RC23779 (W0511-041) November 4, 2005
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

Critical Sizing of LRU Caches with Dependent Requests

Predrag R. Jelenković
Department of Electrical Engineering
Columbia University
New York, NY 10027, USA
predrag@ee.columbia.edu

Ana Radovanović and Mark S. Squillante
Mathematical Sciences Department
IBM Thomas J. Watson Research Center
Yorktown Heights, NY 10598
{aradovan, mss}@us.ibm.com

November 1, 2005

Abstract

It was recently proved in [12] 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 preceding result is asymptotic, it remains unclear how small the cache size can become while still retaining this insensitivity property. In this paper, assuming requests come from a nearly completely decomposable Markov-modulated process, we characterize the critical cache size below which the dependency of requests dominates the cache performance. It appears surprising that the critical cache size is very small, and in fact it is sub-linear to the sojourn times of the modulated chain that determines the dependency structure.

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

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 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 contents 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 high hit ratio, low complexity, and flexibility to dynamically adapt to possible changes in the request patterns.

Due to its importance, the LRU caching has received a significant attention from the literature, both from the context of the combinatorial (worst-case) [2, 3, 13, 14] and probabilistic (average-case) analysis; the letter 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, 4, 11] 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 in [12] we develop a novel, analytically explicit method for analyzing LRU caches in the presence of semi-Markov modulated requests. This way of modeling dependency provides the flexibility for capturing possibly strong statistical correlation, including the widely reported long-range dependence of the access patterns for Web documents. The main results from [12, 11] imply that asymptotically, for large cache sizes, the cache fault probability behaves the same as in the corresponding LRU system with i.i.d. requests [10]. This surprising insensitivity was further validated experimentally in [12, 11] where we found an excellent agreement between the asymptotic results and those from simulation, even for relatively small cache sizes.

Since the results from [12] are asymptotic, they do not provide information on how small the cache sizes can become while still retaining the recently 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. 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 (locality) dominates the cache performance. It is somewhat unexpected that this critical cache size is very small in comparison to the time scale of transitions in the modulating process; in fact, it is sub-linear in $1/\delta$.

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 2 and 3 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 the 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 the 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 [9], [6], and [10]. We therefore proceed with a description of the MTF algorithm.

More formally, consider a finite set of documents $L = \{1, \ldots, N\}$, 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 taking on values in $\{1, \ldots, M\}$ and independent of N^δ. 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 $q_i > 0$, $1 \leq i \leq N$.

3 Preliminary results

The model described in the previous section is a special case of the more general one introduced in [12] and, therefore, the results established therein hold and some of them are applied in this paper to prove our main results. In particular, Lemma 1 of [12] 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. For the reason of completeness, we state this result below. Then, in the following subsection we provide results that characterize the tail of the stationary search cost distribution and 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 [12] and will be used in proving our main theorems.

Lemma 1 *If the request process $\{R_n\}$ is stationary and ergodic, then for any initial permutation Π_0 of the list, the search cost C_n^N converges in distribution to C^N as $n \to \infty$, where*

$$C^N \triangleq \sum_{i=1}^{N} \sum_{m=1}^{\infty} (1 + S_i(m-1)) 1[R_{-m} = i, \mathcal{R}_i(m-1), R_0 = i],$$

$S_i(m)$ is the number of distinct documents, different from i, among $R_{-m}, \ldots, R_{-1}$, and event $\mathcal{R}_i(m) \triangleq \{R_{-j} \neq i, 1 \le j \le m\}$, $m \ge 1$; $S_i(0) \equiv 0$, $\mathcal{R}_i(0) \equiv \Omega$ (where Ω represents a sample space).

3.1 Representation results

In this subsection we state the representation result for the stationary search cost C^N that was derived in [12] and represents the starting point for our analysis. Before stating the theorem we introduce the necessary notation. Let σ_t be the σ-algebra $\sigma(J_u, -t \le u \le 0)$ 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 $[-u, 0)$, $0 < u \le t$, and define an indicator function $B_j(t; J) = 1[N_j(t; J) > 0]$, $j \ge 1$, 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

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

where

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

Empirical request probabilities $\hat{q}_j \equiv \hat{q}_j(t)$, $j \ge 1$, and probabilities $\hat{\pi}_k \triangleq \hat{\pi}_k(t)$, $1 \le k \le M$, are defined as

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

Next, we state the main representation theorem.

