RC24058 (W0609-092) September 21, 2006
Mathematics

IBM Research Report

Critical Sizing of LRU Caches with Dependent Requests

Predrag R. Jelenkovic
Department of Electrical Engineering
Columbia University
New York, NY 10027

Ana Radovanovic, Mark S. Squillante
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Applied Probability Trust (20 September 2006)

CRITICAL SIZING OF LRU CACHES WITH DEPENDENT REQUESTS

PREDRAG R. JELENKOVIĆ,[*] *Columbia University*

ANA RADOVANOVIĆ AND MARK S. SQUILLANTE,[**] *IBM T.J. Watson Research Center*

Abstract

It was recently proved in [13] that the Least-Recently-Used (LRU) caching policy, in the presence of semi-Markov modulated requests that have a generalized Zipf's law popularity distribution, is asymptotically insensitive to the correlation in the request process. However, since the previous result is asymptotic, it remains unclear how small the cache size can become while still retaining the preceding insensitivity property. In this paper, assuming that requests are generated by a nearly completely decomposable Markov-modulated process, we characterize the critical cache size below which the dependency of requests dominates the cache performance. This critical cache size is small relative to the dynamics of the modulating process, and in fact it is sublinear with respect to the sojourn times of the modulated chain that determines the dependence structure.

Keywords: network caching, least-recently-used caching, move-to-front searching, nearly completely decomposable Markov processes, Markov-modulated processes, Zipf's law

2000 Mathematics Subject Classification: Primary 60J25; 60F10
Secondary 68P10; 60G50

1. Introduction

The basic idea of caching is to maintain high-speed access to a subset of x popular items out of a larger collection of N documents that are otherwise accessed at a slower rate. In the context of Internet applications and services, such as Web access and content delivery, caching

[*] Postal address: Department of Electrical Engineering, Columbia University, New York, NY 10027, USA. predrag@ee.columbia.edu. The work of this author was supported by NSF Grant No. 0092113.

[**] Postal address: Mathematical Sciences Department, IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598. {aradovan, mss}@us.ibm.com

has been widely recognized as an effective way to reduce the latency for downloading Internet documents. This is achieved by keeping the popular documents in high-speed caches that are located close to the users requesting these documents. Naturally, the problem of selecting and possibly dynamically updating the content of a cache is central to the efficient operation of any caching system. The broad popularity of the LRU policy stems from its many desirable characteristics, including a high hit ratio, low complexity, and flexibility to dynamically adapt to possible changes in the request patterns.

Due to its importance in practice, LRU caching has received significant attention in the research literature, both in the context of combinatorial (worst-case) [3, 4, 15, 16] and probabilistic (average-case) analysis; the latter is the focus of this paper. In particular, we consider the LRU algorithm in the presence of strong statistical correlation that often characterizes the access patterns for Internet documents; e.g., see [1, 5, 12] and the references therein. However, most of the existing work on the average-case analysis of LRU caches is either performed under the assumption of independent and identically distributed (i.i.d.) requests or it is computationally intractable. To alleviate this problem, in our recent work [13] we develop a novel, analytically explicit asymptotic method for analyzing LRU caches in the presence of semi-Markov modulated requests. This way of modeling dependency in the request process provides the desired flexibility for capturing possibly strong statistical correlation, including the widely reported long-range dependence of the access patterns for Web documents. In fact, Markov modulation techniques are widely used to model dependencies in the arrival processes in the context of queueing and insurance risk theories; e.g., see [2, 11] and the references therein. The main results from [12, 13] imply that asymptotically, for large cache sizes, the cache fault probability in the presence of semi-Markov modulated requests behaves the same as in the corresponding LRU system with i.i.d. requests [10]. This surprising insensitivity was further validated experimentally in [12] where we found excellent agreement with the asymptotic results, even in the cases of actual trace-driven simulations and for relatively small cache sizes, which further supports our way of modeling dependency in the request process.

Since the results from [13] are asymptotic, they do not provide information on how small the cache sizes can become while still retaining the discovered insensitivity property. Our present work attempts to answer this question by studying the cache performance through a joint scaling of the dependence structure of the requests and the cache size. In this paper, the request sequence is modeled as a nearly completely decomposable (NCD) Markov-modulated process with the modulating Markov process having transition rates linearly proportional to a scaling parameter δ. The jumps in this modulating process occur on a time scale of the order $1/\delta$, which implies that the dependency in the request process increases as $\delta \downarrow 0$. We scale the cache size as an increasing function of $1/\delta$ and identify a critical cache sizing below which the dependency dominates the cache performance. Our main results show that this critical cache size is sublinear in comparison with the time scale of transitions in the modulating process, i.e., $1/\delta$. Thus, informally, our results show that the discovered insensitivity property is indeed robust.

The remainder of this paper is organized as follows. In Section 2 we define the model used in our study, while in Section 3 we present a summary of results that are used in our main theorems. The main results are provided in Theorems 1 and 2 of Section 4, together with a discussion of their implications. In Section 5 we conclude the paper.

2. Model description

A LRU cache of size x can be described as follows. Consider a universe of N documents (items), from which x can be placed in an efficiently accessible location called the cache. Each time a request for a document is made, the cache is searched first. If the document is not found in the cache (cache fault), additional delay is incurred to access the item from the outside universe and it is added to the cache by replacing the least recently accessed document in the cache. The performance measure of interest for this algorithm is the LRU fault probability, i.e., the probability that the requested document is not found in the cache.

Analyzing the LRU policy is equivalent to investigating the Move-To-Front (MTF) searching algorithm. In order to justify this claim, we assume that the x documents in the cache, under the LRU rule, are arranged in increasing order of their last access times. Every time there is a request for a document that is not in the cache, the document is brought to the first position of the cache and the last document in the cache is moved to the outside universe. Clearly, the fault probability stays the same if the remaining $N - x$ documents in the outside universe are arranged in any specific order. In particular, they can be arranged in increasing order of their last access times. The obtained searching scheme performed on the ordered list of all documents is called the MTF algorithm. Furthermore, it is clear from the previous arguments that the LRU fault probability is equal to the tail of the MTF search cost, i.e., the position of the requested document evaluated at the cache size. Additional arguments that justify the connection between the MTF search cost distribution and the LRU cache fault probability can be found in [7, 9, 10]. We therefore proceed with a description of the MTF algorithm.

