# Cache Miss Estimation for Non-Stationary Request Processes

Felipe Olmos, Carl Graham, Alain Simonian

# CACHE MISS ESTIMATION FOR NON-STATIONARY REQUEST PROCESSES

By Felipe Olmos[*,†], Carl Graham[†] and Alain Simonian[*]

*Orange Labs[*] and CMAP, École Polytechnique[†]*

The goal of the paper is to evaluate the miss probability of a Least Recently Used (LRU) cache, when it is offered a non-stationary request process given by a Poisson cluster point process. First, we construct a probability space using Palm theory, describing how to consider a tagged document with respect to the rest of the request process. This framework allows us to derive a fundamental integral formula for the expected number of misses of the tagged document. Then, we consider the limit when the cache size and the arrival rate go to infinity in proportion, and use the integral formula to derive an asymptotic expansion of the miss probability in powers of the inverse of the cache size. This enables us to quantify and improve the accuracy of the so-called *Che approximation*.

**1. Introduction.** Since the early days of the Web, cache servers have been used to provide users faster document retrieval while saving network resources. In recent years, there has been a renewed interest in the study of these systems, since they are the building bricks of *Content Delivery Networks* (CDNs), a key component of today's Internet. In fact, these systems handle nowadays around 60% of all video traffic, and it is predicted that this quantity will increase to more than 70% by 2019 [4]. Caches also play an important role in the emergent *Information Centric Networking* (ICN) architecture, that incorporates them ubiquitously into the network in order to increase its overall capacity [1].

In order to improve network efficiency, cache servers are placed close to the users, and store a subset of the *catalog* of available documents. Upon a user request for a document:

- If the document is already stored in the cache, then the cache uploads it directly to the user. This event is called a *cache hit*.
- Else, the request is forwarded to the repository server, which uploads a copy to the user, and possibly to the cache for future requests. This event is called a *cache miss*.

*MSC 2010 subject classifications:* 68B20, 60G55, 60K30

*Keywords and phrases:* performance evaluation, LRU cache policy, Poisson cluster process, Cox process, scaling limit expansion, Che approximation

Since the cost of storage is a constraint, each cache contains only a fraction of the document catalog, and needs to eliminate some documents to free space for new ones. Since the caches must decide to do so in real time, they use simple distributed elimination algorithms, called *cache eviction policies*.

Hereafter, we will focus our efforts on the *Least Recently Used* (LRU) cache eviction policy. To simplify the analysis, we will assume that all documents have the same size, and therefore that the disk of the cache can be represented as a list of documents of size $C \geq 1$. The LRU policy evicts content upon a user request as follows (see Fig. 1):

- If the requested document is already stored in the cache, then it is moved to the front of the list, while all documents which were in front of it are shifted down by one slot.
- Else, a copy of the requested document is downloaded from the server and placed at the front of the list, while all documents which were already in the list are shifted down by one slot except the last one which is eliminated.

Intuitively, this simple policy should perform well, since highly requested documents should stay near the front of the list, whereas unpopular ones should be quickly eliminated.

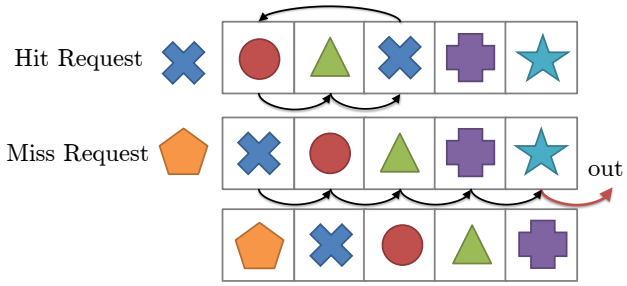



Fig 1: *The LRU eviction policy handling a hit and a miss request on a cache of size* $C = 5$.

Early theoretical studies on LRU caching performance further assumed that the catalog is fixed and finite, and that documents there have each an intrinsic probability to be requested independently, thus defining a popularity distribution. The request process is then modeled as an i.i.d. sequence, where at each time step a document is requested according to its popularity. This framework is commonly referred to as the *Independent Reference Model* (IRM) in the literature, see for instance [13].

While the IRM setting has been proved to be a good model for short time-scales, it is not accurate for larger ones. In fact, other phenomena occurring within longer time-scales must also be taken into account, notably the dynamic nature of the catalog and of user preferences.

In order to capture these phenomena, a new model based on *Poisson cluster point processes* has been independently proposed by Traverso et al. [20] and Olmos et al. [19]. It allows to address the catalog and preference dynamics, and thus to obtain more accurate results in both large and small time scales. Its properties have received only heuristic analysis in these works.

The object of the present paper is to build a sound mathematical framework for the analysis of this model, and to provide rigorous proofs for the estimation of the hit probability, which corresponds to the asymptotic proportion of cache hits among all requests when the number of these requests goes to infinity, or equivalently of the complementary miss probability.

Before describing our main contributions, we briefly review the literature on caching performance, mentioning only papers relevant to our present work; see [10] and the references therein for a more comprehensive bibliography of the subject. The modern treatment of the subject started with Fill and Holst [9], which introduced the embedding of the request sequence into a marked Poisson process, in order to analyze the related problem of the search cost for the *Move-to-Front* list. Independently, Che et al. [3] also used marked Poisson processes to model the requests. In their work, they express the hit probability of a LRU cache in terms of a family of exit times of the documents from the cache. In order to simplify the analysis, they approximated this family by a single constant called the *characteristic time*. This heuristic, called the *Che approximation* in the literature, proved to be empirically accurate even outside of its original setting. The question of quantifying the error incurred in the approximation has been partially answered by Fricker et al. [13], where the authors provide a justification for a Zipf popularity distribution when the cache size $C$ grows to infinity and scales linearly with the catalog size. The error incurred by the approximation is estimated for the exit times, but not, however, for the hit probability.