Theorem 1 *The stationary distribution of the searching cost C^N satisfies*

$$\mathbb{P}[C^N > x] = \mathbb{E} \int_0^{\infty} \sum_{i=1}^{N} q_i^{(J_0)} q_i^{(J_{-t})} e^{-\hat{q}_i t} \mathbb{P}_{\sigma_t}[S_i(t; J) > x - 1] dt, \tag{4}$$

with $S_i(t; J)$, $B_i(t; J)$ and $\hat{q}_i$ satisfying equations (1), (2) and (3), respectively.

In the proposition that follows we investigate the limiting search cost distribution when the number of items $N \to \infty$. Now, assume that the probability mass functions $q_i^{(k)}, 1 \le k \le M$ are defined for all $i \ge 1$. Using these probabilities, for a given modulating process J and each $1 \le N \le \infty$ we define a sequence of request processes $\{R_n^N\}$, whose conditional request probabilities are equal to

$$q_{i,N}^{(k)} = \frac{q_i^{(k)}}{\sum_{i=1}^{N} q_i^{(k)}}, \quad 1 \le i \le N;$$

then, for each finite N, let C^N be the corresponding stationary search cost. In the case of the limiting request process $R_n = R_n^\infty$, similarly as in (1), introduce $S_i(t;J) = \sum_{j \ne i} B_j(t;J)$ to be equal to the number of different items, not equal to i, that are requested in $[-t,0)$; $B_j(t;J)$ is the Bernoulli variable representing the event that item j was requested at least once in $[-t,0)$. Then, the following proposition states the convergence of the stationary search cost C^N in distribution as $N \to \infty$ and provides the representation formula (5) used in our analysis.

Proposition 1 *The constructed sequence of stationary search costs C^N converges in distribution to C as $N \to \infty$, where the distribution of C is given by*

$$\mathbb{P}[C > x] \; = \; \mathbb{E} \int_0^\infty \hat{f}(t) \mathbb{P}_{\sigma_t}[S_i(t;J) > x-1] \, dt, \tag{5}$$

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

$$\hat{f}(t) \triangleq \sum_{i=1}^{\infty} q_i^{(J_0)} q_i^{(J_{-t})} e^{-\hat{q}_i t}. \tag{6}$$

Remark 1 Throughout this paper we will exploit the properties that the variables $S_j(t;J)$, $B_j(t;J)$, $j \ge 1$, are monotonically increasing in t and that the variables $B_j(t;J)$, $j \ge 1$, are conditionally independent given σ_t. This conditional independence arises from the Poisson arrival structure, as is apparent from the derivation in [12]. In general, when the request times are not Poisson, e.g., discrete-time arrivals, these variables may not be conditionally independent. For i.i.d. requests, the Poisson embedding technique was first introduced in [8]. ◇

Remark 2 It is clear that the derivation of the above results does not rely on the fact that the requests arrive at a constant rate [12]. 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. ◇

Remark 3 For the i.i.d. case, Proposition 1 was proved in Proposition 4.4 of [7]. ◇

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 follows from our general problem formulation when the modulating process is assumed to be a constant, i.e., $J_t \equiv \text{constant}$. 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 Proposition 1 reduces to

$$\mathbb{P}[C > x] \ = \ \int_0^\infty \sum_{i=1}^\infty q_i^2 e^{-q_i t} \mathbb{P}[S_i(t) > x - 1] dt. \tag{7}$$

The following two results, originally proved in Lemmas 1 and 2 of [10], are restated here for convenience. 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.

Lemma 2 *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 3 *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 $t \to \infty$,*

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

The next two lemmas are originally proved in [12]. They are repeatedly used in establishing our main results. 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.

Lemma 4 *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 5 *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 2.

