

LRU Caching with Moderately Heavy Request Distributions

Predrag R. Jelenković and Xiaozhu Kang
Department of Electrical Engineering
Columbia University
New York, NY 10027
{predrag, xiaozhu}@ee.columbia.edu

Abstract

Majority of practical caching algorithms, in particular those used in the World Wide Web applications, are based on the so-called Least-Recently-Used (LRU) cache replacement heuristic whose desirable attributes include low complexity, quick adaptability and high cache hit (low fault) probability. Recent studies have developed asymptotic characterization of the LRU fault probability for the generalized Zipf's (power) law request distributions. In this paper, we extend these results to include the distributions that decay faster than power laws but slower than exponential, hence named moderately heavy distributions. Informally, for these types of distributions and the independent reference model, the main result of this paper shows that the ratio between the cache fault probabilities of the LRU heuristic and the optimal static algorithm is, for large caches, equal to $e^\gamma \approx 1.78$, where γ is Euler's constant. Interestingly enough, this limiting ratio is constant, i.e., it is invariant to the underlying characteristics of the request distributions.

Keywords: least-recently-used caching, move-to-front searching, moderately heavy distributions, Web caching, cache fault probability, average-case analysis

1 INTRODUCTION

Renewed interest in caching algorithms stems from their widespread use for content delivery over the World Wide Web (Web) since storing popular documents in proxy caches close to end-users significantly reduces the document download latency and network congestion. Traditional application of caching was in computer engineering, where it is used to speed-up the data transfer between the central processor unit and slow local memory. The basic idea of caching is to enable high-speed access to a subset of x items out of a larger collection of N documents that are stored in a slow access medium, i.e., they cannot be accessed quickly.

One of the fundamental issues of caching is the problem of selecting and possibly dynamically updating the x items that need to be stored in the fast memory (cache). The optimal solution to this problem is often very difficult to find and, therefore, a number of heuristic, usually dynamic, cache

updating algorithms have been proposed. Among the most popular algorithms are those based on the Least-Recently-Used (LRU) cache replacement rule. The wide popularity of this rule is primarily due to its high performance and ease of implementation. LRU algorithm tends to both keep more frequent items in the cache as well as quickly adapt to the potential changes in document popularity, resulting in efficient performance.

In order to further the insight into designing network caching algorithms, it is important to gain a thorough understanding of the baseline LRU cache replacement policy. In the analysis of LRU caching scheme there have been two approaches: combinatorial and probabilistic studies. In this paper we focus on the average-case or probabilistic analysis of MTF and LRU algorithms, e.g., see [1, 2, 3, 5] and the references therein.

In [5], a new analytical technique was developed for the asymptotic analysis of the LRU cache fault probability under the independent reference model with generalized Zipf's law requests. The results from [5] show that the LRU fault probability is asymptotically at most by a constant factor ($e^\gamma \approx 1.78$) away from the optimal frequency algorithm that keeps most frequently used documents in the cache, i.e., replaces Least-Frequently-Used (LFU) items. Within this context and the independent reference model, it is well known that the static LFU policy that stores the most popular documents in the cache is optimal. For direct arguments that justify this intuitively apparent statement see the first paragraph of Subsection 4.1 in [8]. Recently, motivated to close this gap in performance between the LRU and LFU caching algorithms, a new algorithm, termed Persistent Access Caching, was introduced in [6, 9]. This algorithm, in addition to desirable low complexity and adaptability, achieves nearly optimal performance for the independent reference model and generalized Zipf's law request probabilities. Also in the context of generalized Zipf's law requests, the work in [7] relaxes the independent reference assumption and shows that the LRU fault probability is asymptotically, for large caches, insensitive to the possibly strong dependency structure of the request process, i.e., for large cache sizes, the LRU fault probability behaves exactly the same as in the case of independent request sequences.

All of these prior studies analyze the LRU performance

in the context of power (generalized Zipf's) law request distributions. The main objective of this paper is to extend these results to include the distributions that decay faster than power laws but slower than exponential, hence named moderately heavy distributions. Our main result is stated in Theorem 3.1 of Section 3. The formal description of the model and preliminary results are stated in Section 2. In Section 4, we present numerical illustrations of our main result. Concluding remarks and more technical proofs are presented in Section 5 and 6, respectively.

1.1 Empirical Motivation Several recent experimental studies show that the request distributions may not follow Zipf's law, e.g., this was reported in [4] for some P2P file-sharing networks. In this paper, we examine a number of real traces from the log files collected by National Laboratory for Applied Network Research, where we find Web traces that exhibit Weibull (non-Zipf's) characteristics, which provides additional motivation for our theoretical study; here, we present one such example. During the experiment, we exclude the statistically insignificant items that are requested only once, and, thus, use the top 405 requested items only. We analyze one day long trace of 376532 requests to a proxy caches in MAE-West San Jose, California. The measured experimental distribution is matched with a class of Weibull distributions of the form $q_i = ce^{-\lambda i^\beta}$. During the fitting process, we first estimate c by calculating frequencies of the most popular documents (q_0, q_1, q_2) , then we use MATLAB's least square linear fitting tool to obtain λ and β . The fitted distribution $q_i = e^{-0.81-2.387i^{0.17}}$ is presented in Figure 1 in red dashed lines compared with the measured distribution in solid blue line. The accuracy of the fitting is apparent from the figure.

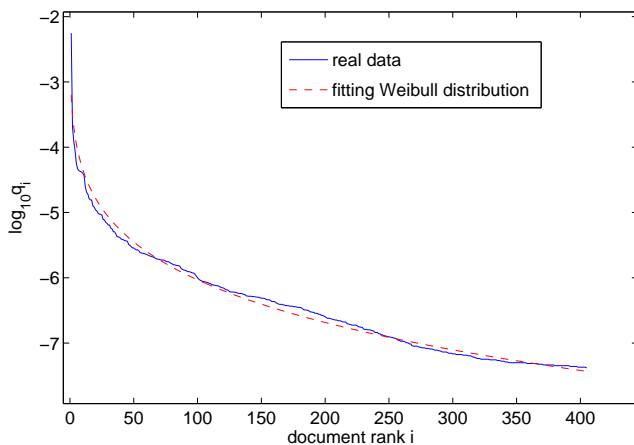


Figure 1: Fitting a real document request trace with Weibull distribution $q_i = e^{-0.81-2.387i^{0.17}}$.

2 MODEL DESCRIPTION AND PRELIMINARY RESULTS

Consider N items, out of which x are kept in a fast memory (cache) and the remaining $N - x$ are stored in a slow memory. Each time a request for an item is made, the cache is searched first. If the item is not found there, it is brought in from the slow memory and replaced with the least recently accessed item from the cache. Such a replacement policy is commonly referred to as LRU, as previously stated in the introduction. The performance quantity of interest for this algorithm is the LRU fault probability, i.e., the probability that the requested item is not in the cache. Our goal in this paper is to asymptotically characterize this probability. The fault probability of the LRU caching is equivalent to the tail of the searching cost distribution for the MTF searching algorithm. In order to justify this claim, we note that x elements in the cache, under the LRU rule, are arranged in the increasing order of their last access times. Each time there is a request for an item that is not in the cache, the item is brought to the first position of the cache and the last element of the cache is moved to the slow memory. We argue that the fault probability stays the same if the remaining items in the slow memory are arranged in any specific order. In particular, they can be arranged in the increasing order of their last access times. The obtained algorithm is then the same as the MTF searching algorithm.

More formally, consider a possibly infinite list of items $\{1, 2, \dots, N\}$ and a sequence of requests that arrive at moments $\{\tau_n\}_{n>-\infty}$ with increments $\{\tau_{n+1} - \tau_n\}_{n>-\infty}$, $\tau_0 = 0$, being stationary and ergodic having $\mathbb{E}\tau_1 = 1/\lambda$ for some $\lambda > 0$, and $\tau_{n+1} - \tau_n > 0$ a.s.. Furthermore, define a sequence of i.i.d. random variables $\{R_n\}_{n>-\infty}$, independent from $\{\tau_n\}$, where $\{R_n = i\}$ represents a request for item i at time τ_n . We denote request probabilities as $\mathbb{P}[R_n = i] = q_i$ and, without loss of generality, assume $q_1 \geq q_2 \geq \dots$.