In the present paper, we succeed in adapting the *Che approximation* to the more complex setting of the cluster point process model. The approximation accuracy has been considered first by Leonardi and Torrisi [16], which provide limit theorems for the exit time as $C$ goes to infinity, as well as an upper bound of the error on the hit probability. However, the latter bound depends on an additional variable, of which the optimal value is not explicitly given in terms of system parameters.

The contribution of our paper is threefold. Firstly, in Sections 2 and 3, we

use the Palm distribution for the system in order to provide a probability space where an "average document" can be tagged and analyzed independently from the rest. Secondly, in Section 4, we use the latter independence structure to obtain an integral formula for the expected number of misses for a document, generalizing the development in [19]. Thirdly, in Section 5, using scaling methods, we deduce from this formula an asymptotic expansion for the average number of misses, showing that the error term in the *Che approximation* is of order $O(1/C)$. In contrast to the upper bound provided in [16], our error estimation depends simply on system parameters and can be readily calculated. Section 6 is devoted to a numerical study validating the accuracy of the asymptotic expansion. Section 7 contains some concluding remarks, and Section 8 contains all proofs.

**2. Document Request Model.** Our request model consists in the following cluster point process on the real line $\mathbb{R}$, illustrated in Fig. 2.

A ground process $\Gamma^g$, hereafter called the **catalog arrival process**, gives the consecutive arrival times of new documents to the catalog. We assume it to be a homogeneous Poisson process with intensity $\gamma > 0$, and denote its generic arrival time by $a$.

The cluster at an arrival time $a$ of $\Gamma^g$ is denoted by $\xi_a$, and is an element of the space $\mathcal{M}^{\#}(\mathbb{R})$ of point processes on $\mathbb{R}$. It represents the **document request process** for the document arriving to the catalog at that time $a$. We assume that $\xi_a$ is a Cox process directed by a stochastic intensity function $\lambda_a \geq 0$ having the following properties.

- Given $\Gamma^g$, the intensities $\lambda_a$ for $a \in \Gamma^g$ are jointly independent.
- The intensities $\lambda_a$ are *causal*: each function $t \mapsto \lambda_a(t)$ is zero for $t < a$. Requests for a document thus occur only after its arrival at the catalog.
- The distribution of $\lambda_a$ is *stationary*: for each arrival time $a \in \mathbb{R}$, the processes $\lambda_a(\cdot)$ and $\lambda_0(\cdot - a)$ have the same distribution.

These three conditions allow to sample the sequence $(\lambda_a)_{a\in\Gamma^g}$ using independent samples of a **canonical intensity function** $\lambda$ with support in $[0, \infty)$, adequately shifted to every arrival time $a$.

For a document arriving at time $a$, we denote by $\Lambda_a$ both the **mean function** associated to the request intensity $\lambda_a$ and the **average number of requests** (with abuse of notation for conciseness)

$$\Lambda_a(t) = \int_a^t \lambda_a(u)\,\mathrm{d}u\,, \quad t \geq a\,, \qquad \Lambda_a = \Lambda_a(\infty)\,.$$

We assume that $\Lambda_a < \infty$ almost surely, and denote by $\bar{\Lambda}_a$ the **complemen-**

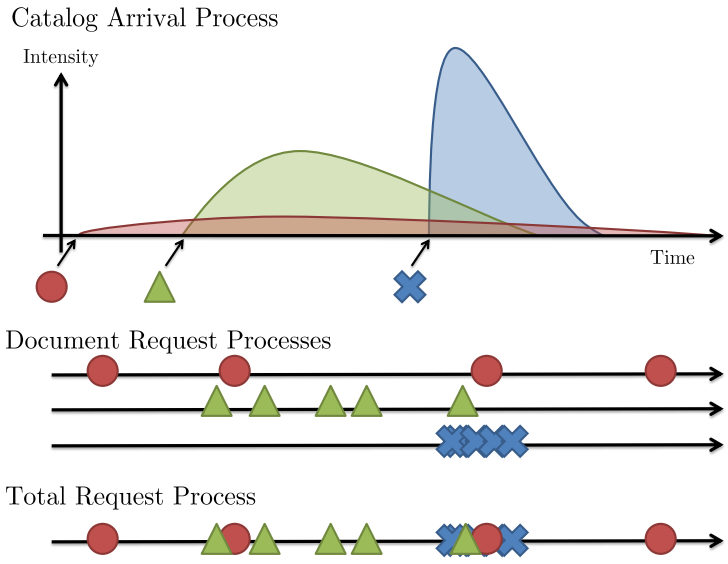



Fig 2: *A sample of the document arrival and request processes.* ***Top****: Each catalog arrival triggers a function representing the request intensity for the corresponding document.* ***Bottom****: A sample of the document request processes. Their superposition generates the total request process.*

**tary mean function**

$$\bar{\Lambda}_a(t) = \Lambda_a - \Lambda_a(t) = \int_t^\infty \lambda_a(u)\,\mathrm{d}u\,, \quad t \geq a\,.$$

When referring to the canonical document, which corresponds to an arrival at time zero, we remove the time index $a$; for instance we write $\Lambda$ and $\bar{\Lambda}(t)$. The superposition of all processes $\xi_a$ for $a \in \Gamma^g$ given by

$$\Gamma = \sum_{a \in \Gamma^g} \xi_a$$

constitutes the **total request process** for all documents. We assume that

$$\int_{-\infty}^{t} \mathbb{E}\Big[1 - e^{-(\Lambda_a(t) - \Lambda_a(s))}\Big]\,\mathrm{d}a < \infty \tag{1}$$

for all times $s \leq t$. This is a necessary and sufficient condition for the process $\Gamma$ to be locally finite almost surely, see [5, Theorem 6.3.III].