More formally, consider a finite set of documents $L = \{1, \dots, N\}$ and a sequence of document requests that arrive at time points $\{\tau_n, -\infty < n < \infty\}$ which represent a Poisson process of unit rate. At each point τ_n, we use R_n to denote the document that has been requested, i.e., the event $\{R_n = i\}$ represents a request for document i at time τ_n. The sequence $\{R_n\}$ is assumed to be independent of the Poisson arrival points $\{\tau_n\}$. The dynamics of the MTF algorithm are defined as follows. Suppose that the system starts at the arrival instant τ_0 of the 0th request with an initial permutation Π_0 of the MTF list. Then, every time τ_n $(n \geq 0)$ that a document is requested, its position in the list is first determined and this value represents the searching cost $C_n^{(N)}$ at time τ_n. The list is then updated by moving the requested document to the first position of the list and shifting one position down those documents that were in front of the requested item. Note that, according to the discussion in the preceding paragraph, $\mathbb{P}[C_n^{(N)} > x]$ represents the fault probability of a cache of size x at time τ_n.

Next, we characterize the dependence structure of the request process. Let $N_\delta = \{T_n, -\infty < n < \infty\}$, $T_0 \leq 0 < T_1$, be a Poisson point process with rate $\delta > 0$. Furthermore, let $\{\mathcal{J}_n, -\infty < n < \infty\}$ be a finite-state, irreducible, aperiodic Markov chain, independent of N_δ, taking values in $\{1, \dots, M\}$, where M is a finite, positive integer. This process is assumed to be stationary with marginal distribution $\pi_k = \mathbb{P}[\mathcal{J}_n = k]$. Then, by embedding this Markov chain into the Poisson process N_δ, we construct a piecewise constant right-continuous *modulating process* J, where J is defined as $J_t = \mathcal{J}_n$ for $T_n \leq t < T_{n+1}$. Note that the transition rates in J are linearly proportional to δ and, therefore, this is a NCD process for small δ.

For each $1 \leq k \leq M$, let $q_i^{(k)}$ be a probability mass function where $q_i^{(k)}$ is used to denote the probability of requesting document i when the underlying process J is in state k, $1 \leq i \leq N$. The dynamics of R_n are then uniquely determined by the modulating process J according to the equation

$$\mathbb{P}[R_l = i_l,\, 1 \leq l \leq n \,|\, J_t,\, t \leq \tau_n] \;=\; \prod_{l=1}^{n} q_{i_l}^{(J_{\tau_l})},$$

where $n \geq 1$; that is, the sequence of requests R_n is conditionally independent given the modulating process J. We use $q_i = \mathbb{P}[R = i] = \sum_{k=1}^{M} \pi_k q_i^{(k)}$ to express the marginal request distribution and assume that $q_i > 0$ for all $1 \leq i \leq N$.

3. Preliminary results

The model described in the previous section, for a fixed δ, is a special case of the more general one introduced in [13] and, therefore, some of the results from [13] are used in this paper. In particular, Lemma 1 of [13] shows that the search cost $C_n^{(N)}$, $N < \infty$, converges in distribution to the stationary value $C^{(N)}$ when the request process $\{R_n\}$ is stationary and ergodic. Then, in the following subsection, we outline this convergence and provide a characterization of the tail of the limiting search cost distribution when the number of documents $N \to \infty$. Next, Subsection 3.2 contains results on MTF searching with i.i.d. requests that were stated and proved in [10] and [13] and will be used in proving our main theorems.

3.1. Representation results

Section 3.1 of [13] contains a general characterization of the stationary distribution of $C^{(N)}$, $N < \infty$. Assume that the probability mass functions $q_i^{(k)}$ are defined for every $i \geq 1, 1 \leq k \leq M$. Then, let σ_t be the σ-algebra $\sigma(J_u, -t \leq u \leq 0)$ containing the history of the process J_t in the interval $[-t, 0]$ and denote the conditional probability $\mathbb{P}_{\sigma_t}[\cdot] = \mathbb{P}[\cdot|\sigma_t]$. Furthermore, let $N_j(u; J)$ be the number of requests for document j in $[-u, 0)$, $0 < u \leq t$, and define an indicator function $B_j(t; J) = 1[N_j(t; J) > 0]$, $j \geq 1$, being equal to 1 if item j was requested in $[-t, 0)$. Then, the number of distinct documents $S_i(t; J)$, different from i, that were requested in $[-t, 0)$ can be expressed as

$$S_i(t; J) \triangleq \sum_{j \neq i} B_j(t; J), \tag{1}$$

where

$$\mathbb{P}_{\sigma_t}[B_j(t; J) = 1] = 1 - e^{-\hat{q}_j t}. \tag{2}$$

We use $\hat{q}_j \equiv \hat{q}_j(t)$, $j \geq 1$, in (2) to denote the empirical probabilities of requesting document j in the interval $[-t, 0)$, i.e.,

$$\hat{q}_j \triangleq \sum_{k=1}^{M} q_j^{(k)} \hat{\pi}_k \quad \text{and} \quad \hat{\pi}_k \equiv \frac{1}{t} \int_{-t}^{0} 1[J_u = k]\, du. \tag{3}$$

Now, one can construct a sequence of finite lists and show that their search costs $C^{(N)}$ converge in distribution, as $N \to \infty$, to

$$\mathbb{P}[C > x] \;=\; \mathbb{E} \int_0^\infty \hat{f}(t)\mathbb{P}_{\sigma_t}[S_i(t; J) > x - 1]\, dt, \tag{4}$$

where $\hat{f}(t)$ is defined as

$$\hat{f}(t) \triangleq \sum_{i=1}^{\infty} q_i^{(J_0)} q_i^{(J_{-t})} e^{-\hat{q}_i t}, \tag{5}$$

and $\hat{q}_i$, $S_i(t; J)$ are as introduced in (1) - (3). The reader is referred to the proof of Proposition 1 in [13] for details, where the above results are established under more general model assumptions on the request process $\{R_n\}$ than the ones introduced in Section 2. The preceding representation of the distribution of C is the starting point of our analysis.

Remark 1. (i) Throughout this paper we will exploit the properties that the variables $S_j(t; J)$, $B_j(t; J)$, $j \geq 1$, are monotonically increasing in t and that the variables $B_j(t; J)$, $j \geq 1$, are conditionally independent given σ_t. This conditional independence arises from the Poisson arrival structure, as is apparent from the derivation in [13]. In general, when the request times are not Poisson, e.g., integer time arrivals, these variables may not be conditionally independent. For i.i.d. requests, the Poisson embedding technique was first introduced in [8]. (ii) It is clear that the derivation of the above results does not rely on the fact that the requests arrive at a constant rate [13]. Thus, our results can be generalized to the case where the arrival rate depends on the state of the modulating process J, i.e., the rate can be set to λ_k when $J_t = k$. We do not consider this extension, since it further complicates the notation without providing any significant new insight.