The dynamics of the MTF algorithm is defined as follows. Suppose that the system starts at moment τ_0 of 0th request with an initial permutation of the list. Then, at every time instant, that an item, say i , is requested, its position in the list is first determined; if i is in the k th position, we say that the search cost C_n^N for this item is equal to k . Now, the list is updated by moving item i to the first position of the list and items in positions $1, \dots, k - 1$ are moved one position down. Note that, according to the discussion in the preceding paragraph, $\mathbb{P}[C_n^N > x]$ represents the stationary fault probability for a cache of size x .

Let $-T_i$ to be the last time before $t = 0$ that item i was requested and define Bernoulli variables $\{B_i(t)\}_{i \geq 1}$ that indicate an item i was requested in $[-t, 0)$, i.e., $B_i(t) = \mathbf{1}[T_i < t]$, $S_i(t) = \sum_{j \neq i} B_j(t)$ and $S(t) = \sum_{i=1}^{\infty} B_i(t)$. The following representation result is a special case of Lemma 1 from [6].

LEMMA 2.1. For any $1 \leq N \leq \infty$, arbitrary initial conditions (Π_0) and any $x \geq 0$, the search cost $C_n^{(N)}$ converges in distribution to $C^{(N)}$ as $n \rightarrow \infty$, where

$$\mathbb{P}[C^{(N)} > x] \triangleq \sum_{i=1}^N q_i \mathbb{P}[S_i(T_i) \geq x]$$

and $S_i(t) \triangleq \sum_{j \neq i} \mathbf{1}[T_j < t]$, $i \geq 1$.

As noted in Remark 1 of [6], the distribution of $C^{(N)}$ does not depend on the distribution of renewal interarrivals $(\tau_{n+1} - \tau_n)$. Thus, without loss of generality, we can assume that these points $\{\tau_n\}$ are Poisson. Thus, by the Poisson decomposition property, $\{B_i(t)\}_{i \geq 1}$ are independent with success probabilities $\mathbb{P}[B_i(t) = 1] = 1 - e^{-q_i t}$. The Poisson embedding technique for LRU policy with i.i.d. requests was first introduced in [2].

For the rest of the paper, we assume that $N = \infty$ and denote $C \equiv C^{(\infty)}$. We use H to denote a sufficiently large positive constant and h 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. Also, we use the following 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 $t \rightarrow t_0$ to denote $\lim_{t \rightarrow t_0} [a(t)/b(t)] = 1$. Similarly, we say that $a(t) \gtrsim b(t)$ as $t \rightarrow t_0$ if $\liminf_{t \rightarrow t_0} a(t)/b(t) \geq 1$; $a(t) \lesssim b(t)$ has a complementary definition.

In the context of the independent reference model and heavy-tailed Zipf's law requests $\mathbb{P}[R = i] \sim c/i^\alpha$ as $i \rightarrow \infty$, $\alpha > 1$ in Theorem 3 of [5] it was proved

$$\lim_{x \rightarrow \infty} \frac{\mathbb{P}[C > x]}{\mathbb{P}[R > x]} = \left(1 - \frac{1}{\alpha}\right) \left[\Gamma\left(1 - \frac{1}{\alpha}\right)\right]^\alpha \nearrow e^\gamma$$

as $\alpha \rightarrow \infty$, where Γ is the Gamma function and $\gamma = 0.5772 \dots$ is Euler's constant. This result was extended in Theorem 2 and Theorem 3 of [6] to Zipf's law distributions with $0 < \alpha \leq 1$.

However, the asymptotic behavior of $\mathbb{P}[C > x]$ is not known for request distributions that are lighter than power laws, e.g., Lognormal and Weibull. Although, in [5] a fluid limit approximation C_f of C was introduced and it was shown that, for request distribution $\mathbb{P}[R = i] \sim ce^{-\lambda i^\beta}$, $(c, \lambda, \beta > 0)$ as $i \rightarrow \infty$,

$$(2.1) \quad \lim_{x \rightarrow \infty} \frac{\mathbb{P}[C_f > x]}{\mathbb{P}[R > x]} = e^\gamma.$$

This result suggests that $\mathbb{P}[C > x]$ can have a similar (or even the same) asymptotic behavior, but there is no rigorous way of making this intuitive argument precise. Hence, in this paper we prove (2.1) directly in Theorem 3.1 without the fluid limit approximation. In this regard, we develop a more refined analytical technique, as compared to those in [5, 7].

3 MAIN RESULTS

The following theorem is our main result that will be proved in the remainder of this section.

THEOREM 3.1. Assume that $q_i \sim ce^{-\lambda i^\beta}$, $0 < \beta < 1$, or $q_i \sim ce^{-\lambda(\log i)^\alpha}$, $\alpha > 1$ as $i \rightarrow \infty$, $(c, \lambda) > 0$, then

$$\lim_{x \rightarrow \infty} \frac{\mathbb{P}[C > x]}{\mathbb{P}[R > x]} = e^\gamma,$$

where $\gamma (= 0.5772 \dots)$ is Euler's constant.

Remark: This theorem shows that the ratio between the fault probability of LRU and the tail of the request distribution is asymptotically invariant to the parameters of the considered class of request frequencies. The result also implies that the LRU fault probability is asymptotically equal to e^γ (≈ 1.78) times the optimal static fault probability.

In order to prove the main theorem, we need the following technical lemmas. To this end, let us define $m(t) = \mathbb{E}S(t) = \sum_{i=1}^{\infty} (1 - e^{-q_i t})$. Also note that, by the Dominated Convergence Theorem, $m'(t) = \sum_{i=1}^{\infty} q_i e^{-q_i t}$ and the inverse $m^{-1}(t)$ is well defined since $m(t)$ is strictly increasing. Our first Lemma 3.1 is a direct consequence of Lemma 3 and 6 of [5].

LEMMA 3.1. If $q_i \sim ce^{-\lambda i^\beta}$ as $i \rightarrow \infty$, where $(c, \lambda, \beta) > 0$, then $m(t)$, its derivative and inverse behave asymptotically, as $t \rightarrow \infty$,

$$m(t) \sim \frac{(\log t)^{\frac{1}{\beta}}}{\lambda^{\frac{1}{\beta}}}, \quad m'(t) \sim \frac{(\log t)^{\frac{1}{\beta}-1}}{t\beta\lambda^{\frac{1}{\beta}}}$$

and

$$m^{-1}(t) \sim e^{-\gamma} c^{-1} e^{\lambda t^\beta}.$$

Next, we define the variance of $S(t)$

$$\begin{aligned} \sigma^2(t) &\triangleq \text{Var}(S(t)) = \sum_{i=1}^{\infty} e^{-q_i t} (1 - e^{-q_i t}) \\ &= \sum_{i=1}^{\infty} (1 - e^{-2q_i t}) - \sum_{i=1}^{\infty} (1 - e^{-q_i t}) \\ &= m(2t) - m(t), \end{aligned}$$

where the second equality follows from the Dominated Convergence Theorem since $1 - e^{-x} \leq x$, $x \geq 0$ implies $m(t) \leq \sum_{i=1}^{\infty} q_i t = t$.

LEMMA 3.2. If $q_i \sim ce^{-\lambda i^\beta}$ as $i \rightarrow \infty$, where $(c, \lambda, \beta) > 0$, then

$$\sigma^2(t) \sim \frac{(\log t)^{\frac{1}{\beta}-1} \log 2}{\lambda^{\frac{1}{\beta}} \beta} \quad \text{as } t \rightarrow \infty.$$

The proof is given in Section 6.

LEMMA 3.3. If $q_i \sim ce^{-\lambda(\log i)^\alpha}$, $\alpha > 1$ as $i \rightarrow \infty$, then $m(t), m'(t), m^{-1}(t), \sigma^2(t)$ behave asymptotically, as $t \rightarrow \infty$,

$$m(t) \sim e^{(\frac{\log t}{\lambda})^{1/\alpha}}, \quad m^{-1}(t) \sim e^{-\gamma} c^{-1} e^{\lambda(\log t)^\alpha}$$

and

$$m'(t) \sim \frac{e^{(\frac{\log t}{\lambda})^{1/\alpha}} (\log t)^{1/\alpha-1}}{\alpha \lambda^{1/\alpha} t},$$

$$\sigma^2(t) \sim \frac{e^{(\frac{\log t}{\lambda})^{1/\alpha}} (\log t)^{1/\alpha-1} \log 2}{\alpha \lambda^{1/\alpha}}.$$

Similarly, the **proof** is given in Section 6.