Lemma 6 *Let $c_2/i^\alpha \leq q_i \leq c_1/i^\alpha$, $\alpha > 1$, for some positive constants c_1, c_2. 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, the tail of the search cost C can be lower bounded as

$$\begin{aligned}\mathbb{P}[C > x] &= \int_0^\infty \sum_{i=1}^\infty q_i^2 e^{-q_i t} \mathbb{P}[S(t) > x-1] dt \\ &\geq \mathbb{P}[S(Hx^\alpha) > x-1] \int_{Hx^\alpha}^\infty \sum_{i=1}^\infty q_i^2 e^{-q_i t} dt \\ &\geq (1-\epsilon) \sum_{i=1}^\infty q_i e^{-Hx^\alpha q_i}, \end{aligned} \tag{8}$$

where the first inequality follows from the monotonicity of $S(t)$, while the second inequality is obtained by applying Lemmas 3, 4 and integration. Next, from the assumptions of this lemma, we have

$$\begin{aligned}\sum_{i=1}^\infty q_i e^{-Hq_i x^\alpha} &\geq \sum_{\lfloor x \rfloor + 1}^\infty \frac{c_2}{i^\alpha} e^{-H\frac{x^\alpha}{i^\alpha}} \\ &\geq c_2 e^{-H} \int_{x+1}^\infty \frac{1}{u^\alpha} du \\ &\geq \frac{h}{x^{\alpha-1}}, \end{aligned}$$

which in conjunction with (8) proves the result. ◇

4 Main results

In this section we state and prove our main results. We show that depending on the way the cache size x and parameter δ scale, we obtain different performance characteristics with respect to the cache fault probability.

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 2 *Let $q_i \leq c_1/i^\alpha$, $\alpha > 1$, and 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*

$$\mathbb{P}[C > x_\delta] \sim \sum_{k=1}^M \pi_k \mathbb{P}[C^{(k)} > x_\delta] \;\; as \;\; x_\delta \to \infty. \tag{9}$$

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 (5) implies

$$\begin{aligned}
&\mathbb{P}[C > x] = \mathbb{E}\int_0^\infty \hat{f}(t)\mathbb{P}_{\sigma_t}[S_i(t;J) > x-1]dt \\
&\geq \mathbb{E}\int_0^{\mathcal{T}_0} \sum_{i=1}^{\infty} (q_i^{(J_0)})^2 e^{-q_i^{(J_0)}t}\mathbb{P}[S_i^{(J_0)}(t) > x-1|J_0]dt \\
&\geq \sum_{k=1}^{M}\mathbb{P}[J_0 = k, \mathcal{T}_0 > Hx^\alpha]\int_0^\infty \sum_{i=1}^{\infty}(q_i^{(k)})^2 e^{-q_i^{(k)}t}\mathbb{P}[S_i^{(k)}(t) > x-1]dt - \sum_{k=1}^{M}\pi_k\int_{Hx^\alpha}^{\infty}\sum_{i=1}^{\infty}(q_i^{(k)})^2 e^{-q_i^{(k)}t}dt.
\end{aligned} \tag{10}$$

Now, since $q_i^{(k)} \leq q_i/\min_k \pi_k \triangleq \bar{q}_i$, $1 \leq k \leq M$, $q_i \leq c_1/i^\alpha$ and $xe^{-x} \leq e^{-1}$ (for $x \geq 0$), the second summand in (10) can be bounded as

$$\begin{aligned}
\sum_{k=1}^{M}\pi_k\int_{Hx^\alpha}^{\infty}\sum_{i=1}^{\infty}(q_i^{(k)})^2 e^{-q_i^{(k)}t}dt &\leq \sum_{k=1}^{M}\pi_k\frac{1}{Hx^\alpha}\sum_{i=1}^{\lfloor H^{1/\alpha}x\rfloor} q_i^{(k)}Hx^\alpha e^{-q_i^{(k)}Hx^\alpha} + \frac{1}{(\min_k \pi_k)}\int_{H^{1/\alpha}x}^{\infty}\frac{c_1}{y^\alpha}dy \\
&\leq \frac{e^{-1}}{H^{1-1/\alpha}x^{\alpha-1}} + \frac{c_1}{(\min_k \pi_k)H^{1-1/\alpha}(\alpha-1)x^{\alpha-1}} \\
&\leq \frac{1}{H^{\frac{\alpha-1}{2\alpha}}}\frac{1}{x^{\alpha-1}},
\end{aligned} \tag{11}$$

for H large enough. 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 (10) and (11) we obtain

$$\mathbb{P}[C > x] \gtrsim \sum_{k=1}^{M}\pi_k\mathbb{P}[C^{(k)} > x] - \frac{1}{H^{\frac{\alpha-1}{2\alpha}}}\frac{1}{x^{\alpha-1}} \;\text{ as }\; x \to \infty.$$