3.2. Results for i.i.d. requests

We next provide several lemmas that consider the LRU caching scheme under independent requests, which will be used in proving our main theorems. The MTF model with i.i.d. requests is equivalent to our general problem formulation when the modulating process is assumed to be a constant, i.e., $J_t \equiv \mathrm{constant}$. In this case the Bernoulli variables $\{B_j(t),\, j \geq 1\}$ indicating that a document j was requested in $[-t, 0)$ are independent with success probabilities $\mathbb{P}[B_i(t) = 1] = 1 - e^{-q_i t}$. Then, using the notation $S_i(t) \triangleq \sum_{j \neq i} B_j(t)$, it is easy to see that the distribution of the limiting stationary search cost C from (4) reduces to

$$\mathbb{P}[C > x] = \int_0^\infty \sum_{i=1}^{\infty} q_i^2 e^{-q_i t}\mathbb{P}[S_i(t) > x - 1] dt. \tag{6}$$

Throughout this paper we shall use some standard notation. For any two real functions $a(t)$ and $b(t)$ and fixed $t_0 \in \mathbb{R} \cup \{\infty\}$, we will use $a(t) \sim b(t)$ as $t \to t_0$ to denote $\lim_{t\to t_0}[a(t)/b(t)] = 1$. Similarly, we say that $a(t) \gtrsim b(t)$ as $t \to t_0$ if $\liminf_{t\to t_0} a(t)/b(t) \geq 1$; $a(t) \lesssim b(t)$ has a complementary definition. The following two results, originally proved in Lemmas 1 and 2 of [10], are restated here for convenience.

Lemma 1. *Assume that $q_i \sim c/i^\alpha$ as $i \to \infty$, with $\alpha > 1$ and $c > 0$. Then, as $t \to \infty$,*

$$\sum_{i=1}^{\infty} q_i^2 e^{-q_i t} \sim \frac{c^{\frac{1}{\alpha}}}{\alpha}\Gamma\left(2 - \frac{1}{\alpha}\right) t^{-2+\frac{1}{\alpha}},$$

where Γ is the Gamma function.

Lemma 2. *Let $S(t) = \sum_{i=1}^{\infty} B_i(t)$ and assume $q_i \sim c/i^\alpha$ as $i \to \infty$, with $\alpha > 1$ and $c > 0$. Then, as $t \to \infty$,*

$$m(t) \triangleq \mathbb{E}S(t) \sim \Gamma\left(1 - \frac{1}{\alpha}\right) c^{\frac{1}{\alpha}} t^{\frac{1}{\alpha}}.$$

Throughout this paper we shall use H to be a sufficiently large positive constant, whereas h will be used to denote a sufficiently small positive constant. The values of H and h are generally different in different places. For example, $H/2 = H$, $H^2 = H$, $H + 1 = H$, etc. Now, the next two lemmas, which are repeatedly used in establishing our main results, were originally proved in [13].

Lemma 3. *Let $\{B_i, i \geq 1\}$ be a sequence of independent Bernoulli random variables, $S = \sum_{i=1}^{\infty} B_i$ and $m = \mathbb{E}[S]$. Then for any $\epsilon > 0$, there exists $\theta_\epsilon > 0$, such that*

$$\mathbb{P}[|S - m| > m\epsilon] \leq He^{-\theta_\epsilon m}.$$

Lemma 4. *If $0 \leq q_i \leq H/i^\alpha$ for some fixed $\alpha > 1$, then for any $x \geq 1$,*

$$\mathbb{P}[C > x] \leq \frac{H}{x^{\alpha-1}}.$$

Finally, the result established in the following lemma is repeatedly used in the proof of Theorem 1.

Lemma 5. *Let $c/i^\alpha \leq q_i \leq c^{-1}/i^\alpha$, $\alpha > 1$, for some positive constant c. Then, for any $x > 0$,*

$$\mathbb{P}[C > x] \geq \frac{h}{x^{\alpha-1}}.$$

Proof. Note that for any $\epsilon > 0$ and x large enough, by using (6), the tail of the search cost C can be lower bounded as

$$\begin{aligned}\mathbb{P}[C > x] &\geq \mathbb{P}[S(Hx^\alpha) > x - 1] \int_{Hx^\alpha}^{\infty} \sum_{i=1}^{\infty} q_i^2 e^{-q_i t} dt \\ &\geq (1-\epsilon) \sum_{i=1}^{\infty} q_i e^{-Hx^\alpha q_i}, \qquad (7)\end{aligned}$$

where the first inequality follows from the monotonicity of $S(t)$, while the second inequality is obtained by applying Lemmas 2, 3 and integration. Next, from the assumptions of the lemma, we have

$$\sum_{i=1}^{\infty} q_i e^{-Hq_i x^\alpha} \geq \sum_{\lfloor x \rfloor + 1}^{\infty} \frac{c}{i^\alpha} e^{-H\frac{x^\alpha}{i^\alpha}} \geq ce^{-H} \int_{x+1}^{\infty} \frac{1}{u^\alpha} du \geq \frac{h}{x^{\alpha-1}},$$

which in conjunction with (7) proves the result.

4. Main results

In this section we state and prove our main results. We show that the cache fault probability exhibits different performance characteristics depending on the scaling between the cache size x and the parameter δ.

In preparation for these proofs we denote the epochs of reversed jump points $\mathcal{T}_n \triangleq -T_{-n}$, $n \geq 0$; this notation is convenient since C depends on J_t for values $t \leq 0$. Furthermore, we define $S^{(k)}(t) \triangleq B_i^{(k)}(t) + S_i^{(k)}(t) = B_i^{(k)}(t) + \sum_{j \neq i} B_j^{(k)}(t)$, $1 \leq k \leq M$, where $B_i^{(k)}(t)$, $i \geq 1$, are Bernoulli random variables with $\mathbb{P}[B_i^{(k)}(t) = 1] = 1 - e^{-q_i^{(k)} t}$. In addition, let $C^{(k)}$ correspond to the stationary search cost with i.i.d. requests when $J_t \equiv k$.

4.1. Asymptotic decomposability

The following theorem establishes the critical cache size scaling as a function of the parameter δ below which the dependency in the request process dominates cache performance, i.e., the insensitivity result does not hold.

Theorem 1. *Let $q_i \leq c_1/i^\alpha$, $\alpha > 1$, and suppose there exists k, $1 \leq k \leq M$, such that $q_i^{(k)} \geq c_2/i^\alpha$, $c_2 > 0$. If x_δ satisfies $x_\delta \delta^{1/\alpha} \to 0$ as $\delta \to 0$, then*

$$\mathbb{P}[C > x_\delta] \sim \sum_{k=1}^{M} \pi_k \mathbb{P}[C^{(k)} > x_\delta] \;\; as \;\; x_\delta \to \infty. \qquad (8)$$