LEMMA 3.4. Let $\{B_i\}_{i \geq 1}$ be a sequence of independent Bernoulli random variables with $S = \sum_{i=1}^{\infty} B_i$, $m = \mathbb{E}[S]$ and $\sigma^2 = \text{Var}[S]$. Then, for any $\theta > 0$ and $y > \theta e^\theta \sigma^2$, we have

$$\mathbb{P}[S > m + y] \leq e^{-\frac{\theta y}{2}}.$$

Proof: Using Markov's inequality, for any $\theta > 0$, we obtain

$$(3.2) \quad \mathbb{P}[S > m + y] \leq e^{-\theta y} \mathbb{E}[e^{\theta(S-m)}].$$

Let $m_i \triangleq \mathbb{E}B_i$, then Taylor's expansion for $\varphi(\theta) \triangleq \mathbb{E}e^{\theta(B_i - m_i)}$ yields

$$\varphi(\theta) = 1 + \frac{\varphi''(\zeta_\theta)\theta^2}{2}, \quad 0 \leq \zeta_\theta \leq \theta,$$

where $\varphi''(\zeta_\theta) = \mathbb{E}[(B_i - m_i)^2 e^{\zeta_\theta(B_i - m_i)}] \leq \text{Var}(B_i)e^\theta$. The last bound and $1 + x \leq e^x$ for $x \geq 0$ imply

$$\mathbb{E}e^{\theta(B_i - m_i)} \leq e^{\text{Var}(B_i)e^\theta/2}.$$

Using the preceding bound and the fact that $\{B_i\}_{i \geq 1}$ are independent Bernoulli random variables, we derive

$$\begin{aligned} \mathbb{E}e^{\theta(S-m)} &= \prod_{i=1}^{\infty} \mathbb{E}e^{\theta(B_i - m_i)} \\ &\leq \prod_{i=1}^{\infty} e^{\text{Var}(B_i)\theta^2 e^\theta/2} \\ &= e^{\sigma^2 \theta^2 e^\theta/2}. \end{aligned}$$

By replacing the preceding bound in (3.2), we arrive at

$$\begin{aligned} \mathbb{P}[S > m + y] &\leq e^{\sigma^2 \theta^2 e^\theta/2 - \theta y} \\ &= e^{-\theta y(1 - \frac{\sigma^2 \theta e^\theta}{2y})} \\ &\leq e^{-\frac{\theta y}{2}}, \end{aligned}$$

where the last inequality is implied by $y > \theta e^\theta \sigma^2$. \diamond

Proof of Theorem 3.1: Weibull case $q_i \sim ce^{-\lambda i^\beta}$. First, we prove the lower bound. To this end, note that

$$\begin{aligned} \mathbb{P}[C > x] &\geq \sum_{i=1}^{\infty} q_i \mathbb{P}[S_i(t_x) > x - 1, T_i > t_x] \\ &= \sum_{i=1}^{\infty} q_i e^{-q_i t_x} \mathbb{P}[S(t_x) > x] \\ (3.3) \quad &= m'(t_x) \mathbb{P}[S(t_x) > x]. \end{aligned}$$

Then, choose t_x such that $m(t_x) = x + x^{(1-\beta)(1-\delta/2)}$ for some $0 < \delta < 1$. By monotonicity of $m(t)$, we have $t_x \leq m^{-1}(x(1 + \delta))$, which implies $\sigma^2(t_x) \leq Hx^{1-\beta}$ by Lemma 3.2. Hence, this estimate and Chebyshev's inequality yield

$$\begin{aligned} \mathbb{P}[S(t_x) > x] &= \mathbb{P}[S(t_x) - m(t_x) > -x^{(1-\beta)(1-\delta/2)}] \\ &\geq 1 - \frac{\sigma^2(t_x)}{x^{(1-\beta)(2-\delta)}} \\ (3.4) \quad &\geq 1 - \frac{H}{x^{(1-\beta)(1-\delta)}}. \end{aligned}$$

Now, by Lemma 3.1, we obtain

$$\begin{aligned} t_x &\leq (1 + \varepsilon/2)e^{-\gamma} c^{-1} e^{\lambda(x+x^{(1-\beta)(1-\delta/2)})^\beta} \\ &= (1 + \varepsilon/2)e^{-\gamma} c^{-1} e^{\lambda x^\beta} e^{\beta \lambda x^{-(1-\beta)\delta/2}} \\ &\sim (1 + \varepsilon/2)e^{-\gamma} c^{-1} e^{\lambda x^\beta} \text{ as } x \rightarrow \infty. \end{aligned}$$

Therefore, by Lemma 3.1,

$$\begin{aligned} m'(t_x) &\geq m' \left((1 + \varepsilon)e^{-\gamma} c^{-1} e^{\lambda x^\beta} \right) \\ &\sim \frac{(\log((1 + \varepsilon)e^{-\gamma} c^{-1} e^{\lambda x^\beta}))^{\frac{1}{\beta}-1}}{((1 + \varepsilon)e^{-\gamma} c^{-1} e^{\lambda x^\beta})\beta \lambda^{\frac{1}{\beta}}} \\ (3.5) \quad &\sim \frac{cx^{1-\beta} e^{\gamma} e^{-\lambda x^\beta}}{(1 + \varepsilon)\beta \lambda} \text{ as } x \rightarrow \infty. \end{aligned}$$

Now, easy calculations show that

$$(3.6) \quad \mathbb{P}[R > x] \sim \frac{cx^{1-\beta} e^{-\lambda x^\beta}}{\beta \lambda},$$

which, together with (3.3), (3.4), (3.5), yields

$$\liminf_{x \rightarrow \infty} \frac{\mathbb{P}[C > x]}{\mathbb{P}[R > x]} \geq \frac{e^\gamma}{1 + \varepsilon};$$

now, letting $\varepsilon \rightarrow 0$, completes the proof of the lower bound.

For the upper bound, choose t_0 to be large and redefine t_x such that $m(t_x) = x - \varepsilon x^{(1-\beta)}$ for some $\varepsilon > 0$, then we

have, for $t_0 < t_x$,

$$\begin{aligned}
\mathbb{P}[C > x] &= \sum_{i=1}^{\infty} q_i \mathbb{P}[S_i(T_i) \geq x] \\
&\leq \mathbb{P}[S(t_0) \geq x] \\
&\quad + \int_{t_0}^{t_x} \sum_{i=1}^{\infty} q_i^2 e^{-q_i t} \mathbb{P}[S(t) > x] dt \\
&\quad + \sum_{i=1}^{\infty} q_i \mathbb{P}[T_i > t_x] \\
(3.7) \quad &\triangleq I_1(x) + I_2(x) + I_3(x).
\end{aligned}$$

First, let $x_0 \triangleq m(t_0)$; then, for any $t_0 > 0$, we can choose x large enough such that $x - x_0 > 2e^2 \text{Var}(S(t_0))$ holds, therefore we can apply Lemma 3.4 with $\theta = 2$ to obtain

$$\begin{aligned}
I_1(x) &= \mathbb{P}[S(t_0) > x] \\
&\leq e^{x_0 - x} \\
(3.8) \quad &= o(e^{-\lambda x^\beta}) \text{ as } x \rightarrow \infty.
\end{aligned}$$

Next, since $m'(t) = \int_t^\infty m''(u) du$ and $m''(t)$ is monotonic, by using the asymptotics of $m'(t)$ in Lemma 3.1 and the Monotone Density Theorem on p. 39 of [10], for $t_0 \leq t \leq t_x$, we have

$$-m''(t) \leq \frac{H(\log t)^{\frac{1}{\beta}-1}}{t^2} \leq \frac{Hm'(t)}{t}.$$

Thus, $I_2(x)$ is bounded by

$$\begin{aligned}
I_2(x) &= \int_{t_0}^{t_x} -m''(t) \mathbb{P}[S(t) \geq x] dt \\
&\leq H \int_{t_0}^{t_x} \frac{m'(t)}{t} \mathbb{P}[S(t) \geq x] dt.
\end{aligned}$$

Now by changing variable $y = x - m(t)$ in the preceding integral, using $t = m^{-1}(x - y)$, the asymptotic expression for $m^{-1}(x - y)$ from Lemma 3.1 and $-m'(t)dt = dy$, we obtain

$$I_2(x) \leq H \int_{\epsilon x^{1-\beta}}^{x-x_0} \frac{\mathbb{P}[S(m^{-1}(x-y)) - (x-y) > y]}{e^{\lambda(x-y)^\beta}} dy;$$

recall $x_0 = m(t_0)$.