**3. Tagging a Document via Palm Theory.** The key of our analysis is to *tag* one document of the system and treat the remaining process as an external *environment*. To do this, we follow [6, p.279]. Let $Q_{u,\nu}$ be the local Palm distribution at point $(u,\nu)$ in $\mathbb{R}\times\mathcal{M}^{\#}(\mathbb{R})$ for the point process

$$\widetilde{\Gamma}=\sum_{a\in\Gamma^g}\delta_{a,\xi_a}$$

constituted by the ground process $\Gamma^g$ marked with the document request processes, which constitutes a Poisson point process on $\mathbb{R}\times\mathcal{M}^{\#}(\mathbb{R})$. Define the mark-averaged Palm distribution $\overline{Q}_u$ on $\mathcal{M}^{\#}(\mathbb{R})$ by

$$\overline{Q}_u(\cdot)=\mathbb{E}[Q_{u,\xi_u}(\cdot)]\,.$$

Under this distribution $\overline{Q}_u$, the process has the structure given by the following proposition, illustrated by Fig. 3.

**Proposition 1** (Palm Decomposition for Tagged Document)
*Under the distribution $\overline{Q}_u$, the process $\tilde{\Gamma}$ has almost surely a point at time $u$. Furthermore:*

- *The distribution of the mark $\xi_u$ is the same as the original one.*
- *The distribution of the remaining process $\tilde{\Gamma}\setminus\delta_{u,\xi_u}$ is the same than that of the original process $\tilde{\Gamma}$.*
- *The mark $\xi_u$ and the process $\tilde{\Gamma}\setminus\delta_{u,\xi_u}$ are independent.*

We refer to Section 8.1 for the proof of this proposition.

Proposition 1 allows us to consider a probability space for which there is a document arrival at time $a=0$, almost surely. We call this document the **tagged** document, and the complementary process **the rest**.

In the next section, we shall see that for the LRU caching discipline, the independence of the tagged document from the rest allows us to derive a general integral formula for the miss probability.

**4. Fundamental Integral Formula.** As stated in the previous section, we will consider a tagged document at time zero, so that its associated distribution is the canonical one. For a LRU cache with size $C$, let $N$ and $\mu_C$ be the random number of requests and number of misses for the tagged document, respectively. The total miss probability is defined by

$$p_C=\frac{\mathbb{E}[\mu_C]}{\mathbb{E}[N]}\,,$$

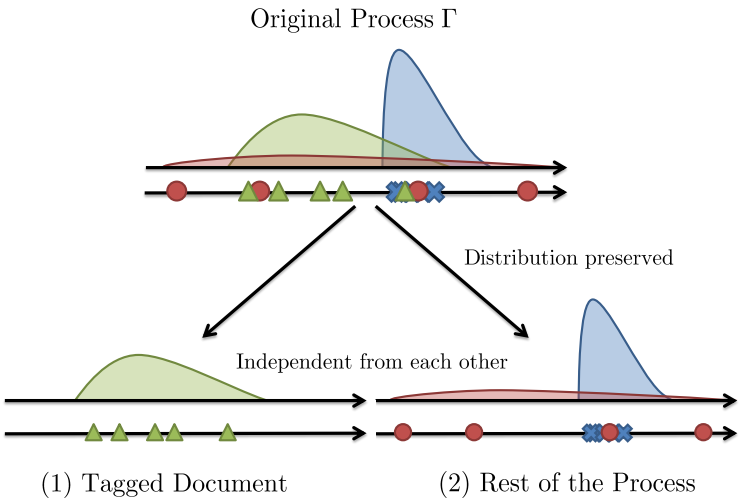



Fig 3: *Illustration of request process $\xi$ under the averaged Palm distribution. The original process is decomposed into: (1) the tagged document and (2) the rest of the process. These are mutually independent, and the rest of the process has the same distribution as the original process.*

which is also the average per-document miss probability $\mu_C/N$ under the size biased distribution of $N$. The mixed Poisson random variable $N$ with random mean $\Lambda$ has expectation

$$\mathbb{E}[N] = \mathbb{E}[\mathbb{E}[N \,|\, \Lambda]] = \mathbb{E}[\Lambda]\,,$$

and it remains to study $\mu_C$.

Let $(\Theta_r)_{r=1}^N$ be the sequence of request times for the tagged document, with the understanding that it is the empty set if $N = 0$. The first request being always a miss, the number of misses can be written as

$$\mu_C = \mathbb{1}\{N \geq 1\} + \mathbb{1}\{N \geq 2\} \sum_{r=2}^{N} \mathbb{1}\{\text{Request at } \Theta_r \text{ is a miss}\}. \tag{2}$$

Under the LRU policy, a document requested at time $s$ will be erased from the cache at the first time, after the last request for this document, that $C$ distinct other documents have been requested.

For each $s$ in $\mathbb{R}$, let us define the process $X^s = (X_t^s)_{t \geq s}$ which counts the number of **distinct** documents in the **rest** of the process which are requested

on the time interval $[s,t]$, and its **exit time** $T_C^s$ to level $C$. Hence, $T_C^s$ is the time that a document requested at time $s$ spends in the cache before being evicted. Denoting by $F^s(\xi_a)$ the first arrival time of $\xi_a$ in $[s,\infty)$, the process $X^s = (X_t^s)_{t\geq s}$ and exit time $T_C^s$ can be expressed as

$$\begin{cases} X_t^s = \#\{(a,\xi_a) \text{ in } \tilde{\Gamma} \setminus \delta_{0,\xi_0} : F^s(\xi_a) \leq t\}\,, \\ T_C^s = \inf\{t \geq s : X_t^s = C\}\,. \end{cases} \tag{3}$$

These definitions allow us to express the miss events as

$$\{\text{Request at } \Theta_r \text{ is a miss}\} = \left\{X_{\Theta_r}^{\Theta_{r-1}} \geq C\right\} = \left\{\Theta_r > T_C^{\Theta_{r-1}}\right\}\,, \quad r \geq 2\,,$$

since such a miss occurs if and only if at least $C$ distinct other documents have been requested in the interval $[\Theta_{r-1},\Theta_r]$. Hence (2) can be written as

$$\mu_C = \mathbb{1}\{N \geq 1\} + \mathbb{1}\{N \geq 2\} \sum_{r=2}^{N} \mathbb{1}\left\{\Theta_r > T_C^{\Theta_{r-1}}\right\}. \tag{4}$$

To proceed further, we study the consequences of the structure of the cluster point process on the structure of the families $X^s$ and $T_C^s$.