Next, by applying Lemma 6 and letting $H \to \infty$, we conclude

$$\mathbb{P}[C > x] \gtrsim \sum_{k=1}^{M}\pi_k\mathbb{P}[C^{(k)} > x] \;\text{ as }\; x \to \infty. \tag{12}$$

Next, we prove the upper bound. After splitting the integral in (5), we define

$$\mathbb{P}[C > x] = \mathbb{E}\int_0^{\mathcal{T}_0} + \mathbb{E}\int_{\mathcal{T}_0}^{\infty} \triangleq I_1(x) + I_2(x). \tag{13}$$

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

$$\begin{aligned}
I_1(x) &= \mathbb{E}\sum_{k=1}^{M}1[J_0 = k]\int_0^{\mathcal{T}_0}\sum_{i=1}^{\infty}(q_i^{(k)})^2 e^{-q_i^{(k)}t}\mathbb{P}[S_i^{(k)}(t) > x-1]dt \\
&\leq \sum_{k=1}^{M}\pi_k\int_0^{\infty}\sum_{i=1}^{\infty}(q_i^{(k)})^2 e^{-q_i^{(k)}t}\mathbb{P}[S_i^{(k)}(t) > x-1]dt \\
&= \sum_{k=1}^{M}\pi_k\mathbb{P}[C^{(k)} > x],
\end{aligned} \tag{14}$$

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

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

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

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

$$\sum_{i=1}^{\infty} \frac{q_i^{(k)}}{q_i^{(k)} + \delta} \leq \sum_{i=1}^{\infty} \frac{1}{1 + h\delta i^\alpha} \leq \int_0^\infty \frac{1}{1 + h\delta z^\alpha} dz \leq \frac{1}{(h\delta)^{1/\alpha}} \int_0^\infty \frac{1}{1 + y^\alpha} dy, \tag{15}$$

we obtain

$$I_2(x) \leq H\delta^{1-1/\alpha} = o\left(\frac{1}{x^{\alpha-1}}\right) \quad \text{as} \quad x \to \infty, \tag{16}$$

where the last equality is implied by the assumption of the theorem since $x\delta^{1/\alpha} \to 0$ as $\delta \to 0$ yields $\delta^{1-1/\alpha} = o(1/x^{\alpha-1})$ as $x \to \infty$. Finally, this last observation together with (14) and Lemma 6 imply $I_2(x) = o(I_1(x))$ as $x \to \infty$, which, in conjunction with (13) and (12), concludes the proof of the theorem. $\diamond$

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 3 *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*

$$\mathbb{P}[C > x_\delta] \sim K(\alpha) \mathbb{P}[R > x_\delta] \ \ \text{as} \ \ x_\delta \to \infty, \tag{17}$$

where

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

Γ is the Gamma function.

Proof: Again, to simplify the notation, we set $x \equiv x_\delta$. First we prove the upper bound. After splitting the integral in (5), we define

$$\begin{aligned} \mathbb{P}[C > x] &= \mathbb{E} \int_0^{\mathcal{T}_{\lfloor hx^\alpha \delta \rfloor}} + \mathbb{E} \int_{\mathcal{T}_{\lfloor hx^\alpha \delta \rfloor}}^{\infty} \\ &\triangleq I_1(x) + I_2(x). \end{aligned} \tag{18}$$

Now, we estimate an upper bound for $I_1(x)$. After conditioning on the value of $\mathcal{T}_{\lfloor hx^\alpha\delta\rfloor}$, we obtain

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

where $\hat{f}(t)$ is defined in (6). Note that $\hat{f}(t) \leq \sum_{i=1}^{\infty} q_i^{(J_0)} = 1$ 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),

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

Next, note that the empirical distributions are uniformly bounded by $\hat{q}_i = \sum_{k=1}^M \hat{\pi}_k q_i^{(k)} \leq \sum_{k=1}^M q_i^{(k)} \leq \bar{q}_i \triangleq q_i/\min_k \pi_k < \infty$, since $\min_k \pi_k > 0$. Then, we define independent Bernoulli random variables $\bar{B}_i(t)$, $i \geq 1$, with $\mathbb{P}[\bar{B}_i(t) = 1] = 1 - e^{-\bar{q}_i t}$ and $\bar{S}(t) \triangleq \sum_{i=1}^\infty \bar{B}_i(t)$; $\bar{S}(t)$ is constructed to be non-decreasing in t. Note that for every ω, $\mathbb{P}_{\sigma_t}[B_i(t;J) = 1] \leq \mathbb{P}[\bar{B}_i(t) = 1]$ and, therefore, $\mathbb{P}_{\sigma_t}[S(t;J)\geq x] \leq \mathbb{P}[\bar{S}(t)\geq x]$ uniformly in ω. Thus, we further upper bound the second integral in (20) as