At this point, recall the definition $\sigma^2(t) = \text{Var}(S(t))$ and note that, by Lemma 3.2,

$$\max_{0 \leq y \leq x-x_0} \sigma^2(m^{-1}(x-y)) \leq Hx^{1-\beta}.$$

Thus, we can choose $\theta > 0$ small enough such that

$$\begin{aligned}
\max_{\epsilon x^{1-\beta} \leq y \leq x-x_0} \theta e^\theta \sigma^2(m^{-1}(x-y)) &\leq H\theta e^\theta x^{1-\beta} \\
&< \epsilon x^{1-\beta},
\end{aligned}$$

which ensures that Lemma 3.4 applies to $\mathbb{P}[S(m^{-1}(x-y)) - (x-y) > y]$ and, therefore,

$$(3.9) \quad I_2(x) \leq H \int_{\epsilon x^{1-\beta}}^{x-x_0} e^{-\frac{\theta y}{2} - \lambda(x-y)^\beta} dy.$$

Next, we observe that the exponent $f(y) = \theta y/2 + \lambda(x-y)^\beta$ in the preceding integral is concave since $f''(y) = -\beta(1-\beta)\lambda/(x-y)^{2-\beta} < 0$ for $0 \leq y \leq x - x_0$. Using the concavity property, we obtain

$$(3.10) \quad \min_{\epsilon x^{1-\beta} \leq y \leq x-x_0} f(y) = \min(f(\epsilon x^{1-\beta}), f(x-x_0)).$$

To this end, for large x ,

$$(3.11) \quad f(\epsilon x^{1-\beta}) \geq h x^{1-\beta} + \lambda x^\beta,$$

and

$$(3.12) \quad f(x) \geq hx/2.$$

Combining (3.10), (3.11) and (3.12), the bound in (3.9) becomes

$$(3.13) \quad I_2(x) \leq Hx e^{-hx^{1-\beta} - \lambda x^\beta} = o(e^{-\lambda x^\beta}).$$

In order to estimate $I_3(x)$, note that

$$\begin{aligned}
t_x &= m^{-1}(x - \epsilon x^{(1-\beta)}) \\
&\gtrsim (1-\epsilon) e^{-\gamma} c^{-1} e^{\lambda(x-\epsilon x^{(1-\beta)})^\beta} \\
(3.14) \quad &\gtrsim (1-H\epsilon) e^{-\gamma} c^{-1} e^{\lambda x^\beta} \text{ as } x \rightarrow \infty.
\end{aligned}$$

Next, by Lemma 3.1 and (3.14),

$$\begin{aligned}
I_3(x) &= m'(t_x) \leq m' \left((1-H\epsilon) e^{-\gamma} c^{-1} e^{\lambda x^\beta} \right) \\
&= \frac{(\log((1-H\epsilon) e^{-\gamma} c^{-1} e^{\lambda x^\beta}))^{\frac{1}{\beta}-1}}{((1-H\epsilon) e^{-\gamma} c^{-1} e^{\lambda x^\beta}) \beta \lambda^{\frac{1}{\beta}}} \\
&\sim \frac{c x^{1-\beta} e^{\gamma} e^{-\lambda x^\beta}}{(1-H\epsilon) \beta \lambda} \text{ as } x \rightarrow \infty.
\end{aligned}$$

Finally, combining the last asymptotic expression with (3.6), (3.8), (3.13) and (3.7), we obtain

$$\limsup_{x \rightarrow \infty} \frac{\mathbb{P}[C > x]}{\mathbb{P}[R > x]} \leq \frac{e^\gamma}{1-H\epsilon}.$$

Now, letting $\epsilon \rightarrow 0$ completes the proof of the upper bound and the theorem for the case of $q_i \sim ce^{-\lambda i^\beta}$.

Lognormal-like case $q_i \sim ce^{-\lambda(\log i)^\alpha}$. We prove the lower bound first. Hence, we choose t_x , different from in the Weibull case, such that

$$(3.15) \quad m(t_x) = x + \frac{x}{(\log x)^{(\alpha-1)(1+\delta)}}$$

for some $0 < \delta < 1$, which implies $t_x \leq He^{\lambda(\log x)^\alpha}$ and $\sigma^2(t_x) \leq Hx/(\log x)^{\alpha-1}$. This, combined with Chebyshev's inequality, yields

$$\begin{aligned} \mathbb{P}[S(t_x) > x] &= \mathbb{P}[S(t_x) - m(t_x) > -x/(\log x)^{(\alpha-1)(1+\delta)}] \\ &\geq 1 - \frac{\sigma^2(t_x)(\log x)^{(\alpha-1)(2+2\delta)}}{x^2} \\ (3.16) \quad &\geq 1 - \frac{H(\log x)^{(\alpha-1)(1+2\delta)}}{x}. \end{aligned}$$

Next, by Lemma 3.3, we obtain

$$(3.17) \quad t_x \leq (1 + \varepsilon/2)e^{-\gamma}c^{-1}e^{\lambda(\log(m(t_x)))^\alpha}.$$

Now, by replacing (3.15), we further simplify the exponent $(\log(m(t_x)))^\alpha$ in the preceding expression for large x as

$$\begin{aligned} &(\log(m(t_x)))^\alpha \\ &= \left(\log x + \log\left(1 + 1/(\log x)^{(\alpha-1)(1+\delta)}\right)\right)^\alpha \\ &\leq (\log x)^\alpha \left(1 + \frac{1}{(\log x)^{(\alpha-1)(1+\delta)+1}}\right)^\alpha \\ &\leq (\log x)^\alpha + \frac{\alpha(1+\varepsilon)}{(\log x)^{(\alpha-1)\delta}}, \end{aligned}$$

where, in the second inequality, we used $\log(1+y) \leq y$, $y > 0$ and in the last inequality, we applied $(1+y)^\alpha \leq 1 + \alpha(1+\varepsilon)y$ for y small enough. When the last bound is replaced in (3.17), we obtain

$$\begin{aligned} t_x &\leq (1 + \varepsilon/2)e^{-\gamma}c^{-1}e^{\lambda(\log x)^\alpha} e^{\frac{\lambda}{\alpha(\log x)^{(\alpha-1)\delta}}} \\ &\sim (1 + \varepsilon/2)e^{-\gamma}c^{-1}e^{\lambda(\log x)^\alpha} \text{ as } x \rightarrow \infty. \end{aligned}$$

Then, by Lemma 3.3, we further derive

$$(3.18) \quad m'(t_x) \gtrsim \frac{cxe^\gamma e^{-\lambda(\log x)^\alpha} (\log x)^{1-\alpha}}{\alpha\lambda(1+\varepsilon)} \text{ as } x \rightarrow \infty.$$

Also, straightforward calculation shows that

$$(3.19) \quad \mathbb{P}[R > x] \sim \frac{cxe^{-\lambda(\log x)^\alpha} (\log x)^{1-\alpha}}{\alpha\lambda} \text{ as } x \rightarrow \infty.$$

Using (3.19), replacing (3.16) and (3.18) in (3.3), and letting $\varepsilon \rightarrow 0$ complete the proof of the lower bound.

For the upper bound, first we define $I_1(x)$, $I_2(x)$ and $I_3(x)$ exactly the same as (3.7) and reset t_x such that

$$m(t_x) = x - \frac{\varepsilon x}{(\log x)^{(\alpha-1)}}$$

for some $\varepsilon > 0$.