**Proposition 2** (Characterization of $X^s$ and $T_C^s$)
*Let $s$ be in $\mathbb{R}$. The process $X^s = (X_t^s)_{t\geq s}$ defined by (3) is an inhomogeneous Poisson process with intensity function*

$$\Xi^s(t) = \mathbb{E}[X_t^s] = \gamma \int_{-\infty}^{t} \mathbb{E}\left[1 - e^{-(\Lambda_a(t)-\Lambda_a(s))}\right] \mathrm{d}a\,, \quad t \geq s\,. \tag{5}$$

*In particular, $T_C^s - s \overset{d}{=} T_C$, where $T_C = T_C^0$ is the exit time of a document requested at time zero.*

We refer to Section 8.2 for the proof.

Equation (4), Proposition 2, and the independence between the tagged document and the rest of the process now yield an integral formula for $\mathbb{E}[\mu_C]$.

**Theorem 3** (Integral Formula for Expected Misses)
*The expected number of misses is given by*

$$\mathbb{E}[\mu_C] = \mathbb{E}[m(T_C)] \tag{6}$$

*where $T_C = T_C^0$ denotes the exit time for a document requested at time zero, see (3), and the function $m$ is defined using the notation in Section 2 by*

$$m(t) = \mathbb{E}\left[\int_0^\infty \lambda(u) e^{-(\Lambda(u+t)-\Lambda(u))}\,\mathrm{d}u\right], \quad t \geq 0\,. \tag{7}$$

*Moreover, $\lim_{t\to\infty} \downarrow m(t) = m_0$, where $m_0 = \mathbb{E}\left[1 - e^{-\Lambda}\right]$ with $\Lambda = \Lambda(\infty)$.*

The proof is postponed to Section 8.3. It uses the following result of independent interest.

**Proposition 4** (Functionals of Holding Times)
*Let $\xi$ be an inhomogeneous Poisson process on $[0,\infty)$ with* deterministic *intensity function $\lambda$. Let the mean function $\Lambda$ satisfy $\Lambda(\infty) < \infty$, so that $\xi$ has a finite random number $N$ of points $(\Theta_r)_{r=1}^N$. Then, for any $F : \mathbb{R}^+ \to \mathbb{R}$,*

$$\mathbb{E}\left[\mathbb{1}\{N \geq 2\} \sum_{r=2}^{N} F(\Theta_r - \Theta_{r-1})\right] = \int_0^\infty \mathrm{d}w\, F(w) \int_0^\infty \mathrm{d}u\, \lambda(u)\lambda(u+w) e^{-(\Lambda(u+w)-\Lambda(u))} \,.$$

We refer to Section 8.4 for the proof of this proposition.

The above analysis would identically apply if the random variable $T_C$ were deterministic and equal to some positive constant $t$. This would correspond to the cache discipline known as *Time to Live (TTL)*, where the cache evicts a document after a fixed amount of time $t$. Therefore, $m(t)$ is simply the average number of misses for a TTL cache of eviction time $t$. We can thus regard the number of misses in a LRU cache as a time randomization of the misses in a TTL cache.

Indeed, the integral formula (7) in Theorem 3 can by rewritten using integration by parts as

$$m(t) = \mathbb{E}\left[\int_0^\infty \lambda(u)\, e^{-(\Lambda(u)-\Lambda(u-t))}\, \mathrm{d}u\right]$$

which can be informally interpreted as follows: The exponential term

$$e^{-(\Lambda(u)-\Lambda(u-t))}$$

is simply the conditional probability $\mathbb{P}[\xi[u-t,u] = 0 \,|\, \lambda]$. Thus a request at time $u$ will contribute to the intensity of the miss process if there were no requests in the interval $[u-t,u]$, which is exactly a miss event in a $t$-TTL cache. This relationship between the miss probabilities of TTL and LRU caches has been already noted by Fofack et al. [11].

**5. Asymptotic Expansion.** It holds that $\lim_{t\to\infty} \downarrow m(t) = m_0$, see Theorem 3. Moreover, Proposition 2 yields that the exit time $T_C$ increases to infinity with $C$. Hence, (6) and dominated convergence yield that

$$\lim_{C\to\infty} \mathbb{E}[\mu_C] = m_0 \,.$$

This formula is not informative, since it basically tells us that the first request for a document is the unique miss for infinite capacity.

A more interesting way to derive asymptotics for $\mathbb{E}[\mu_C]$ is to scale some system parameters with respect to $C$. An intuitively good choice is to scale the catalog arrival rate $\gamma$ proportionally to the cache size $C$. In the following, with help of the results of the previous sections, we shall provide an asymptotic expansion for $\mathbb{E}[\mu_C]$ as $C$ grows large in this scaling.

The canonical exit time $T_C$ is the first passage time to level $C$ of an inhomogeneous Poisson process with mean function $\Xi = \Xi^0$, see Proposition 2. To pursue the analysis, we first prove a key relation between $\Xi$ and $m$.

**Proposition 5** (Relation between $\Xi$ and $m$)
*The functions $\Xi$ and $m$ in (5) and (7) satisfy $\Xi'(t) = \gamma\, m(t)$, and hence*

$$\Xi(t) = \gamma\, M(t)\,, \quad \textit{where} \;\; M(t) = \int_0^t m(s)\,\mathrm{d}s\,, \qquad t \geq 0\,.$$

We refer to Section 8.5 for the proof.