$$2hx^\alpha\mathbb{P}[\bar{S}(2hx^\alpha)\geq x] \leq Hx^\alpha e^{-h\theta_h x^\alpha},$$

where the last inequality follows from Lemmas 3 and 4 for h small enough and some constant $\theta_h > 0$. The preceding inequality and (20) imply, as $x \to \infty$,

$$I_1(x) \leq \mathbb{P}[\mathcal{T}_{\lfloor hx^\alpha\delta\rfloor} > 2hx^\alpha] + o\left(\frac{1}{x^{\alpha-1}}\right). \tag{21}$$

Next, after exploiting the large deviation bound for the sum of exponential i.i.d. random variables, we obtain for some $\theta > 0$,

$$\mathbb{P}[\mathcal{T}_{\lfloor hx^\alpha\delta\rfloor} > 2hx^\alpha] \leq e^{-\theta hx^\alpha\delta} = e^{-\theta h\frac{x^\alpha\delta}{\log x}\log x} = o\left(\frac{1}{x^{\alpha-1}}\right) \text{ as } x\to\infty, \tag{22}$$

since $x^\alpha\delta/\log x \to \infty$ as $x \to \infty$, which can be easily implied from $x\delta^{1/\delta}/\log x \to \infty$ as $\delta \to 0$. Finally, from (21) and (22), we conclude that, as $x \to \infty$,

$$I_1(x) = o\left(\frac{1}{x^{\alpha-1}}\right). \tag{23}$$

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

$$\mathcal{A}(n) \triangleq \cap_{1\leq k\leq M}\left\{\left|\tau_k(\mathcal{T}_n) - \frac{\pi_k(n+1)}{\delta}\right| \leq 2\epsilon\frac{\pi_k(n+1)}{\delta}\right\}, \tag{24}$$

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 $\tau_k(\mathcal{T}_n) \overset{d}{=} \sum_{i=0}^{N_n(k)} \epsilon_i$, 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 $0 \le i \le n$. Then,

$$\mathbb{P}\left[\tau_k(\mathcal{T}_n) > (1+\epsilon)\frac{\pi_k(n+1)}{\delta}\right] \le \mathbb{P}[N_n(k) \ge (1+\epsilon)\pi_k(n+1)] + \mathbb{P}\left[\sum_{i=0}^{\lceil(1+\epsilon)\pi_k(n+1)\rceil} \epsilon_i > (1+2\epsilon)\frac{\pi_k(n+1)}{\delta}\right]. \tag{25}$$

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

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

where in the last expression we applied Markov inequality. Therefore,

$$\mathbb{P}\left[\sum_{i=1}^{n} \epsilon_i > (1+\epsilon)\frac{n}{\delta}\right] \le \inf_{0<u<1} e^{-n[u(1+\epsilon)+\log(1-u)]} = e^{-n(\epsilon+\log(1+\epsilon))}, \tag{26}$$

where u in the previous expression is used instead of θ/δ. Then, after applying a well-known large deviation result on finite state ergodic Markov chains (e.g., see Section 3.1.2 of [5]) to bound the first term of (25) and using (26), we conclude, that there exists a constant $\theta_k(\epsilon) > 0$, independent from δ, that satisfies

$$\mathbb{P}\left[\tau_k(\mathcal{T}_n) > (1+\epsilon)\frac{\pi_k(n+1)}{\delta}\right] \le e^{-\theta_k(\epsilon)n}. \tag{27}$$

Using analogous arguments to the ones in (25), (26) and (27), 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

$$\mathbb{P}[\mathcal{A}^c(n)] \le \max_k \mathbb{P}\left[\left|\tau_k(\mathcal{T}_n) - \frac{\pi_k(n+1)}{\delta}\right| > 2\epsilon\frac{\pi_k(n+1)}{\delta}\right] \le He^{-\theta_\epsilon n}, \tag{28}$$