Similarly as in the Weibull case, we can apply Lemma 3.4, for large x ,

$$\begin{aligned} I_1(x) &= \mathbb{P}[S(t_0) > x] \\ &\leq e^{x_0-x} \\ (3.20) \quad &= o(xe^{-\lambda(\log x)^\alpha} (\log x)^{1-\alpha}) \text{ as } x \rightarrow \infty. \end{aligned}$$

Next, using Lemma 3.3 and the Monotone Density Theorem on p. 39 in [10] for $t_0 \leq t \leq t_x$, we obtain

$$-m''(t) \leq \frac{He^{(\frac{\log t}{\lambda})^{1/\alpha}} (\log t)^{1/\alpha-1}}{t^2} \leq \frac{Hm'(t)}{t}.$$

Thus, $I_2(x)$ is upperbounded by

$$I_2(x) \leq H \int_{t_0}^{t_x} \frac{m'(t)}{t} \mathbb{P}[S(t) \geq x] dt.$$

Again, as in the Weibull case, changing variable $y = x - m(t)$, using $t = m^{-1}(x - y)$ and applying Lemma 3.3, we arrive at

$$I_2(x) \leq H \int_{\varepsilon x(\log x)^{1-\alpha}}^{x-x_0} \frac{\mathbb{P}[S(m^{-1}(x-y)) - (x-y) > y]}{e^{\lambda(\log(x-y))^\alpha}} dy.$$

Since

$$\max_{0 \leq y \leq x-x_0} \sigma^2(m^{-1}(x-y)) \leq \frac{Hx}{(\log x)^{(\alpha-1)}},$$

we can choose $\theta > 0$ small enough such that

$$\begin{aligned} \max_{0 \leq y \leq x-x_0} \theta e^\theta \sigma^2(m^{-1}(x-y)) &\leq H\theta e^\theta x (\log x)^{1-\alpha} \\ &< \varepsilon x (\log x)^{1-\alpha}. \end{aligned}$$

Thus, we can apply Lemma 3.4 to derive

$$I_2(x) \leq H \int_{\varepsilon x(\log x)^{1-\alpha}}^{x-x_0} He^{-\frac{\theta y}{2} - \lambda(\log(x-y))^\alpha} dy.$$

The exponent function in the last bound $f(y) = \frac{\theta y}{2} + \lambda(\log(x-y))^\alpha$ is increasing in the integration interval since $f'(y) = \theta/2 - \alpha\lambda(\log(x-y))^{\alpha-1}/(x-y) > 0$ for $\varepsilon x(\log x)^{1-\alpha} \leq y \leq x-x_0$ and x_0 large enough since $(\log u)^{\alpha-1}/u \rightarrow 0$ as $u \rightarrow \infty$. Therefore, we have

$$\begin{aligned} I_2(x) &\leq Hxe^{-\frac{hx}{(\log x)^{\alpha-1}} - \lambda(\log x)^\alpha} \\ (3.21) \quad &= o(xe^{-\lambda(\log x)^\alpha} (\log x)^{1-\alpha}). \end{aligned}$$

In order to estimate $I_3(x)$, using similar arguments as in proving the lower bound, we derive

$$t_x \gtrsim (1 - H\varepsilon)e^{-\gamma}c^{-1}e^{\lambda(\log x)^\alpha} \text{ as } t \rightarrow \infty,$$

which yields

$$(3.22) \quad I_3(x) \lesssim \frac{cxe^\gamma e^{-\lambda(\log x)^\alpha} (\log x)^{1-\alpha}}{\alpha\lambda(1-H\varepsilon)} \text{ as } x \rightarrow \infty.$$

Finally, combining (3.19), (3.20) (3.21), (3.22) and letting $\varepsilon \rightarrow 0$ complete the proof of the lognormal-like case and the proof of the theorem. \diamond

4 NUMERICAL EXAMPLES

In this section we illustrate our main result stated in Theorem 3.1. Even though the result is valid for large cache sizes, our simulations show that the approximation of the fault probability works very well for small caches as well.

4.1 Convergence to stationarity In the presented experiments, we use a discrete time model without the Poisson embedding, i.e., $\tau_n = n$. In order to ensure that the simulated values of the fault probabilities do not deviate significantly from the stationary ones, we first estimate the difference between the distributions of $C^{(N)}$ and $C_n^{(N)}$, where $C_n^{(N)}$ is the search cost after n requests with arbitrary initial conditions.

Using same argument as in [6], we can upper bound the difference between the tails of these distributions as

$$\sup_x \left| \mathbb{P}[C_n^{(N)} > x] - \mathbb{P}[C^{(N)} > x] \right| \leq e_n \triangleq \sum_{i=1}^N q_i \mathbb{P}[T_i \geq n],$$

where now T_i denotes success times in a Bernoulli process with parameter q_i , thus

$$\mathbb{P}[T_i \geq n] = (1 - q_i)^{n-1}.$$

Using the preceding bound, we obtain

$$(4.23) \quad e_n = \sum_{i=1}^N q_i (1 - q_i)^{n-1}.$$

4.2 Experiments In the presented experiments, the initial permutation of the list is chosen uniformly at random. The simulation results and the probability approximation $e^\gamma \mathbb{P}[R > x]$ are presented with “*” symbols and solid lines on Figures 2, 3 and 4, respectively.

Since our asymptotic formula is obtained for infinite number of documents N , it can be expected that asymptotic expression gives reasonable approximation of the fault probability $\mathbb{P}[C^{(N)} > x]$ only if both N and x are large (with N much larger than x). However, our experiments show that the obtained approximation works well for relatively small values of N and almost all cache sizes $x < N$.

Experiment 1 We set request distribution to be Weibull $\mathbb{P}[R = i] = c_N e^{-i^{1/3}}$, where $c_N = (\sum_{i=1}^N e^{-i^{1/3}})^{-1}$ and $N = 800$. The fault probabilities are measured for cache sizes $x = 50j, 1 \leq j \leq 10$. Before conducting measurements, we let the experiment run for $n = 10^7$ requests to ensure that the system reaches stationarity. Then, the actual measurement time is also set to be 10^7 requests long. After computing e_n using (4.23) for a given warm-up time of 10^7 requests, we obtain that $e_n < 3 \times 10^{-11}$, which is negligible compared to the smallest measured probabilities ($> 10^{-2}$). Therefore, the measured fault probabilities are

essentially the stationary ones. The accuracy of approximation $e^\gamma \mathbb{P}[R > x]$ is apparent from Figure 2.

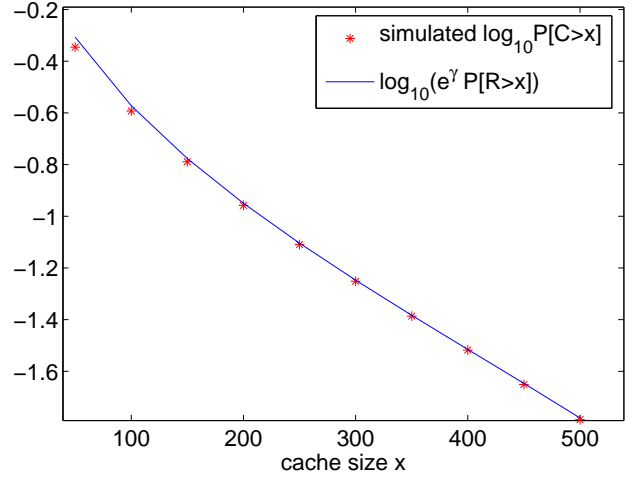


Figure 2: Illustration for Experiment 1.

Experiment 2 In this example, we choose Weibull request distribution $\mathbb{P}[R = i] = c_N e^{-i^{2/3}}$, where $c_N = (\sum_{i=1}^N e^{-i^{2/3}})^{-1}$ and $N = 70$. The fault probabilities are measured for cache sizes $x = 5j, 1 \leq j \leq 8$. Here, we set the warm up time to $n = 2 \times 10^8$ before starting the measurements and then collect the data for 2×10^8 requests as well. Again, evaluating formula (4.23) for $n = 2 \times 10^8$ yields $e_n < 10^{-11}$, which is much smaller than the measured probabilities ($> 10^{-5}$), implying that we are basically observing the stationary probabilities. An excellent fit of approximation $e^\gamma \mathbb{P}[R > x]$ to the simulated data, even for relatively small cache sizes (5 – 40), can be observed from Figure 3.