Proof. To simplify notation we write $x \equiv x_\delta$. First, we prove the lower bound. Since $S(t;J) = S^{(k)}(t)$ a.s. on $\{J_0 = k\}$ for all $-\mathcal{T}_0 \leq t \leq 0$, the representation formula given in (4) implies

$$\begin{aligned}
\mathbb{P}[C > x] &= \mathbb{E}\int_0^\infty \hat{f}(t)\mathbb{P}_{\sigma_t}[S_i(t;J) > x-1]dt \\
&\geq \mathbb{E}\int_0^{\mathcal{T}_0} \sum_{i=1}^\infty (q_i^{(J_0)})^2 e^{-q_i^{(J_0)}t}\mathbb{P}[S_i^{(J_0)}(t) > x-1|J_0]dt \\
&\geq \sum_{k=1}^M \mathbb{P}[J_0 = k, \mathcal{T}_0 > Hx^\alpha]\int_0^\infty \sum_{i=1}^\infty (q_i^{(k)})^2 e^{-q_i^{(k)}t}\mathbb{P}[S_i^{(k)}(t) > x-1]dt \\
&\qquad - \sum_{k=1}^M \pi_k \int_{Hx^\alpha}^\infty \sum_{i=1}^\infty (q_i^{(k)})^2 e^{-q_i^{(k)}t}dt. \qquad (9)
\end{aligned}$$

Now, since $q_i^{(k)} \leq \bar{q}_i \triangleq q_i/\min_k \pi_k$, $1 \leq k \leq M$, $q_i \leq c_1/i^\alpha$ and $xe^{-x} \leq e^{-1}$ (for $x \geq 0$), the second summand in (9) can be bounded as

$$\begin{aligned}
\sum_{k=1}^M \pi_k \int_{Hx^\alpha}^\infty \sum_{i=1}^\infty (q_i^{(k)})^2 e^{-q_i^{(k)}t}dt &\leq \sum_{k=1}^M \pi_k \frac{1}{Hx^\alpha} \sum_{i=1}^{\lfloor H^{1/\alpha}x\rfloor} q_i^{(k)} Hx^\alpha e^{-q_i^{(k)}Hx^\alpha} + \frac{1}{(\min_k \pi_k)}\int_{H^{1/\alpha}x}^\infty \frac{c_1}{y^\alpha}dy \\
&\leq \frac{1}{H^{1-1/\alpha}}\frac{1}{x^{\alpha-1}}\left(e^{-1} + \frac{c_1}{(\min_k \pi_k)(\alpha-1)}\right). \qquad (10)
\end{aligned}$$

Then, by the assumption of the theorem, $\mathbb{P}[J_0 = k, \mathcal{T}_0 > Hx^\alpha] = \pi_k e^{-H\delta x^\alpha} \to \pi_k$ as $\delta \to 0$ ($x \to \infty$), and, therefore, from (9) and (10) we obtain

$$\mathbb{P}[C > x] \gtrsim \sum_{k=1}^M \pi_k \mathbb{P}[C^{(k)} > x] - \frac{1}{H^{1-1/\alpha}}\left(e^{-1} + \frac{c_1}{(\min_k \pi_k)(\alpha-1)}\right)\frac{1}{x^{\alpha-1}} \quad \text{as} \quad x \to \infty.$$

To simplify the notation in the remainder of the paper we will simply write $f(x) \gtrsim, \sim, \lesssim g(x)$ without explicit reference to $x \to \infty$. Next, by applying Lemma 5 and letting $H \to \infty$, we conclude

$$\mathbb{P}[C > x] \gtrsim \sum_{k=1}^M \pi_k \mathbb{P}[C^{(k)} > x]. \qquad (11)$$

Let us now prove the upper bound. After splitting the integral in (4), we define

$$\mathbb{P}[C > x] = \mathbb{E}\int_0^{\mathcal{T}_0} + \mathbb{E}\int_{\mathcal{T}_0}^\infty \triangleq I_1(x) + I_2(x). \qquad (12)$$

First, we provide an upper bound for $I_1(x)$. Since $S(t;J) = S^{(k)}(t)$ a.s. on $\{J_0 = k\}$, we derive

$$I_1(x) = \mathbb{E}\sum_{k=1}^{M} 1[J_0 = k] \int_0^{\mathcal{T}_0} \sum_{i=1}^{\infty} (q_i^{(k)})^2 e^{-q_i^{(k)}t} \mathbb{P}[S_i^{(k)}(t) > x-1]dt$$
$$\leq \sum_{k=1}^{M} \pi_k \int_0^{\infty} \sum_{i=1}^{\infty} (q_i^{(k)})^2 e^{-q_i^{(k)}t} \mathbb{P}[S_i^{(k)}(t) > x-1]dt = \sum_{k=1}^{M} \pi_k \mathbb{P}[C^{(k)} > x], \quad (13)$$

where the inequality is obtained after replacing $\mathcal{T}_0$ with ∞.

Next, in deriving an asymptotic estimate of $I_2(x)$, we use $q_i^{(J_{-t})} e^{-\hat{q}_i t} dt = -d(e^{-\hat{q}_i t})$ as follows

$$I_2(x) \leq \mathbb{E}\sum_{i=1}^{\infty} q_i^{(J_0)} \int_{\mathcal{T}_0}^{\infty} q_i^{(J_{-t})} e^{-\hat{q}_i t} dt = \mathbb{E}\sum_{i=1}^{\infty} q_i^{(J_0)} \int_{\mathcal{T}_0}^{\infty} -d(e^{-\hat{q}_i t})$$
$$= \mathbb{E}\sum_{i=1}^{\infty} q_i^{(J_0)} e^{-q_i^{(J_0)}\mathcal{T}_0} = \sum_{k=1}^{M} \pi_k \sum_{i=1}^{\infty} \frac{q_i^{(k)}\delta}{q_i^{(k)} + \delta}.$$

Since the first assumption of the theorem implies $q_i^{(k)} \leq H/i^\alpha$, $1 \leq k \leq M$, using the inequality

$$\sum_{i=1}^{\infty} \frac{q_i^{(k)}}{q_i^{(k)} + \delta} \leq \sum_{i=1}^{\infty} \frac{1}{1 + h\delta i^\alpha} \leq \int_0^{\infty} \frac{1}{1 + h\delta z^\alpha} dz \leq \frac{1}{(h\delta)^{1/\alpha}} \int_0^{\infty} \frac{1}{1 + y^\alpha} dy, \quad (14)$$

we obtain

$$I_2(x) \leq H\delta^{1-1/\alpha} = o\left(\frac{1}{x^{\alpha-1}}\right), \quad (15)$$

where the last equality is implied by the assumption of the theorem since $x\delta^{1/\alpha} \to 0$ as $\delta \to 0$ yields $\delta^{1-1/\alpha} = o(1/x^{\alpha-1})$. Finally, this last observation together with (13) and Lemma 5 imply $I_2(x) = o(I_1(x))$, which, in conjunction with (12) and (11), concludes the proof of the theorem.

4.2. Asymptotic insensitivity

The following theorem establishes the scaling of the cache size as a function of the parameter δ for which the insensitivity result holds.