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

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

$$\begin{aligned} I_2(x) &\le \mathbb{E}\sum_{n=\lfloor hx^\alpha\delta\rfloor}^{\infty}\int_{\mathcal{T}_n}^{\mathcal{T}_{n+1}} \hat{f}(t)\mathbb{P}_{\sigma_t}[S(t;J)\ge x]dt \\ &= \mathbb{E}\sum_{n=\lfloor hx^\alpha\delta\rfloor}^{\infty} 1[\mathcal{A}^c(n)]\int_{\mathcal{T}_n}^{\mathcal{T}_{n+1}} \hat{f}(t)\mathbb{P}_{\sigma_t}[S(t;J)\ge x]dt + \mathbb{E}\sum_{n=\lfloor hx^\alpha\delta\rfloor}^{\infty} 1[\mathcal{A}(n)]\int_{\mathcal{T}_n}^{\mathcal{T}_{n+1}} \hat{f}(t)\mathbb{P}_{\sigma_t}[S(t;J)\ge x]dt \\ &\triangleq I_{21}(x) + I_{22}(x). \end{aligned} \tag{29}$$

Then, by using (28), we obtain

$$I_{21}(x) \le \sum_{n=\lfloor hx^\alpha\delta\rfloor}^{\infty} \mathbb{P}[\mathcal{A}^c(n)] \le He^{-\theta_\epsilon hx^\alpha\delta} = o\left(\frac{1}{x^{\alpha-1}}\right) \text{ as } x\to\infty. \tag{30}$$

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

$$I_{22}^{(2)}(x) \leq \mathbb{E}\sum_{n=\lfloor hx^{\alpha}\delta\rfloor}^{\lfloor g_{\epsilon}x^{\alpha}\delta\rfloor} 1[\mathcal{A}(n)]\int_{\mathcal{T}_n}^{\mathcal{T}_{n+1}} \hat{f}(t)\mathbb{P}_{\sigma_{\mathcal{T}_{n+1}}}[S(\mathcal{T}_{n+1};J)\geq x]dt + \mathbb{E}\sum_{n=\lfloor g_{\epsilon}x^{\alpha}\delta\rfloor+1}^{\infty} 1[\mathcal{A}(n)]\int_{\mathcal{T}_n}^{\mathcal{T}_{n+1}} \hat{f}(t)dt, \tag{31}$$

where g_{ϵ} will be defined later. We define $S^{*}(n) = \sum_{i=1}^{\infty} B_i^{*}(n)$, where $\{B_i^{*}(n), i\geq 1\}$ is a sequence of independent Bernoulli random variables with $\mathbb{P}[B_i^{*}(n)=1] = 1-e^{-(1+2\epsilon)q_i(n+1)/\delta}$; $S^{*}(n)$ is constructed to be non-decreasing in n. Then, for every $\omega\in\mathcal{A}(n)$,

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

and, therefore, by stochastic dominance

$$\mathbb{P}_{\sigma_{\mathcal{T}_n}}[S(\mathcal{T}_n;J)\geq x] \leq \mathbb{P}[S^{*}(n)\geq x]. \tag{32}$$

Now, using (19) and the monotonicity of $S^{*}(n)$ in n, we upper-bound the first term in (31) as

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

Finally, if we pick g_{ϵ} to be

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

and then apply Lemmas 3 and 4 to estimate the upper bound for the first term in (31), we derive

$$\mathbb{E}\sum_{n=\lfloor hx^{\alpha}\delta\rfloor}^{\lfloor g_{\epsilon}x^{\alpha}\delta\rfloor} 1[\mathcal{A}(n)]\int_{\mathcal{T}_n}^{\mathcal{T}_{n+1}} \hat{f}(t)\mathbb{P}_{\sigma_{\mathcal{T}_{n+1}}}[S(\mathcal{T}_{n+1};J)\geq x]dt = o\left(\frac{1}{x^{\alpha-1}}\right) \text{ as } x\to\infty. \tag{33}$$