Experiment 3 Now, we exemplify the lognormal-like case with $\mathbb{P}[R = i] = c_N e^{-(\log i)^2/2}$, where $c_N = (\sum_{i=1}^N e^{-(\log i)^2/2})^{-1}$, $N = 100$, and measure the fault probabilities for cache sizes $x = 5j, 1 \leq j \leq 12$. We use the first 2×10^6 requests to warm up the system to reach stationarity and then measure the cache fault probabilities for the following 2×10^6 requests. After estimating e_n in (4.23) for $n = 2 \times 10^6$, we obtain that $e_n < 6 \times 10^{-11}$, which justifies that the measured probabilities ($> 10^{-3}$) are the stationary ones. Similarly, further validation of our theoretical result is depicted in Figure 4.

5 Concluding Remarks

Motivated by recent empirical findings that the request distributions may not follow Zipf’s law, we extend the asymp-

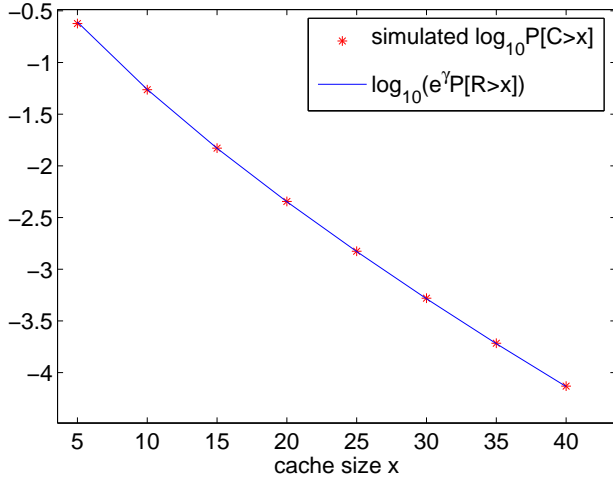


Figure 3: Illustration for Experiment 2.

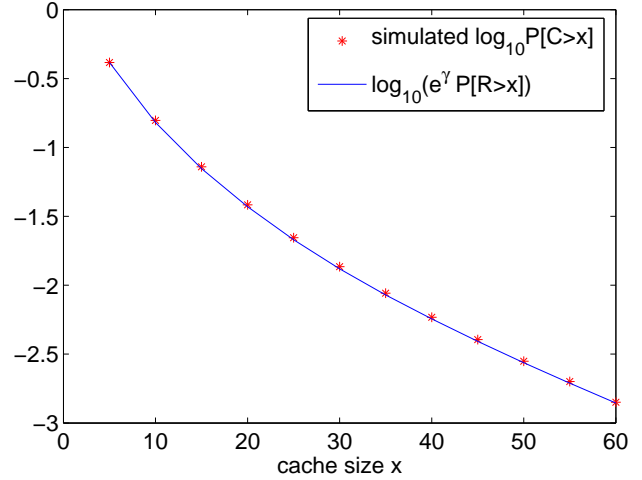


Figure 4: Illustration for Experiment 3.

otic results on LRU fault probability to classes of moderately heavy distributions (e.g., some Weibull and lognormal) that decay faster than power law but slower than exponential. We prove that the ratio between the cache fault probability of LRU and the optimal static algorithm is asymptotically $e^\gamma \approx 1.78$ for large caches. It is interesting that this limiting ratio is invariant to the specific parameters of the considered request distributions. We verify the insensitivity result of the ratio with numerical experiments in Section 4 as well.

6 Proofs

Proof of Lemma 3.2: Suppose first that $0 < \beta \leq 1$. Then, by Lemma 3.1 and using the fact that $(\log u)^{1/\beta-1}$ is nondecreasing for $0 < \beta \leq 1$, we can prove the upper bound as follows

$$\begin{aligned}
 \sigma^2(t) &\leq (1 + \varepsilon) \int_t^{2t} \frac{(\log u)^{\frac{1}{\beta}-1}}{u \lambda^{\frac{1}{\beta}} \beta} du \\
 &\leq (1 + \varepsilon) \frac{(\log 2t)^{\frac{1}{\beta}-1}}{\lambda^{\frac{1}{\beta}} \beta} \int_t^{2t} \frac{1}{u} du \\
 &= (1 + \varepsilon) \frac{(\log 2t)^{\frac{1}{\beta}-1} \log 2}{\lambda^{\frac{1}{\beta}} \beta} \\
 (6.24) \quad &\sim (1 + \varepsilon) \frac{(\log t)^{\frac{1}{\beta}-1} \log 2}{\lambda^{\frac{1}{\beta}} \beta} \text{ as } t \rightarrow \infty.
 \end{aligned}$$

Similarly, for the lower bound, we can easily show

$$\sigma^2(t) \gtrsim (1 - \varepsilon) \frac{(\log t)^{\frac{1}{\beta}-1} \log 2}{\lambda^{\frac{1}{\beta}} \beta}$$

as $t \rightarrow \infty$. Now, combining the preceding asymptotic inequality with (6.24) and letting $\varepsilon \rightarrow 0$ imply the result for

$\beta \leq 1$. Similar arguments can be repeated for $\beta > 1$, we omit the details. \diamond

The following result will be used in proving Lemma 3.3.

LEMMA 6.1. For any $-1 < d < 0$, and $t > 0$,

$$\begin{aligned}
 \int_0^t \frac{1 - e^{-x}}{x} e^{(\frac{\log(t/x)}{\lambda})^{d+1}} \left(\log \frac{t}{x} \right)^d dx - \frac{e^{(\log t/\lambda)^{d+1}} \lambda^{d+1}}{d+1} \\
 \sim \gamma e^{(\log t/\lambda)^{d+1}} (\log t)^d \text{ as } t \rightarrow \infty.
 \end{aligned}$$

Proof: First, elementary calculations yield

$$\begin{aligned}
 \int_0^t \frac{1 - e^{-x}}{x} e^{(\frac{\log(t/x)}{\lambda})^{d+1}} \left(\log \frac{t}{x} \right)^d dx - \frac{e^{(\log t/\lambda)^{d+1}} \lambda^{d+1}}{d+1} \\
 = \int_0^1 \frac{1 - e^{-x}}{x} e^{(\frac{\log(t/x)}{\lambda})^{d+1}} \left(\log \frac{t}{x} \right)^d dx \\
 - \int_1^t \frac{e^{-x}}{x} e^{(\frac{\log(t/x)}{\lambda})^{d+1}} \left(\log \frac{t}{x} \right)^d dx - \frac{\lambda^{d+1}}{d+1} \\
 \triangleq I_1(t) - I_2(t) - \frac{\lambda^{d+1}}{d+1};
 \end{aligned}$$

note that these I_1 and I_2 are different from those in (3.7).

Then, by changing the variable of integration to $u = t/x$ in $I_1(t)$, we obtain

$$\begin{aligned}
 I_1(t) &= \left(\int_t^{t \log t} + \int_{t \log t}^\infty \right) \frac{1 - e^{-t/u}}{u} e^{(\frac{\log u}{\lambda})^{d+1}} (\log u)^d du \\
 (6.25) \quad &\triangleq I_{11}(t) + I_{12}(t).
 \end{aligned}$$

Next, it is easy to see that

$$\begin{aligned} I_{11}(t) &\leq e^{(\frac{\log(t \log t)}{\lambda})^{d+1}} (\log t)^d \int_t^{t \log t} \frac{1 - e^{-t/u}}{u} du \\ (6.26) \quad &\sim e^{(\frac{\log t}{\lambda})^{d+1}} (\log t)^d \int_0^1 \frac{1 - e^{-x}}{x} dx \text{ as } t \rightarrow \infty, \end{aligned}$$

since $-1 < d < 0$ and $\log \log t / (\log t)^{-d} \rightarrow 0$ for large t . Similarly,

$$\begin{aligned} I_{11}(t) &\geq e^{(\frac{\log t}{\lambda})^{d+1}} (\log(t \log t))^d \int_t^{t \log t} \frac{1 - e^{-t/u}}{u} du \\ (6.27) \quad &\sim e^{(\frac{\log t}{\lambda})^{d+1}} (\log t)^d \int_0^1 \frac{1 - e^{-x}}{x} dx \text{ as } t \rightarrow \infty. \end{aligned}$$

Also, by using $1 - e^{-x} \leq x$ for $x > 0$, we obtain

$$\begin{aligned} I_{12}(t) &\leq \int_{t \log t}^{\infty} e^{(\frac{\log u}{\lambda})^{d+1}} (\log u)^d \frac{1 - e^{-t/u}}{u} du \\ &\leq t \int_{t \log t}^{\infty} \frac{1}{u^2} e^{(\frac{\log u}{\lambda})^{d+1}} (\log u)^d du \\ &\leq 2t \int_{t \log t}^{\infty} \frac{1}{u^2} e^{(\frac{\log u}{\lambda})^{d+1}} du \\ &\sim \frac{2e^{(\frac{\log(t \log t)}{\lambda})^{d+1}}}{\log t} \\ (6.28) \quad &\sim \frac{2e^{(\frac{\log t}{\lambda})^{d+1}}}{\log t} = o(I_{11}(t)) \text{ as } t \rightarrow \infty, \end{aligned}$$

where second to the last relationship follows from Proposition 1.5.10 in [10].