Theorem 2. *Let $q_i \sim c/i^\alpha$ as $i \to \infty$, $\alpha > 1$. If x_δ satisfies $x_\delta \delta^{1/\alpha}/\log x_\delta \to \infty$ as $\delta \to 0$, then*

$$\mathbb{P}[C > x_\delta] \sim K(\alpha)\mathbb{P}[R > x_\delta] \ \ as \ \ x_\delta \to \infty, \quad (16)$$

where

$$K(\alpha) \triangleq \left(1 - \frac{1}{\alpha}\right)\left[\Gamma\left(1 - \frac{1}{\alpha}\right)\right]^\alpha,$$

and Γ *is the Gamma function.*

Remark 2. The following proof is based on Theorem 3 of [13]. The main technical novelty of the present paper is to demonstrate that the estimates from this proof hold *uniformly* for all small δ. In this regard, the proof below draws these parallels and emphasizes the derivation of the uniform bounds.

Proof. Again, to simplify the notation, we set $x \equiv x_\delta$. First we prove the upper bound. After splitting the integral in (4), we define

$$\begin{aligned}\mathbb{P}[C > x] &= \mathbb{E}\int_0^{\mathcal{T}_{\lfloor hx^\alpha\delta\rfloor}} + \mathbb{E}\int_{\mathcal{T}_{\lfloor hx^\alpha\delta\rfloor}}^\infty \\ &\triangleq I_1(x) + I_2(x). \end{aligned} \tag{17}$$

Next, we show that $I_1(x)$ is negligible for large x, i.e.,

$$I_1(x) = o\left(\frac{1}{x^{\alpha-1}}\right). \tag{18}$$

To this end, after conditioning on the value of $\mathcal{T}_{\lfloor hx^\alpha\delta\rfloor}$, we obtain

$$\begin{aligned} I_1(x) \le \mathbb{E}&\left[1[\mathcal{T}_{\lfloor hx^\alpha\delta\rfloor} > 2hx^\alpha]\int_0^{\mathcal{T}_{\lfloor hx^\alpha\delta\rfloor}} \hat{f}(t)\mathbb{P}_{\sigma_t}[S(t;J) \ge x]dt\right] \\ &+ \mathbb{E}\left[1[\mathcal{T}_{\lfloor hx^\alpha\delta\rfloor} \le 2hx^\alpha]\int_0^{\mathcal{T}_{\lfloor hx^\alpha\delta\rfloor}} \hat{f}(t)\mathbb{P}_{\sigma_t}[S(t;J) \ge x]dt\right], \end{aligned}$$

where $\hat{f}(t)$ is defined in (5). Note that $\hat{f}(t) \le \sum_{i=1}^\infty q_i^{(J_0)} = 1$ and

$$\int_0^\infty \hat{f}(t)dt = 1, \tag{19}$$

since $-d(e^{-\hat{q}_i t}) = e^{-\hat{q}_i t}d(\sum_{k=1}^M q_i^{(k)}\int_{-t}^0 1[J_u = k]du) = e^{-\hat{q}_i t}q_i^{(J_{-t})}dt$. Then, using (19),

$$\begin{aligned} I_1(x) &\le \mathbb{E}\left[1[\mathcal{T}_{\lfloor hx^\alpha\delta\rfloor} > 2hx^\alpha]\int_0^\infty \hat{f}(t)dt\right] + \mathbb{E}\int_0^{2hx^\alpha} \mathbb{P}_{\sigma_t}[S(t;J) \ge x]dt \\ &\le \mathbb{P}[\mathcal{T}_{\lfloor hx^\alpha\delta\rfloor} > 2hx^\alpha] + 2hx^\alpha\mathbb{P}[\bar{S}(2hx^\alpha) \ge x], \end{aligned} \tag{20}$$

where $\bar{S}(t) \triangleq \sum_{i\ge 1}\bar{B}_i$, and $\bar{B}_i$, $i \ge 1$, are independent Bernoulli random variables with $\mathbb{P}[\bar{B}_i = 1] = 1 - e^{-\bar{q}_i t}$, $\bar{q}_i \triangleq q_i/(\min_k \pi_k)$, similarly as in (28) of [13]. Now, by Lemmas 2 and 3, the last term of (20) is $o(1/x^{\alpha-1})$.

Next, after using the Chernoff bound for the sum of exponential i.i.d. random variables, we obtain for any $\delta > \theta > 0$,

$$\mathbb{P}[\mathcal{T}_{\lfloor hx^\alpha \delta \rfloor} > 2hx^\alpha] \leq e^{-\theta 2hx^\alpha} e^{-\lfloor hx^\alpha \delta \rfloor \log(1-\frac{\theta}{\delta})}$$
$$\leq e^{-\frac{1}{2}h\frac{x^\alpha \delta}{\log x} \log x} = o\left(\frac{1}{x^{\alpha-1}}\right) \quad \text{as} \quad x \to \infty, \tag{21}$$

since $x^\alpha \delta / \log x \to \infty$ as $x \to \infty$, which follows directly from the assumption $x\delta^{1/\alpha}/\log x \to \infty$ as $\delta \to 0$. Finally, using the preceding estimates in (20), we have proved (18).

In order to estimate $I_2(x)$, we define the set $\mathcal{A}(n)$ as

$$\mathcal{A}(n) \triangleq \cap_{1 \leq k \leq M} \left\{ \left| \tau_k(\mathcal{T}_n) - \frac{\pi_k(n+1)}{\delta} \right| \leq 2\epsilon \frac{\pi_k(n+1)}{\delta} \right\}, \tag{22}$$

where $\tau_k(\mathcal{T}_n)$ represents the total time that process J spends in state k in the interval $(-\mathcal{T}_n, 0)$. Next, due to the memoryless property of the exponential distribution, note that $\tau_k(\mathcal{T}_n) \stackrel{d}{=} \sum_{i=0}^{N_n(k)} \epsilon_i$, where $N_n(k)$ is equal to the number of times that the Markov chain $\{J_{-\mathcal{T}_i}\}$ visits state k and ϵ_i are exponential i.i.d. random variables with mean $1/\delta$, both for $0 \leq i \leq n$. Then,

$$\mathbb{P}\left[\tau_k(\mathcal{T}_n) > (1+\epsilon)\frac{\pi_k(n+1)}{\delta}\right] \leq \mathbb{P}[N_n(k) \geq (1+\epsilon)\pi_k(n+1)]$$
$$+\mathbb{P}\left[\sum_{i=0}^{\lceil (1+\epsilon)\pi_k(n+1)\rceil} \epsilon_i > (1+2\epsilon)\frac{\pi_k(n+1)}{\delta}\right]. \tag{23}$$

