)$ has a unique solution in $(0, \infty)$, which we call T_0. Define $\mu_0 = C'(T_0) = \sum_{i=1}^n \lambda_i(1-G_i(T_0))$, the miss rate in a TTL cache with timer T_0. Assume that $\mu_0 > 0$. Let $\delta_i = 0$ if the request process for content $i = 1, \ldots, n$ is Poisson and $\delta_i = 1$ otherwise.

PROPOSITION 2. *Let $\delta = \max_{1\le i\le n} \delta_i$ and $\mu \in (0, \mu_0)$. Suppose there exist a constant B and $\rho \in \left[\frac{\delta}{\mu T_0}, \min\left\{1, \frac{\mu_0-\mu}{B}\right\}\right]$ such that*

$$|\bar{G}_i(T_0) - \bar{G}_i(T_0 \pm \varepsilon T_0)| \le B\varepsilon \quad \text{for } \varepsilon \in \left[\frac{\delta}{\mu T_0}, \rho\right]. \qquad (12)$$

Let $D_0 = \sqrt{\frac{2}{n}}\mu T_0$. Assume $D_0/B \ge \sqrt{\frac{e}{2}}$ and

$$1 + \sqrt{\log \frac{D_0}{B}} \le D_0\rho - \sqrt{\frac{2}{n}}\delta.$$

Then,

$$\begin{aligned}\Delta_n(T_0) &:= |H^{\rm LRU} - H^{\rm TTL}(T_0)| \\ &\le \max_{1\le i\le n} \left|\mathbb{P}_i^0[Y_i(\tau) = 1] - \mathbb{P}_i^0[Y_i(T_0) = 1]\right| \\ &\le \frac{B\delta}{\mu T_0} + \frac{B}{D_0}\left(\sqrt{\log\frac{D_0}{B}} + 1 + \frac{1}{\sqrt{2}}\right).\end{aligned}$$

See [7] for the proof. Note that (12) holds for a large class of distributions.

Example 1. For Poisson arrivals, $\bar{G}_i(t) = e^{-\lambda_i t}$. For any $\varepsilon \ge 0$,

$$\begin{aligned}0 &\le \bar{G}_i(T_0) - \bar{G}_i(T_0 + \varepsilon T_0) = e^{-\lambda_i T_0}(1 - e^{-\lambda_i \varepsilon T_0}) \\ &\le \lambda_i T_0 e^{-\lambda_i T_0}\varepsilon \le e^{-1}\varepsilon.\end{aligned}$$

For $\varepsilon \in [0, 1]$,

$$\begin{aligned}0 &\le \bar{G}_i(T_0 - \varepsilon T_0) - \bar{G}_i(T_0) \le \sup_{x\ge 0} e^{-x}(e^{\varepsilon x} - 1) \\ &= (1-\varepsilon)^{\frac{1}{\varepsilon}-1}\varepsilon \le \varepsilon.\end{aligned}$$

Thus (12) holds with $B = 1$ and $\rho = \mu_0 - \mu$.

Example 2. Suppose G_i's are continuously differentiable. By the mean value theorem, it exists $\xi_i \in [1, 1+\varepsilon]$ such that

$$\begin{aligned}0 &\le \bar{G}_i(T_0) - \bar{G}_i(T_0 + \varepsilon T_0) = G_i'(\xi_i T_0)\varepsilon T_0 \\ &\le \xi_i T_0 G_i'(\xi_i T_0)\varepsilon \le \left[\sup_{t\ge 0} tG_i'(t)\right]\varepsilon.\end{aligned}$$

Similarly, there exists $\zeta_i \in [1-\varepsilon, 1]$ such that

$$\begin{aligned}0 &\le \bar{G}_i(T_0 - \varepsilon T_0) - \bar{G}_i(T_0) = G_i'(\zeta_i T_0)\varepsilon T_0 \\ &\le \frac{\zeta_i}{1-\varepsilon} T_0 G_i'(\zeta_i T_0)\varepsilon \le \frac{\varepsilon}{1-\rho}\left[\sup_{t\ge 0} tG_i'(t)\right].\end{aligned}$$

Thus (12) holds with $\rho \in [\frac{\delta}{\mu_0 T_0}, \frac{\mu_0-\mu}{A+\mu_0-\mu}]$ and $B = \frac{A}{1-\rho}$, where $A = \max_{1\le i\le n} \sup_{t\ge 0} tG_i'(t)$.

Below is the rate of convergence for Poisson arrivals.

COROLLARY 1. *Assume that $\delta_i = 0$ for each i. If $C \sim \beta_0 n$, $\min_i \lambda_i = \Omega(n^{-\gamma})$ for $1 \le \gamma < 3/2$,*

$$\begin{aligned}\Delta_n(T_0) &\le \max_{1\le i\le n} \left|\mathbb{P}_i^0[Y_i(\tau) = 1] - \mathbb{P}_i^0[Y_i(T_0) = 1]\right| \\ &= O\left(n^{\gamma-3/2}\sqrt{\log n}\right).\end{aligned}$$

5. CONCLUSIONS

In this paper, we developed an approximation for the aggregate and individual content hit rates of an LRU cache for the case that content requests are described by independent stationary and ergodic processes. This approximation extends one first proposed and studied by Fagin [4] for the independent reference model and provides the theoretical basis for approximations introduced in [6] for content requests described by independent renewal processes. We showed that the approximations become exact in the limit as the number of contents goes to infinity while the ratio of this and the cache size remains constant. Last, we established the rate of convergence for the approximation as number of contents increases.

Future directions include extension of these results to other cache policies such as FIFO and random and to networks of caches. In addition, it is desirable to relax independence between different content request streams.

6. REFERENCES

[1] F. Baccelli and P. Brémaud. *Elements of Queueing Theory: Palm Martingale Calculus and Stochastic Recurrences*, volume 26 of *Applications of Mathematics, Stochatic Modelling and Applied Probability*. Springer-Verlag Berlin Heidelberg, 2nd edition, 2003.

[2] H. Che, Y. Tung, and Z. Wang. Hierarchical Web caching systems: Modeling, design and experimental results. *IEEE J. on Selected Areas in Communications*, 20(7):1305–1314, 2002.

[3] A. Dan and D. Towsley. An approximate analysis of the LRU and FIFO buffer replacement schemes. In *Proc. SIGMETRICS*, pages 143–152, Boulder, CO, USA, May 1990.

[4] R. Fagin. Asymptotic miss ratios over independent references. *Journal of Computer and System Sciences*, 14(2):222–250, 1977.

[5] C. Fricker, P. Robert, and J. Roberts. A versatile and accurate approximation for LRU cache performance. In *Proc. 24th International Teletraffic Congress (ITC 24)*, Kraków, Poland, September 4-7 2012.

[6] M. Garetto, E. Leonardi, and V. Martina. A unified approach to the performance analysis of caching systems. *ACM Trans. on Modeling and Performance Evaluation of Computing Systems (TOMPECS)*, 1(3), May 2016.

[7] B. Jiang, P. Nain, and D. Towsley. LRU cache under stationary requests. Technical report, University of Massachusetts, April 2017.