Similarly, by breaking the integral at point $\log t$, we obtain

$$I_2(t) \lesssim e^{(\log t / \lambda)^{d+1}} (\log t)^d \int_1^{\infty} \frac{e^{-x}}{x} dx \text{ as } t \rightarrow \infty$$

and

$$\begin{aligned} I_2(t) &\geq e^{(\frac{\log(t/\log t)}{\lambda})^{d+1}} (\log t)^d \int_1^{\log t} \frac{e^{-x}}{x} dx \\ &\sim e^{(\log t / \lambda)^{d+1}} (\log t)^d \int_1^{\infty} \frac{e^{-x}}{x} dx \text{ as } t \rightarrow \infty. \end{aligned}$$

Combining the previous two asymptotic bounds for $I_2(t)$ with (6.25), (6.26), (6.27) and (6.28) yields

$$I_1(t) - I_2(t) \sim \gamma e^{(\log t / \lambda)^{d+1}} (\log t)^d \text{ as } t \rightarrow \infty,$$

which completes the proof of Lemma 6.1. \diamond

Proof of Lemma 3.3: Let us first consider the case $q_i = ce^{-\lambda(\log i)^\alpha}$, $\alpha > 1$. Since $1 - \exp(-cte^{-\lambda(\log i)^\alpha})$

is monotonically decreasing in i , we have

$$\begin{aligned} m(t) &= \sum_{i=1}^{\infty} \int_i^{i+1} (1 - \exp(-cte^{-\lambda(\log i)^\alpha})) du \\ &\geq \sum_{i=1}^{\infty} \int_i^{i+1} (1 - \exp(-cte^{-\lambda(\log u)^\alpha})) du \\ (6.29) \quad &= \int_1^{\infty} (1 - \exp(-cte^{-\lambda(\log u)^\alpha})) du \\ &\triangleq f(ct). \end{aligned}$$

Similarly,

$$(6.30) \quad m(t) \leq 1 + f(ct).$$

Changing the variable of integration to $x = ct \exp(-\lambda(\log u)^\alpha)$, we obtain

$$\begin{aligned} (6.31) \quad f(ct) &= \frac{1}{\alpha \lambda^{1/\alpha}} \int_0^{ct} \frac{1 - e^{-x}}{x} \exp\left(\left(\frac{\log(ct/x)}{\lambda}\right)^{1/\alpha}\right) \\ &\quad \times \left(\log \frac{ct}{x}\right)^{\frac{1}{\alpha}-1} dx. \end{aligned}$$

Next, using Lemma 6.1, combined with (6.29) and (6.30), we obtain

$$(6.32) \quad m(t) \sim e^{(\frac{\log t}{\lambda})^{1/\alpha}} \text{ as } t \rightarrow \infty,$$

which completes the proof for the case $q_i = ce^{-\lambda(\log i)^\alpha}$.

To prove the general case $q_i \sim ce^{-\lambda(\log i)^\alpha}$, for any $0 < \epsilon < c$, we choose $i_0 > 0$, such that for all $i \geq i_0$, $-\epsilon < q_i e^{\lambda(\log i)^\alpha} - c < \epsilon$. Hence,

$$-i_0 + f((c - \epsilon)t) \leq m(t) \leq i_0 + 1 + f((c + \epsilon)t),$$

which concludes the result for $m(t)$ since $f(c, t) \sim e^{(\frac{\log t}{\lambda})^{1/\alpha}}$ as $t \rightarrow \infty$ for any fixed c .

Next we prove the asymptotic result for $m^{-1}(t) \sim e^{-\gamma} c^{-1} e^{\lambda(\log t)^\alpha}$. In this regard, for any $\epsilon > 0$, choose i_0 such that for all $i > i_0$, $c(1 - \epsilon) \exp(-\lambda(\log i)^\alpha) \leq q_i \leq c(1 + \epsilon) \exp(-\lambda(\log i)^\alpha)$; let $c_\epsilon = (1 + \epsilon)c$, then as $t \rightarrow \infty$,

$$\begin{aligned} m(t) &\leq i_0 + 1 + f(c_\epsilon t) \\ &= \frac{\gamma e^{(\frac{\log c_\epsilon t}{\lambda})^{1/\alpha}} (\log c_\epsilon t)^{1/\alpha-1}}{\alpha \lambda^{1/\alpha}} (1 + o(1)) + e^{(\frac{\log c_\epsilon t}{\lambda})^{1/\alpha}}, \end{aligned}$$

where the last asymptotic relationship is implied by (6.31) and Lemma 6.1. Using the above expression, we obtain

$$\begin{aligned} m(e^{-\gamma} c_\epsilon^{-1} e^{\lambda(\log u)^\alpha}) &\leq \exp\left(\log u \left(1 - \frac{\gamma}{\lambda(\log u)^\alpha}\right)^{1/\alpha}\right) \\ &\quad \times \left(1 + \frac{\gamma}{\alpha \lambda (\log u)^{\alpha-1}} (1 - \gamma/(\lambda(\log u)^\alpha))^{1/\alpha-1}\right). \end{aligned}$$

Now, applying the fact that $(1 - y)^\eta = 1 - \eta y(1 + o(y^\delta))$ as $y \downarrow 0$ for $1/\alpha < \delta < 1$ in the preceding expression with $y = \gamma/(\lambda(\log u)^\alpha)$ and $\eta = 1/\alpha, \eta = 1/\alpha - 1$, we derive

$$\begin{aligned} & m(e^{-\gamma} c_\epsilon^{-1} e^{\lambda(\log u)^\alpha}) \\ & \leq u \exp\left(\frac{-\gamma(1 + o((\log u)^{-\alpha\delta}))}{\alpha\lambda(\log u)^{\alpha-1}}\right) \\ & \quad \times \left(1 + \frac{\gamma(1 - h(\log u)^{-\alpha})}{\alpha\lambda(\log u)^{\alpha-1}}\right) \\ & \leq u \exp\left(\frac{-\gamma(1 + o((\log u)^{-\alpha\delta}))}{\alpha\lambda(\log u)^{\alpha-1}} + \frac{\gamma(1 - h(\log u)^{-\alpha})}{\alpha\lambda(\log u)^{\alpha-1}}\right) \\ (6.33) \quad & = u \exp(o(-(\log u)^{-\alpha})) \text{ as } u \rightarrow \infty, \end{aligned}$$

where we use $1 + x \leq e^x$ in the second inequality and the last equality follows from $\alpha\delta > 1$.

Now, we define a new variable v equal to the expression in (6.33)

$$(6.34) \quad v = u \exp(o(-(\log u)^{-\alpha})) \text{ (as } u \rightarrow \infty).$$

This implies

$$\begin{aligned} u &= v e^{o((\log u)^{-\alpha})} \\ &\sim v e^{o((\log v)^{-\alpha})} \text{ as } v \rightarrow \infty \end{aligned}$$

since (6.34), in particular, yields $u \sim v$ as $u \rightarrow \infty$. Therefore,

$$\begin{aligned} e^{\lambda(\log u)^\alpha} &= \exp\left(\lambda\left(\log\left(v e^{o((\log v)^{-\alpha})}\right)\right)^\alpha\right) \\ &\sim e^{\lambda(\log v)^\alpha} \text{ as } v \rightarrow \infty. \end{aligned}$$

Applying $m^{-1}(\cdot)$ on both sides of (6.33) and using the last asymptotic relationship, we obtain

$$(6.35) \quad \liminf_{v \rightarrow \infty} m^{-1}(v) e^\gamma c \exp(-\lambda(\log v)^\alpha) \geq 1,$$

which proves the lower bound.