Proposition 5 implies that $\Xi(t) = y \Leftrightarrow M(t) = y/\gamma$ and thus that

$$\Xi^{-1}(y) = M^{-1}\left(\frac{y}{\gamma}\right), \quad y \geq 0\,, \tag{8}$$

(for definiteness, we consider left-continuous inverses). Moreover, the exit time $T_C$ is the first passage time to level $C$ of an inhomogeneous Poisson process with mean function $\Xi$ (see Proposition 2), and can be expressed as

$$T_C = \Xi^{-1}(\widehat{T}_C) \tag{9}$$

where $\widehat{T}_C$ is the first passage time to level $C$ of an unit Poisson process and has a Gamma$(C, 1)$ distribution. From Theorem 3 and (9), we derive that $\mathbb{E}[\mu_C] = \mathbb{E}[m(T_C)] = \mathbb{E}\Big[m(\Xi^{-1}(\widehat{T}_C))\Big]$, and (8) eventually yields that

$$\mathbb{E}[\mu_C] = \mathbb{E}\left[m\left(M^{-1}\left(\frac{\widehat{T}_C}{\gamma}\right)\right)\right]. \tag{10}$$

Now, the strong law of large numbers yields that $\lim_{C\to\infty} \widehat{T}_C/C = 1$ almost surely, and thus (10) strongly suggests to consider the scaling

$$C = \gamma\theta \;\; \text{for some} \;\; \theta > 0\,. \tag{11}$$

This scaling is quite natural, since Little's law ([2, Section 3.1.2]) applied to the cache system yields that $C = \gamma\,\mathbb{E}\big[T_C^{\text{in}}\big]$, where

$$T_C^{\text{in}} = \int_0^\infty \mathbb{1}\{\text{Object is in the cache at } t\}\,\mathrm{d}t$$

is the sojourn time of an object in the cache. Note that we do consider the objects without any requests as entering the system, but we set their sojourn time to $T_C^{\text{in}} = 0$.

As a consequence, the asymptotic analysis under the scaling (11) amounts to fixing the average sojourn time $\theta = \mathbb{E}\big[T_C^{\text{in}}\big] = C/\gamma$ and the distribution of the canonical intensity function $\lambda$ while letting $C$ grow to infinity.

Under the scaling (11), eq. (10) and $\lim_{C\to\infty} \widehat{T}_C/C = 1$ a.s. imply using dominated convergence that

$$\lim_{C\to\infty} \mathbb{E}[\mu_C] = m(t_\theta)\,, \quad t_\theta = M^{-1}(\theta)\,.$$

In the following, the quantity $t_\theta$ will be called the **characteristic time**. The asymptotics of $\mathbb{E}[\mu_C]$ will be expressed in terms of $t_\theta$. In this aim, we first recall two basic results regarding the Gamma$(C, 1)$ distribution.

**Lemma 6** (Classical Bounds on Gamma Laws)
*Let $\widehat{T}_C$ follow a* Gamma$(C, 1)$ *distribution, and $X_C = \widehat{T}_C/C$. Then:*

(i) *For any $C > 1$ and $\eta > 0$,*

$$\mathbb{P}[|X_C - 1| \geq \eta] \leq 2e^{-C\cdot\varphi(1+\eta)}\,,$$

*where $\varphi(x) = x - 1 - \log x$ is the large deviations rate function for the law of large numbers for exponential random variables of mean* 1.

(ii) *For any $C > 1$ and $k > 1$,*

$$\mathbb{E}\Big[(X_C - 1)^k\Big] = O(C^{-\lceil k/2\rceil})\,.$$

We refer to Section 8.6 for the classical proofs. We now formulate our central result concerning the asymptotics for the average number of misses.

**Theorem 7** (Expected Number of Misses Expansion)
*Assume that the function $m$ is twice continuously differentiable in $(0, \infty)$. Let $t_\theta = M^{-1}(\theta)$ (see Proposition 5) and*

$$e(t_\theta) = \frac{\theta^2}{2m(t_\theta)^2}\left(m''(t_\theta) - \frac{m'(t_\theta)^2}{m(t_\theta)}\right).$$

*Then, as $C$ goes to infinity with the scaling $C = \gamma\theta$ for fixed $\theta > 0$, we have*

$$\mathbb{E}[\mu_C] = m(t_\theta) + \frac{e(t_\theta)}{C} + o\left(\frac{1}{C}\right). \tag{12}$$

We refer to Section 8.7 for the proof.

Theorem 7 justifies the accuracy of the estimations that use the *Che approximation*. In the present setting, this heuristic consists in replacing the exit time $T_C$ in (6) by the constant $\widetilde{t}_C = \Xi^{-1}(C)$, therefore estimating $\mathbb{E}[\mu_C]$ by $m(\widetilde{t}_C)$. Now, under the scaling $C = \gamma\theta$, the identity (8) entails that

$$\widetilde{t}_C = \Xi^{-1}(C) = M^{-1}\left(\frac{C}{\gamma}\right) = M^{-1}(\theta) = t_\theta.$$

The quantity $\widetilde{t}_C$ is called in the literature the "characteristic time", and this identity justifies this naming for $t_\theta$ as well. More importantly, the asymptotic expansion of $\mathbb{E}[\mu_C]$ in Theorem 7 shows that the error in the *Che approximation* is of order $1/C$ and specifies it precisely, for large $C$ and fixed average sojourn time $\theta$.