Next, we derive the asymptotics of the second term in (31). Note that for every $\omega\in\mathcal{A}(n)$ and $t\in(\mathcal{T}_n,\mathcal{T}_{n+1}]$

$$\begin{aligned}\hat{f}(t) &= \sum_{i=1}^{\infty} q_i^{(J_0)} q_i^{(J-\mathcal{T}_{n+1})} e^{-q_i^{(1)}\tau_1(\mathcal{T}_n)-\cdots-q_i^{(M)}\tau_M(\mathcal{T}_n)} e^{-q_i^{(J-\mathcal{T}_{n+1})}(t-\mathcal{T}_n)}\\ &\leq \sum_{i=1}^{\infty} q_i^{(J_0)} q_i^{(J-\mathcal{T}_{n+1})} e^{-(1-2\epsilon)q_i\frac{n}{\delta}} e^{-q_i^{(J-\mathcal{T}_{n+1})}(t-\mathcal{T}_n)}\end{aligned} \tag{34}$$

and, therefore, after integration with respect to t and applying the bound $1-e^{-x}\leq x$, $x\geq 0$, we obtain that the second term in (31) can be upper bounded as

$$\begin{aligned}&\mathbb{E}\sum_{n=\lfloor g_{\epsilon}x^{\alpha}\delta\rfloor+1}^{\infty} 1[\mathcal{A}(n)]\int_{\mathcal{T}_n}^{\mathcal{T}_{n+1}} \hat{f}(t)dt \leq \mathbb{E}\sum_{n=\lfloor g_{\epsilon}x^{\alpha}\delta\rfloor+1}^{\infty} 1[\mathcal{A}(n)]\sum_{i=1}^{\infty} q_i^{(J_0)} q_i^{(J-\mathcal{T}_{n+1})}(\mathcal{T}_{n+1}-\mathcal{T}_n)e^{-(1-2\epsilon)q_i\frac{n}{\delta}}\\ &\leq \sum_{n=\lfloor g_{\epsilon}x^{\alpha}\delta\rfloor+1}^{\infty}\sum_{i=1}^{\infty}\mathbb{E}[\mathbb{E}[q_i^{(J_0)}|J_0]\mathbb{E}[q_i^{(J-\mathcal{T}_{n+1})}(\mathcal{T}_{n+1}-\mathcal{T}_n)|J_0]]e^{-nq_i(1-2\epsilon)/\delta},\end{aligned} \tag{35}$$

where in the last inequality we used conditional independence of the Markov chain. Due to the asymptotic independence of the aperiodic, finite state Markov chains, i.e., $\mathbb{P}[J_{-\mathcal{T}_n} = k | J_0 = l] \to \mathbb{P}[J_{-\mathcal{T}_n} = k]$ as $n \to \infty$, together with (35), (33) and (31), we conclude, as $x \to \infty$,

$$I_{22}^{(2)}(x) \le (1+\epsilon) \sum_{n=\lfloor g_\epsilon x^\alpha \delta \rfloor + 1}^{\infty} \sum_{i=1}^{\infty} q_i^2 \frac{1}{\delta} e^{-(1-2\epsilon) q_i \frac{n}{\delta}} + o\left(\frac{1}{x^{\alpha-1}}\right). \tag{36}$$

The sum from the previous inequality can be further upper-bounded as

$$\sum_{n=\lfloor g_\epsilon x^\alpha \delta \rfloor + 1}^{\infty} \sum_{i=1}^{\infty} q_i^2 \frac{1}{\delta} e^{-(1-2\epsilon) q_i \frac{n}{\delta}} \le \int_{g_\epsilon x^\alpha \delta}^{\infty} \sum_{i=1}^{\infty} q_i^2 \frac{1}{\delta} e^{-(1-2\epsilon) \frac{q_i}{\delta} t} dt. \tag{37}$$

Next, after applying Lemma 2, computing the integral, multiplying it with $x^{\alpha-1}$ and then taking the $\limsup$ as $x \to \infty$, we derive

$$\limsup_{x \to \infty} I_{22}^{(2)}(x) x^{\alpha-1} \le K(\alpha) \frac{(1+\epsilon)^2 (1+2\epsilon)^{1-\frac{1}{\alpha}}}{(1-2\epsilon)^{1+\alpha-\frac{1}{\alpha}}}, \tag{38}$$

which by passing $\epsilon \to 0$, and in conjunction with (36), (33), (31), (30), (29) and (18) proves the upper bound.