Next, note that for any $0 < \theta < \delta$ and any positive integer n

$$\mathbb{P}\left[\sum_{i=1}^n \epsilon_i > (1+\epsilon)\frac{n}{\delta}\right] = \mathbb{P}\left[e^{\theta \sum_{i=1}^n \epsilon_i} > e^{\theta(1+\epsilon)\frac{n}{\delta}}\right] \leq e^{-n[\frac{\theta}{\delta}(1+\epsilon)+\log(1-\frac{\theta}{\delta})]},$$

where in the last expression we applied the Markov inequality. Therefore, by setting $u = \theta/\delta$ in the preceding expression,

$$\mathbb{P}\left[\sum_{i=1}^n \epsilon_i > (1+\epsilon)\frac{n}{\delta}\right] \leq \inf_{0<u<1} e^{-n[u(1+\epsilon)+\log(1-u)]} = e^{-n(\epsilon+\log(1+\epsilon))}, \tag{24}$$

where the minimum is achieved for $u = \epsilon/(1+\epsilon)$. Then, after applying a well-known large deviation result on finite-state ergodic Markov chains (e.g., see Section 3.1.2 of [6]) to bound the first term of (23) and using (24), we conclude that there exists a constant $\theta_k(\epsilon) > 0$, independent of δ, which satisfies

$$\mathbb{P}\left[\tau_k(\mathcal{T}_n) > (1+\epsilon)\frac{\pi_k(n+1)}{\delta}\right] \leq e^{-\theta_k(\epsilon)n}. \tag{25}$$

Using arguments analogous to those in (23), (24) and (25) for estimating the exponential upper bound for $\mathbb{P}\left[\tau_k(\mathcal{T}_n) < (1-\epsilon)\frac{\pi_k(n+1)}{\delta}\right]$, in conjunction with the union bound, we conclude

$$\mathbb{P}[\mathcal{A}^c(n)] \leq \max_k \mathbb{P}\left[\left|\tau_k(\mathcal{T}_n) - \frac{\pi_k(n+1)}{\delta}\right| > 2\epsilon\frac{\pi_k(n+1)}{\delta}\right] \leq He^{-\theta_\epsilon n}, \tag{26}$$

for some positive constant $\theta_\epsilon > 0$, independent of δ.

At this point, we are ready to proceed with estimating the integral $I_2(x)$. After multiplying $I_2(x)$ with $1[\mathcal{A}(n)]$ and $1[\mathcal{A}^c(n)]$, we define

$$\begin{aligned} I_2(x) &\leq \mathbb{E}\sum_{n=\lfloor hx^\alpha\delta\rfloor}^{\infty}\int_{\mathcal{T}_n}^{\mathcal{T}_{n+1}} \hat{f}(t)\mathbb{P}_{\sigma_t}[S(t;J)\geq x]dt \\ &= \mathbb{E}\sum_{n=\lfloor hx^\alpha\delta\rfloor}^{\infty} 1[\mathcal{A}^c(n)]\int_{\mathcal{T}_n}^{\mathcal{T}_{n+1}} \hat{f}(t)\mathbb{P}_{\sigma_t}[S(t;J)\geq x]dt \\ &\qquad + \mathbb{E}\sum_{n=\lfloor hx^\alpha\delta\rfloor}^{\infty} 1[\mathcal{A}(n)]\int_{\mathcal{T}_n}^{\mathcal{T}_{n+1}} \hat{f}(t)\mathbb{P}_{\sigma_t}[S(t;J)\geq x]dt \\ &\triangleq I_{21}(x) + I_{22}(x). \end{aligned} \tag{27}$$

Then, by using (26), we obtain

$$I_{21}(x) \leq \sum_{n=\lfloor hx^\alpha\delta\rfloor}^{\infty}\mathbb{P}[\mathcal{A}^c(n)] \leq He^{-\theta_\epsilon hx^\alpha\delta} = o\left(\frac{1}{x^{\alpha-1}}\right) \text{ as } x\to\infty. \tag{28}$$

Next, we estimate $I_{22}(x)$. Since $S(t;J)$ is a.s. non-increasing in t, after splitting the sum we obtain

$$\begin{aligned} I_{22}(x) &\leq \mathbb{E}\sum_{n=\lfloor hx^\alpha\delta\rfloor}^{\lfloor g_\epsilon x^\alpha\delta\rfloor} 1[\mathcal{A}(n)]\int_{\mathcal{T}_n}^{\mathcal{T}_{n+1}} \hat{f}(t)\mathbb{P}_{\sigma_{\mathcal{T}_{n+1}}}[S(\mathcal{T}_{n+1};J)\geq x]dt \\ &\qquad + \mathbb{E}\sum_{n=\lfloor g_\epsilon x^\alpha\delta\rfloor+1}^{\infty} 1[\mathcal{A}(n)]\int_{\mathcal{T}_n}^{\mathcal{T}_{n+1}} \hat{f}(t)dt, \end{aligned} \tag{29}$$

where g_ϵ will be defined later. Note that for every $\omega\in\mathcal{A}(n)$ and k, $1\leq k\leq M$,

$$(1-2\epsilon)\pi_k\frac{n+1}{\delta} \leq \tau_k(\mathcal{T}_n) \leq (1+2\epsilon)\pi_k\frac{n+1}{\delta}. \tag{30}$$

Therefore, by definition (2),

$$\mathbb{P}_{\sigma_{\mathcal{T}_n}}[B_i(\mathcal{T}_n;J)=1] = 1 - e^{-\sum_{k=1}^M q_i^{(k)}\tau_k(\mathcal{T}_n)} \leq \mathbb{P}[B_i^*(n)=1],$$

and

$$\mathbb{P}_{\sigma_{\mathcal{T}_n}}[S(\mathcal{T}_n; J) \geq x] \leq \mathbb{P}[S^*(n) \geq x], \tag{31}$$

where we define $S^*(n) \triangleq \sum_{i=1}^{\infty} B_i^*(n)$ with $\{B_i^*(n), i \geq 1\}$ representing a sequence of independent Bernoulli random variables and $\mathbb{P}[B_i^*(n) = 1] = 1 - e^{-(1+2\epsilon)q_i(n+1)/\delta}$; $S^*(n)$ is constructed to be non-decreasing in n. Then, if we pick g_ϵ to be

$$g_\epsilon \triangleq \frac{(1-2\epsilon)^\alpha}{\left[\Gamma\left(1-\frac{1}{\alpha}\right)\right]^\alpha c(1+2\epsilon)},$$

using the analogous arguments as in (66) - (67) of [13], we conclude that, for any large x, $\mathbb{E}S^*(g_\epsilon x^\alpha \delta) < (1-\epsilon)x$ and, therefore, by Lemma 3,

$$\begin{aligned}\mathbb{E} \sum_{n=\lfloor hx^\alpha\delta \rfloor}^{\lfloor g_\epsilon x^\alpha \delta \rfloor} 1[\mathcal{A}(n)] \int_{\mathcal{T}_n}^{\mathcal{T}_{n+1}} \hat{f}(t)\mathbb{P}_{\sigma_{\mathcal{T}_{n+1}}}[S(\mathcal{T}_{n+1}; J) \geq x]dt& \\ \leq g_\epsilon x^\alpha \delta \mathbb{P}[S^*(g_\epsilon x^\alpha \delta) \geq x] = o\left(\frac{1}{x^{\alpha-1}}\right).&\end{aligned} \tag{32}$$