**Remark 8** (Higher Order Expansions)
*If the function $m$ has derivatives of higher order, the proof of Theorem 7 together with Lemma 6 allow us to derive higher order expansions of $\mathbb{E}[\mu_C]$ in powers of $1/C$. Specifically, to obtain an expansion at order $n$, we must expand $f_\theta$ to the $2n$-th order, since $\mathbb{E}\left[(X_C - 1)^k\right]$ is $O(1/C^{\lceil k/2 \rceil})$ by Lemma 6. We then eventually obtain*

$$\mathbb{E}[\mu_C] = \sum_{k=0}^{2n} \frac{f_\theta^{(k)}(1)}{j!} \frac{\phi_k(C)}{C^k} + o\left(\frac{1}{C^n}\right)$$

*where $\phi_k$ is a polynomial of degree $\lfloor k/2 \rfloor$, as shown in the proof of Lemma 6.*

**Remark 9** (Laplace Asymptotic Method)
*Theorem 7 can be proved by purely analytical methods. Indeed, (20) can be written in integral form, after using the change of variables $w \mapsto w/C$, as*

$$\mathbb{E}[\mu_C] = \frac{C^C}{\Gamma(C)} \int_0^\infty e^{-C(w - \log(w))} \frac{f_\theta(w)}{w} \, \mathrm{d}w \, .$$

*Theorem 7 then follows by expanding this integral using the Laplace method (see [18, (3.15)]) and $\Gamma(C)$ using the Stirling formula. The expansion of the numerator must be performed through a Taylor series of function $f_\theta$ around the extremal point of the argument $w - \log(w)$ of the exponential term, that is, near $w = 1$. This method is, however, more complicated, since it involves the expansion of both numerator and denominator in powers of $\sqrt{C}$.*

The smoothness assumptions on the function $m$ in Theorem 7 can usually be checked readily on a case by case basis, by justifying interchange of derivation and expectation in (7) using dominated convergence. Nevertheless, it is difficult to give a general result.

To conclude this section, we show that these smoothness assumptions hold for a class of random intensities $\lambda$ which is suitable for modeling purposes. This class is built by randomly scaling a deterministic shape function in both domain and range. It includes the families used in previous works [20, 19].

**Proposition 10** (Twice Continuously Differentiable Example)
*Let $f \in \mathcal{C}^1(0,\infty)$ be a strictly positive unimodal function satisfying that $\int f = 1$, $\int f^2 < \infty$, and $\int |f'| < \infty$. Let $(R, L)$ be a couple of positive random variables with a smooth joint density, satisfying that $\mathbb{E}[R] < \infty$ and $\mathbb{E}[RL] < \infty$. If the canonical document request intensity is of the form*

$$\lambda(u) = R \cdot f\left(\frac{u}{L}\right), \quad u \geq 0, \tag{13}$$

*then the function $m$ is $\mathcal{C}^2(0,\infty)$ with derivatives given for $t > 0$ by*

$$\begin{cases} m'(t) = -\mathbb{E}\left[R^2 L \int_0^\infty f(u) f\left(u + \frac{t}{L}\right) e^{-RL\left(F\left(u+\frac{t}{L}\right)-F(u)\right)} \, \mathrm{d}u\right], \\ m''(t) = \mathbb{E}\left[R^3 L \int_0^\infty f(u) f\left(u + \frac{t}{L}\right)^2 e^{-RL\left(F\left(u+\frac{t}{L}\right)-F(u)\right)} \, \mathrm{d}u\right] \\ \qquad\quad - \mathbb{E}\left[R^2 \int_0^\infty f(u) f'\left(u + \frac{t}{L}\right) e^{-RL\left(F\left(u+\frac{t}{L}\right)-F(u)\right)} \, \mathrm{d}u\right], \end{cases} \tag{14}$$

*where $F(u) = \int_0^u f(v)dv$.*

We defer the proof of the proposition to Section 8.8.

Note that Proposition 10 only imposes mild conditions on the distribution of $(R, L)$. The admitted shape functions $f$ include exponential and power law decreasing profiles, and Gaussian curves restricted to $[0,\infty)$. In addition, the assumption of $f$ being strictly positive on $[0,\infty)$ can be weakened to that of being positive only in a compact interval; this in turn implies that $f'$ is not differentiable everywhere and the second derivative of $m$ will thus contain additional terms from the integral of $f'$. These terms can be obtained by integration by parts (see [12, Th. 3.36] for a generalized form).

One example of such a family with compact support is given by the "Box Model", previously analyzed in [19], which can be constructed by simply

taking $f = \mathbb{1}_{[0,1]}$. In this case, $m$ and its derivatives reduce to

(15)
$$\begin{cases} m(t) = \mathbb{E}\left[\left(1 - e^{-RL}\right)\mathbb{1}_{\{L \le t\}} + \left(1 - e^{-Rt} + R(L-t)e^{-Rt}\right)\mathbb{1}_{\{L>t\}}\right], \\ m'(t) = -\mathbb{E}\left[R^2(L-t)e^{-Rt}\,\mathbb{1}_{\{L>t\}}\right], \\ m''(t) = \mathbb{E}\left[(R^2 + R^3(L-t))e^{-Rt}\,\mathbb{1}_{\{L>t\}}\right]. \end{cases}$$

We will use this model for a numerical illustration in the next section.

**6. Numerical Experiments.** We provide some numerical results to validate the accuracy of asymptotic expansion (12), by comparing it to the values obtained from the system simulation. In our experiments, we used the "Box Model" in which the canonical intensity function is given by

$$\lambda(u) = R \cdot \mathbb{1}\{0 \le u \le L\}, \quad u \ge 0,$$

where the random pair $(R, L)$ represents the request rate and lifespan of a document. In view of (12), we obtain the zero order and first order approximations for the hit probability $q_C$, namely

$$q_C = 1 - p_C = 1 - \frac{\mathbb{E}[\mu_C]}{\mathbb{E}[\Lambda]} \approx \begin{cases} 1 - \dfrac{m(t_\theta)}{\mathbb{E}[\Lambda]}, & \text{0-th Order} \\ 1 - \dfrac{m(t_\theta) + e(t_\theta)/C}{\mathbb{E}[\Lambda]}, & \text{1-st Order} \end{cases} \tag{16}$$

where $\mathbb{E}[\Lambda] = \mathbb{E}[RL]$.