The estimation of the lower bound of (5) starts from

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

where $g_\epsilon \triangleq \frac{(1+2\epsilon)^\alpha}{[\Gamma[1-\frac{1}{\alpha}]]^\alpha c(1-\epsilon)}$. Then, using analogous arguments to those in obtaining (32), with redefined $\mathbb{P}[B_i^*(n) = 1] = 1 - e^{-(1-2\epsilon) q_i (n+1)/\delta}$, $i \ge 1$, we obtain

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

Then, after applying Lemmas 3 and 4 for large enough x, lower bounding $\hat{f}(t)$ analogously as in (34), and computing the integral, we obtain

$$\mathbb{P}[C > x] \ge (1-\epsilon) \sum_{n=\lfloor g_\epsilon x^\alpha \delta \rfloor}^{\infty} \sum_{i=1}^{\infty} q_i^{(J_0)} q_i^{(J_{-\mathcal{T}_{n+1}})} (\mathcal{T}_{n+1} - \mathcal{T}_n) e^{-(1+2\epsilon) q_i \frac{n+1}{\delta}} - \sum_{n=\lfloor g_\epsilon x^\alpha \delta \rfloor}^{\infty} \mathbb{P}[\mathcal{A}^c(n)].$$

Finally, after applying analogous arguments as in (35), (36), (37) and (38), to estimate the asymptotic lower bound of the first term in the previous expression, in conjunction with (30), we conclude the proof of this theorem. (The details of the proof of the lower bound are omitted in order to avoid repetitions of the arguments.)

◇

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 3 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 2 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 stated, it is surprising that this value is very small (almost negligible) in comparison with the time scale of jumps $(1/\delta)$ in the modulating process J.

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 (see [12, 11]) that for the semi-Markov modulated requests and generalized Zipf's law marginal request distributions, the caching performance does not depend on the correlation in the request traffic for large cache sizes; more precisely, LRU cache performance is asymptotically identical to the i.i.d. case with the same marginal request distribution. However, it remained unknown what is the critical cache size below which this asymptotic insensitivity does not hold. Our goal in this paper was to determine this critical size. In order to pursue the analysis, we observed a specific case of the model introduced in [12, 11], where requests are modulated by a nearly completely decomposable Markov process. By exploiting the analytic techniques introduced in [12], we discover the critical scaling between the cache size x and the transition rate of the modulating (Markov) process δ. The result is somewhat surprising, since the analysis shows that the critical cache size is sublinear, and, therefore, much smaller than the time scale of transitions in the modulating process $1/\delta$. Furthermore, Theorem 3 identifies a transition region for cache sizes as a function of δ above which the previously mentioned insensitivity result holds.

References

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

[2] J. L. Bentley and C. C. McGeoch. Amortized analysis of self-organizing sequential search heuristics. *Communications of the ACM*, 28(4):404–411, 1985.

[3] Allan Borodin and Ran El-Yaniv, editors. *Online Computation and Competative Analysis*. Cambridge University Press, 1998.

[4] L. Breslau, Pei Cao, Li Fan, G. Phillips, and S. Shenker. Web caching and Zipf-like distributions: Evidence and implications. In *IEEE INFOCOM*, 1999.

[5] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Jones and Bartlett Publishers, 1993.

[6] J. A. Fill. An exact formula for the move-to-front rule for self-organizing lists. *Journal of Theoretical Probability*, 9(1):113–159, 1996.

[7] J. A. Fill. Limits and rate of convergence for the distribution of search cost under the move-to-front rule. *Theoretical Computer Science*, 164:185–206, 1996.

[8] J. A. Fill and L. Holst. On the distribution of search cost for the move-to-front rule. *Random Structures and Algorithms*, 8(3):179, 1996.

[9] P. Flajolet, D. Gardy, and L. Thimonier. Birthday paradox, coupon collector, caching algorithms and self-organizing search. *Discrete Applied Mathematics*, 39:207–229, 1992.

[10] P. R. Jelenković. Asymptotic approximation of the move-to-front search cost distribution and least-recently-used caching fault probabilities. *Annals of Applied Probability*, 9(2):430–464, 1999.

[11] P. R. Jelenković and A. Radovanović. Asymptotic insensitivity of least-recently-used caching to statistical dependency. In *Proceedings of INFOCOM 2003*, San Fransisco, April 2003.

[12] P. R. Jelenković and A. Radovanović. Least-recently-used caching with dependent requests. *Theoretical Computer Science*, 326:293–327, 2004.

[13] D. D. Sleator and R. E. Tarjan. Self-adjusting binary search trees. *Journal of the ACM*, 32(3):652–686, 1985.

[14] Neal E. Young. On-line paging against adversarially biased random inputs. *J. Algorithms*, 37(1):218–235, 2000.