Next, we derive the asymptotics of the second term in (29). Note that for every $\omega \in \mathcal{A}(n)$ and $t \in (\mathcal{T}_n, \mathcal{T}_{n+1}]$, the bound from (30) results in

$$\hat{f}(t) \leq \sum_{i=1}^{\infty} q_i^{(J_0)} q_i^{(\mathcal{J}_{-n-1})} e^{-(1-2\epsilon)q_i \frac{n}{\delta}} e^{-q_i^{(\mathcal{J}_{-n-1})}(t-\mathcal{T}_n)}; \tag{33}$$

recall that $\mathcal{T}_n = -T_{-n}$ from above and that $J_{-\mathcal{T}_{n+1}} = J_{T_{-n-1}} = \mathcal{J}_{-n-1}$ from Section 2. Using the preceding bound in the second term of (29), computing the integration with respect to t and applying $1 - e^{-x} \leq x$, $x \geq 0$, we obtain

$$\begin{aligned}\mathbb{E} \sum_{n=\lfloor g_\epsilon x^\alpha \delta \rfloor+1}^{\infty} 1[\mathcal{A}(n)] \int_{\mathcal{T}_n}^{\mathcal{T}_{n+1}} \hat{f}(t)dt &\leq \mathbb{E} \sum_{n=\lfloor g_\epsilon x^\alpha \delta \rfloor+1}^{\infty} 1[\mathcal{A}(n)] \sum_{i=1}^{\infty} q_i^{(J_0)} q_i^{(\mathcal{J}_{-n-1})} (\mathcal{T}_{n+1} - \mathcal{T}_n) e^{-(1-2\epsilon)q_i \frac{n}{\delta}} \\ &\leq \sum_{n=\lfloor g_\epsilon x^\alpha \delta \rfloor+1}^{\infty} \sum_{i=1}^{\infty} \frac{1}{\delta} \mathbb{E}[q_i^{(J_0)} \mathbb{E}[q_i^{(\mathcal{J}_{-n-1})} | J_0]] e^{-nq_i(1-2\epsilon)/\delta}.\end{aligned} \tag{34}$$

Now, in the last expression, we employ the asymptotic independence of the Markov chain J_n, similarly as in (19) - (20) of [13], and the independence between $\{\mathcal{T}_n\}$ and $\{\mathcal{J}_{-n}\}$, and then we bound the resulting sum by an integral with the change of variable $t = n/\delta$, which, in conjunction with (32), yields

$$I_{22}(x) \leq (1+\epsilon) \sum_{n=\lfloor g_\epsilon x^\alpha \delta \rfloor + 1}^{\infty} \sum_{i=1}^{\infty} q_i^2 \frac{1}{\delta} e^{-(1-2\epsilon) q_i \frac{n}{\delta}} + o\left(\frac{1}{x^{\alpha-1}}\right)$$
$$\leq (1+\epsilon) \int_{g_\epsilon x^\alpha}^{\infty} \sum_{i=1}^{\infty} q_i^2 e^{-(1-2\epsilon) q_i t} dt + o\left(\frac{1}{x^{\alpha-1}}\right).$$

Finally, by applying Lemma 1, we derive

$$\limsup_{x\to\infty} I_{22}(x) x^{\alpha-1} \leq K(\alpha) \frac{(1+\epsilon)^2 (1+2\epsilon)^{1-\frac{1}{\alpha}}}{(1-2\epsilon)^{1+\alpha-\frac{1}{\alpha}}}, \tag{35}$$

which by passing $\epsilon \to 0$, in conjunction with (29), (28), (27) and (17), proves the upper bound.

The estimation of the lower bound of (4) starts from

$$\mathbb{P}[C > x] \geq \mathbb{E} \sum_{n=\lfloor g_\epsilon x^\alpha \delta \rfloor}^{\infty} 1[\mathcal{A}(n)] \int_{\mathcal{T}_n}^{\mathcal{T}_{n+1}} \hat{f}(t) \mathbb{P}_{\sigma_{\mathcal{T}_n}} [S(\mathcal{T}_n; J) \geq x] dt, \tag{36}$$

where $g_\epsilon \triangleq (1+2\epsilon)^\alpha [\Gamma[1-\frac{1}{\alpha}]]^{-\alpha} c^{-1} (1-\epsilon)^{-1}$. Using analogous arguments to those in obtaining (31), with redefined $\mathbb{P}[B_i^*(n) = 1] = 1 - e^{-(1-2\epsilon) q_i (n+1)/\delta}$, $i \geq 1$, we obtain

$$\mathbb{E} \sum_{n=\lfloor g_\epsilon x^\alpha \delta \rfloor}^{\infty} 1[\mathcal{A}(n)] \int_{\mathcal{T}_n}^{\mathcal{T}_{n+1}} \hat{f}(t) \mathbb{P}_{\sigma_{\mathcal{T}_n}} [S(\mathcal{T}_n; J) \geq x] dt$$
$$\geq \mathbb{P}[S^*(g_\epsilon x^\alpha \delta) > x] \mathbb{E} \sum_{n=\lfloor g_\epsilon x^\alpha \delta \rfloor}^{\infty} 1[\mathcal{A}(n)] \int_{\mathcal{T}_n}^{\mathcal{T}_{n+1}} \hat{f}(t) dt.$$

We then complete the proof of this theorem by applying arguments analogous to those used in (33) - (34) to lower bound $\hat{f}(t)$ and, therefore, the integral on the right hand side of (36) for all $\omega \in \mathcal{A}(n)$, $t \in [\mathcal{T}_n, \mathcal{T}_{n+1})$. Then, in conjunction with asymptotic independence and bounding arguments analogous to those used in (19)-(21) of [13] for all δ small enough, one can easily complete the proof of the lower bound, the details of which are omitted since the arguments are repetitive.

4.3. Discussion

Note that when $x < 1/\delta^p$ for some $p > 0$, the condition $x\delta^{1/\alpha}/\log x \to \infty$ of Theorem 2 is implied by $x\delta^{1/\alpha}/\log(1/\delta) \to \infty$. Thus, for H large enough and for all $x > H\log(1/\delta)/\delta^{1/\alpha}$, the cache behaves as the corresponding i.i.d. system with marginal distribution $\{q_i\}$. Hence, under this asymptotic scaling, the correlation structure plays no role. On the other hand, Theorem 1 states that for very small caches, $x \leq 1/(H\delta^{1/\alpha})$, the cache performance is distinctly different from that of the corresponding i.i.d. system; in fact, the fault probability is decomposed into a mixture of i.i.d. systems. Informally, we see that this qualitative transition in the cache performance occurs around cache sizes on the order of $1/\delta^{1/\alpha}$. As previously noted, this value is sublinear (relatively negligible) in comparison to the time scale of jumps $(1/\delta)$ in the modulating process J.