For a given general distribution of $(R, L)$, we cannot deduce explicit expressions for $m, m', m'', M$, and $M^{-1}$ from (15). In particular, there is usually no formula for $t_\theta$ in terms of $\theta$. In consequence, we resorted to numerical integration and inversion to obtain the hit probability estimates in (16).

As argued in [19], actual data traces suggest that the distributions of variable $R$ and $L$ are heavy tailed with infinite variance, that is, with tail index $\alpha \in (1, 2)$. For our experiments, we consequently chose $R$ and $L$ to be distributed as independent Pareto-Lomax variables, with probability density $\alpha\sigma^\alpha/(\sigma + x)^{\alpha+1}$ for $x > 0$, with respective parameters $(\alpha = 1.9, \sigma = 22.5)$ and $(\alpha = 1.7, \sigma = 0.07)$. Such values have been taken so that the simulation time is not excessive; they provide a "box" of average width 0.1 and height 25 with high volatility since neither $R$ nor $L$ have a finite variance.

We generated the request process associated with these intensity functions for various values of $\gamma$ ranging from 10 to 1,000. For each request sequence, we simulated an LRU cache and obtained the empirical hit probability for various capacities $C$.

To obtain reliable results, the heavy tailed nature of the input distributions requires to use the stable-law central limit theorem (see [21, Th. 4.5.1]). Specifically, there exists a so-called stable law $S_\alpha(\sigma,\beta,\mu)$ with scaling parameter $\sigma$ and a constant $K_\alpha$ such that, in distribution,

$$\lim_{n\to\infty} \frac{1}{K_\alpha}\frac{1}{n^{1/\alpha}}\sum_{i=1}^{n}(L_i - n\mathbb{E}[L]) = S_\alpha(1,1,0)\,.$$

This allows to heuristically quantify the convergence rate for the law of large numbers by considering that

$$\frac{1}{n}\sum_{i=1}^{n}(L_i - n\mathbb{E}[L]) \underset{n\to\infty}{\approx} S_\alpha\left(\frac{K_\alpha}{n^{1-1/\alpha}},1,0\right)$$

(in the present case, $\alpha = 1.7$ for $L$). We then chose the simulation time $S$ such that the average number of observed documents $n = \gamma S \times \mathbb{E}\left[1-e^{-RL}\right]$ is such that scaling parameter $K_\alpha/n^{1-1/\alpha}$ is smaller than $10^{-3}$ (such a value of $n$ ensures the same accuracy for the request rate $R$ with larger tail index $\alpha = 1.9$). Besides, we also chose $S$ large enough to ensure that there is enough time for all observable documents to appear in the simulated trace.

We show in Fig. 4 some of the resulting hit probability curves from these experiments. We observe that the zero order approximation in (16) is almost exact already for $\gamma = 500$. The error incurred by the approximation for lower $\gamma$ can be corrected by using the first order approximation in (16), as shown in Fig. 5a for $\gamma = 50$. For even lower intensities, this correction might not be enough to approximate the real hit probability, as illustrated in Fig. 5b for $\gamma = 10$; the higher order expansion of Remark 8 would then be needed.

The above numerical results therefore illustrate the accuracy of the asymptotic expansion for the hit probability.

**7. Concluding Remarks.** In this paper, we have estimated the hit probability of a LRU cache for a traffic model based on a Poisson cluster point process. In this endeavor, we have built using Palm theory a probability space where a tagged document can be analyzed independently from the rest of the process. In the case of the LRU replacement policy, this property is key for the analysis, since it allowed us to derive an integral expression for the expected number of misses of the tagged object.

Using this expression, we were able to obtain an asymptotic expansion of this integral for large $C$ under the scaling $C = \gamma\theta$ for fixed $\theta > 0$. This expansion quantifies rigorously and in precise fashion the error made when

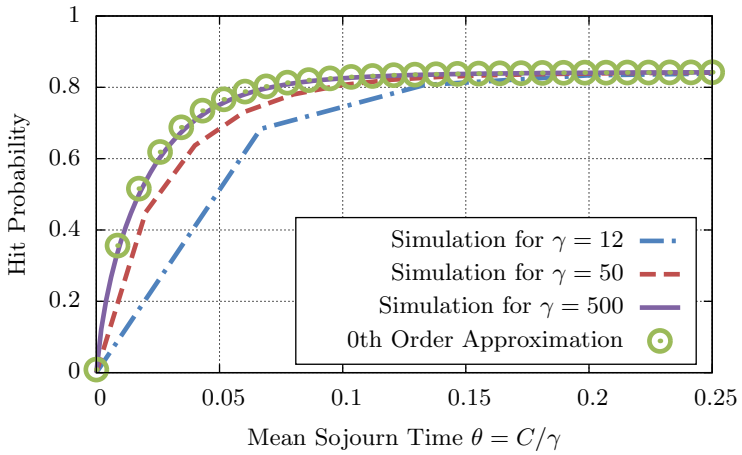
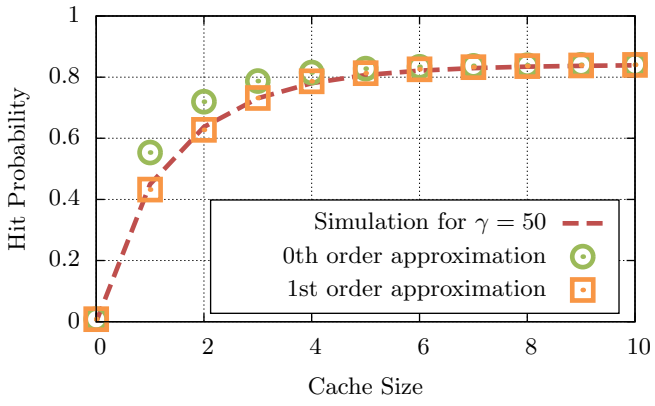



Fig 4: *Convergence of the hit probability curves obtained in the experiments to the 0th order Che approximation.*