For the upper bound, we use

$$m(t) \geq f(c_\epsilon t) - i_0,$$

where $f(c_\epsilon t)$ is as defined in (6.29). Repeating almost the same arguments as for the lower bound, we obtain

$$\limsup_{v \rightarrow \infty} m^{-1}(v) e^\gamma c \exp(-\lambda(\log v)^\alpha) \leq 1,$$

which, combined with (6.35), yields the asymptotic result for $m^{-1}(t)$.

Now we prove the asymptotic result for $m'(t)$,

$$m'(t) \sim \frac{e^{(\frac{\log ct}{\lambda})^{1/\alpha}} (\log ct)^{1/\alpha-1}}{\alpha\lambda^{1/\alpha} t}.$$

By dominated convergence, from the definition of $m(t)$ it follows that

$$m'(t) = \sum_{i=1}^{\infty} q_i e^{-q_i t}.$$

Let us first assume $q_i = c \exp(-\lambda(\log i)^\alpha)$. Observe that for $t > 0$, the function $\exp(-\lambda(\log u)^\alpha - cte^{-\lambda(\log u)^\alpha})$ is increasing in u , for $u < \exp((\frac{\log ct}{\lambda})^{1/\alpha})$; it is decreasing for $u > \exp((\frac{\log ct}{\lambda})^{1/\alpha})$ and has its maximum e^{-1}/t for $u = \exp((\frac{\log ct}{\lambda})^{1/\alpha})$. Next, define $l(t) = \lfloor \exp((\frac{\log ct}{\lambda})^{1/\alpha}) \rfloor$, then

$$\begin{aligned} m'(t) &\leq \sum_{i=1}^{l(t)-1} q_i e^{-q_i t} + \sum_{l(t)+1}^{\infty} q_i e^{-q_i t} + \frac{e^{-1}}{t} \\ &\leq \left(\sum_{i=1}^{l(t)-1} \int_i^{i+1} + \sum_{l(t)+1}^{\infty} \int_{i-1}^i \right) q_i e^{-q_i t} + \frac{e^{-1}}{t} \\ &\leq \int_1^{\infty} q_i e^{-q_i t} + \frac{e^{-1}}{t} \\ &= g(t) + \frac{e^{-1}}{t}, \end{aligned}$$

where

$$g(t) \triangleq \int_1^{\infty} c e^{-\lambda(\log u)^\alpha} e^{-cte^{-\lambda(\log u)^\alpha}} du.$$

Similarly, we have the lower bound

$$m'(t) \geq g(t) - \frac{e^{-1}}{t}.$$

By changing variable $x = te^{-\lambda(\log u)^\alpha}$, we obtain

$$g(t) = \frac{c}{\alpha\lambda^{1/\alpha} t} \int_0^t e^{(\frac{\log \frac{t}{x}}{\lambda})^{1/\alpha}} e^{-cx} \left(\log \frac{t}{x}\right)^{1/\alpha-1} dx.$$

In order to complete the proof, it is enough to prove that

$$\begin{aligned} & \int_0^t e^{(\frac{\log \frac{t}{x}}{\lambda})^{1/\alpha}} e^{-cx} \left(\log \frac{t}{x}\right)^{1/\alpha-1} dx \\ & \sim e^{(\frac{\log t}{\lambda})^{1/\alpha}} (\log t)^{1/\alpha-1} \text{ as } t \rightarrow \infty. \end{aligned}$$

To finish this, we decompose the integral into three parts,

$$\begin{aligned} g(t) &= \int_0^{1/\log t} + \int_{1/\log t}^{\log t} + \int_{\log t}^t \\ &\triangleq I_1(t) + I_2(t) + I_3(t). \end{aligned}$$

First,

$$\begin{aligned} I_2(x) &\leq e^{(\frac{\log(t \log t)}{\lambda})^{1/\alpha}} \left(\log\left(\frac{t}{\log t}\right)\right)^{1/\alpha-1} \int_{1/\log t}^{\log t} e^{-x} dx \\ &\sim \frac{e^{(\frac{\log ct}{\lambda})^{1/\alpha}} (\log ct)^{1/\alpha-1}}{\alpha\lambda^{1/\alpha} t} \text{ as } t \rightarrow \infty; \end{aligned}$$

similarly for the lower bound,

$$I_2(t) \geq e^{\left(\frac{\log(t/\log t)}{\lambda}\right)^{1/\alpha}} (\log(t \log t))^{1/\alpha-1} \int_{1/\log t}^{\log t} e^{-x} dx \\ \sim \frac{e^{\left(\frac{\log ct}{\lambda}\right)^{1/\alpha}} (\log ct)^{1/\alpha-1}}{\alpha \lambda^{1/\alpha} t} \text{ as } t \rightarrow \infty.$$

For $I_1(t)$ we have the following estimate if we let $u = \log(t/x)$ and use Proposition 1.5.10 of [10], we derive

$$I_1(t) \leq t \int_{\log(t \log t)}^{\infty} e^{u^{1/\alpha}} e^{-u} u^{1/\alpha-1} du \\ \sim o(I_2(t)).$$

Similarly, one can prove that

$$I_3(t) = o(I_2(t)),$$

we omit the details. Thus, we complete the proof for the case $q_i = c \exp(-\lambda(\log i)^\alpha)$.

In the general case $q_i \sim c \exp(-\lambda(\log i)^\alpha)$ as $i \rightarrow \infty$, for any $\epsilon > 0$, we can choose i_0 such that for all $i > i_0$, $c(1 - \epsilon) \exp(-\lambda(\log i)^\alpha) \leq q_i \leq c(1 + \epsilon) \exp(-\lambda(\log i)^\alpha)$. Using this in conjunction with the proof for the case $q_i = c \exp(-\lambda(\log i)^\alpha)$, we obtain

$$m'(t) \leq \frac{1 + \epsilon}{1 - \epsilon} \frac{e^{\left(\frac{\log t}{\lambda}\right)^{1/\alpha}} (\log t)^{1/\alpha-1}}{\alpha \lambda^{1/\alpha} t} \text{ as } t \rightarrow \infty.$$

Similarly, we can obtain the lower bound. By passing $\epsilon \rightarrow 0$ we will prove the result for $m'(t)$.

Next, by repeating the similar arguments as in the proof of Lemma 3.2, we obtain, as $t \rightarrow \infty$,

$$\sigma^2(t) = m(2t) - m(t) \\ \sim \frac{e^{\left(\frac{\log t}{\lambda}\right)^{1/\alpha}} (\log t)^{1/\alpha-1} \log 2}{\alpha \lambda^{1/\alpha}},$$

which completes the proof. \diamond

References

- [1] 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.
- [2] 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.
- [3] P. Flajolet, D. Gardy, and L. Thimonier. Birthday paradox, coupon collector, caching algorithms and self-organizing search. *Discrete Applied Mathematics*, 39:207–229, 1992.
- [4] K. P. Gummadi, R. J. Dunn, S. Saroiu, S. D. Gribble, H. M. Levy, J. Zahorjan. Measurement, modeling, and analysis of a peer-to-peer file-sharing workload. In *Proceedings of the 19th ACM symposium on operating systems principles (SOSP-19)*, October 2003.
- [5] 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.
- [6] P. R. Jelenković, X. Kang, and A. Radovanović. Near optimality of the discrete persistent access caching algorithm. *Discrete Mathematics and Theoretical Computer Science*, AD:201–222, 2005.
- [7] P. R. Jelenković and A. Radovanović. Least-recently-used caching with dependent requests. *Theoretical Computer Science*, 326:293–327, 2004.
- [8] P. R. Jelenković and A. Radovanović. Optimizing LRU for variable document sizes. *Combinatorics, Probability & Computing*, 13:1–17, 2004.
- [9] P. R. Jelenković and A. Radovanović. The Persistent-Access-Caching Algorithm. Technical Report EE2004-03-05, Department of Electrical Engineering, Columbia University, New York, April 2004.
- [10] N. H. Bingham, C. M. Goldie and J. L. Teugels. *Regular Variation*. Cambridge University Press, 1987.