In order to gain additional insights into the qualitative behavior underlying Theorems 1 and 2, consider the expected time between two successive requests for a document during which the underlying Markov chain J_t is in a fixed state. Then, the expected length of this time interval is inversely proportional to the document's conditional access frequency, and thus the LRU algorithm has a tendency, in stationarity, to arrange the documents in the cache list in (approximately) descending order of their access probabilities. Therefore, it can be intuitively expected that the access probabilities of documents at the end of the cache list are on the order of $x^{-\alpha}$, which from the above arguments implies that the time period during which every document in the cache is accessed at least once is of the order x^{α}. Hence, if the expected sojourn time that the modulated process spends in a particular state, $1/\delta$, is much greater than x^{α}, i.e., $1/\delta \gg x^{\alpha}$, then the cache content basically goes through many replacement cycles and the cache essentially reaches stationarity while the underlying modulating process remains in the same state, resulting in the decomposition result presented in Theorem 1. On the other hand, if $1/\delta \ll x^{\alpha}$, then the Markov chain J_t undergoes significant mixing between the successive requests for documents that are at the end of the cache list. Therefore, successive requests for documents with probabilities smaller than $1/x^{\alpha}$, which essentially determine the cache fault probability, appear nearly independent, implying that the cache fault probability is the same as if the requests were i.i.d., as we have rigorously shown in Theorem 2.

In the critical regime, when $1/\delta \approx x^{\alpha}$, the above arguments suggest that the time scale of jumps in J_t and the access frequencies of documents at the end of the cache list are comparable. Therefore, the cache fault probability is a result of an intricate and complex interplay between the modulating chain dependence structure and the conditional access frequencies. While deeper understanding of this critical regime is important for a complete mathematical understanding of the problem, we strongly believe that the potential asymptotic results will likely be more difficult to prove and will not be explicit, as those in Theorems 1 and 2. Rather, even asymptotically, the cache fault probability in the critical regime will be a complex functional of the transition probabilities of J_t and the conditional access frequencies $q_i^{(k)}$ whose further understanding will require numerical studies. Therefore, we do not pursue this direction further.

5. Concluding remarks

In this paper we investigate the performance, namely fault probability, of LRU caches in the presence of correlated requests. It has been recently discovered in [12, 13] that, for the semi-Markov modulated requests and generalized Zipf's law marginal access frequencies, the caching performance does not depend on the correlation in the request traffic for large cache sizes. Specifically, LRU cache performance is asymptotically identical to the case of i.i.d. requests that have the same access frequences. However, for small caches this is clearly not the case. Hence, in our present study we investigate the smallest (critical) cache size above which the discovered asymptotic insensitivity property still holds. We answer this question based on the use of a joint scaling between the request process dependence structure and the cache size. More precisely, we consider requests that are modulated by NCD Markov processes with small transition rate δ such that the cache size x_δ grows to infinity as $\delta \downarrow 0$. Then, extending the analytical techniques from [12, 13], we show in Theorems 1 and 2 that, maybe somewhat surprisingly, the critical scaling of the cache size x_δ is very small in relation to the time scale of the request process dependence structure; basically it is sublinear in the average time $1/\delta$ between the jumps in the modulating process. Hence, from a practical perspective, it may not be necessary to model in great detail the request process dependence structure found in Web environments. In addition, it is worth noting that our results can be extended to the case with variable document sizes by exploiting the recent analysis in [14].

Acknowledgments

We are very grateful to the anonymous reviewer for valuable suggestions.

References

[1] Almeida, V., Bestavros, A., Crovella, M. and de Oliviera, A. (1996). Characterizing reference locality in the WWW. In *Proceedings of the Fourth International Conference on Parallel and Distributed Information Systems*. Miami Beach, Florida.

[2] ASMUSSEN, S., HENRIKSEN, L. F. AND KLÜPPELBERG, C. (1994). Large claims approximations for risk processes in a Markovian environment. *Stochastic Processes and their Applications* **54,** 29–43.

[3] BENTLEY, J. L. AND MCGEOCH, C. C. (1985). Amortized analysis of self-organizing sequential search heuristics. *Communications of the ACM* **28,** 404–411.

[4] BORODIN, A. AND EL-YANIV, R. (1998). *Online Computation and Competitive Analysis*. Cambridge University Press.

[5] BRESLAU, L., CAO, P., FAN, L., PHILLIPS, G. AND SHENKER, S. (1999). Web caching and Zipf-like distributions: Evidence and implications. In *IEEE INFOCOM*.

[6] DEMBO, A. AND ZEITOUNI, O. (1993). *Large Deviations Techniques and Applications*. Jones and Bartlett Publishers.

[7] FILL, J. A. (1996). An exact formula for the move-to-front rule for self-organizing lists. *Journal of Theoretical Probability* **9,** 113–159.

[8] FILL, J. A. AND HOLST, L. (1996). On the distribution of search cost for the move-to-front rule. *Random Structures and Algorithms* **8,** 179.

[9] FLAJOLET, P., GARDY, D. AND THIMONIER, L. (1992). Birthday paradox, coupon collector, caching algorithms and self-organizing search. *Discrete Applied Mathematics* **39,** 207–229.

[10] JELENKOVIĆ, P. R. (1999). Asymptotic approximation of the move-to-front search cost distribution and least-recently-used caching fault probabilities. *Annals of Applied Probability* **9,** 430–464.

[11] JELENKOVIĆ, P. R. AND LAZAR, A. A. (1998). Subexponential asymptotics of a Markov-modulated random walk with queueing applications. *Journal of Applied Probability* **35,** 325–347.

[12] JELENKOVIĆ, P. R. AND RADOVANOVIĆ, A. (2003.). Asymptotic Insensitivity of Least-Recently-Used Caching to Statistical Dependency. In *Proceedings of INFOCOM 2003*. San Francisco.

[13] JELENKOVIĆ, P. R. AND RADOVANOVIĆ, A. (2004). Least-Recently-Used Caching with Dependent Requests. *Theoretical Computer Science* **326,** 293–327.

[14] JELENKOVIĆ, P. R. AND RADOVANOVIĆ, A. (2004). Optimizing LRU Caching for Variable Document Sizes. *Combinatorics, Probability* & *Computing* **13,** 1–17.

[15] SLEATOR, D. D. AND TARJAN, R. E. (1985). Self-adjusting binary search trees. *Journal of the ACM* **32,** 652–686.

[16] YOUNG, N. E. (2000). On-line paging against adversarially biased random inputs. *J. Algorithms* **37,** 218–235.