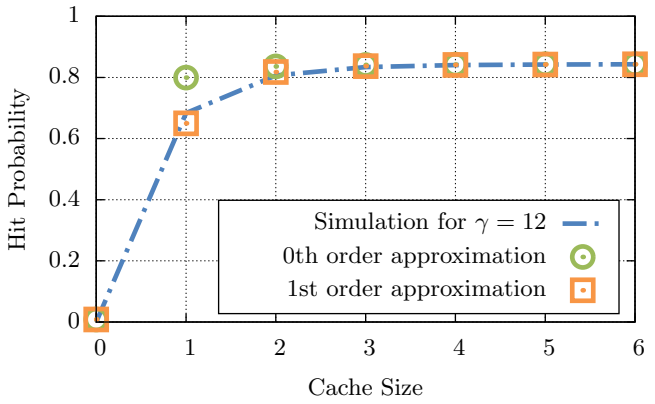
applying the commonly used "Che approximation". We have further shown that the latter expansion is valid for a sub-class of processes suitable for modeling purposes. Finally, the accuracy of our theoretical results has been illustrated by numerical experiments.

Our framework could be used to analyze other caching policies satisfying that the eviction policy for the canonical document depends only on the rest of the document request process. Examples of such caching policies found in the literature are RANDOM, which evicts a uniformly chosen document when adding a new document to the cache, and FIFO, which works as LRU except that it does not move a requested document that is already in the cache to the front of it. Such alternative policies may be relevant in that the replacement operations are somewhat simpler than LRU, and this may compensate their probable lesser performance in terms of the hit probability. However, the miss events for this policies are more intricate to analyze, since they depend on the missed requests in the rest $\widetilde{\Gamma} \setminus \delta_{0,\xi_0}$.

Another possible extension of our study would be to take into account the fact that documents have random sizes. These sizes and the cache size $C$ should be measured for instance in bits, packets, or by a continuous value in $\mathbb{R}^+$. The document sizes can be incorporated as additional marks to the cluster point process. In this case, the process $X$ defining the canonical exit time becomes a compound inhomogeneous Poisson process, summing up these file sizes. The exit time to consider for a canonical document of size $S$




(a) *Approximations for* $\gamma = 50$




(b) *Approximations for* $\gamma = 10$

Fig 5: *Comparison between hit probability curves obtained in the experiments and their analytic approximations in the case of low* $\gamma$.

is then the first passage time of $X$ strictly above $C - S$.

## 8. Proof Section.

8.1. *Proof of Proposition 1.* The Slivnyak-Mecke Theorem characterizes the Laplace functional of Poisson point processes under their Palm distributions, see [6, Prop. 13.1.VII]. Here, for $(u, \nu)$ in $\mathbb{R} \times \mathcal{M}^{\#}(\mathbb{R})$, the Laplace

functional $\mathcal{L}_{u,\nu}$ of $\tilde{\Gamma}$ under the Palm distribution $Q_{u,\nu}$ can be expressed by

$$\mathcal{L}_{u,\nu}[f] = e^{-f(u,\nu)} \cdot \mathcal{L}[f]$$

for any measurable function $f : \mathbb{R} \times \mathcal{M}^{\#}(\mathbb{R}) \to \mathbb{R}^{+}$, where $\mathcal{L}$ is the Laplace functional on the original probability space. The Laplace functional $\mathcal{L}_u$ under $\overline{Q}_u$ is consequently given by

$$\mathcal{L}_u[f] = \mathbb{E}[\mathcal{L}_{u,\xi_u}[f]] = \mathbb{E}\Big[e^{-f(u,\xi_u)}\Big]\, \mathcal{L}[f]\,.$$

Note that the expectation in the right-hand side is the Laplace functional of the point process $\delta_{u,\xi_u}$. Since Laplace functionals characterize point processes, the conclusion follows.

8.2. *Proof of Proposition 2.* Condition (1) implies that $\Xi^s(t) < \infty$ for all $t \geq s$. Among all points $(a, \xi_a)$ in the rest $\widetilde{\Gamma} \setminus \delta_{0,\xi_0}$, the process $(X^s_u)_{s \leq u \leq t}$ counts those such that $F^s(\xi_a)$ falls in $[s,t]$, and for $h \geq 0$ the increment $X^s_{t+h} - X^s_t$ counts those such that $F^s(\xi_a)$ falls in $(t, t+h]$. Since the corresponding two subsets of $\mathbb{R} \times \mathcal{M}^{\#}(\mathbb{R})$ are disjoint and $\widetilde{\Gamma} \setminus \delta_{0,\xi_0}$ is Poisson, we conclude that $X^s$ is a counting process with independent increments. In consequence, it is a inhomogeneous Poisson process.

The mean function for this process is then given by

$$\mathbb{E}[X^s_t] = \mathbb{E}\left[\sum_{a \in \Gamma^g} \mathbb{1}\{F^s(\xi_a) \in [s,t]\}\right] = \mathbb{E}\left[\sum_{a \in \Gamma^g} \mathbb{1}\{\xi_a[s,t] \geq 1\}\right].$$

Formula (5) follows from the latter expression and the fact that the mean measure $\eta$ of $\widetilde{\Gamma} \setminus \delta_{0,\xi_0}$ is defined by

$$\eta([t_1,t_2] \times B) = \gamma \int_{t_1}^{t_2} \mathbb{P}[\xi_a \in B]\, \mathrm{d}a$$

for any Borel subset $B$ of $\mathcal{M}^{\#}(\mathbb{R})$.

8.3. *Proof of Theorem 3.* In the first r.h.s. term of (4), $N$ is a mixed Poisson random variable with random mean $\Lambda$ and thus

$$\mathbb{E}[\mathbb{1}\{N \geq 1\}] = \mathbb{E}[\mathbb{E}[\mathbb{1}\{N \geq 1\} \,|\, \Lambda]] = \mathbb{E}\big[1 - e^{-\Lambda}\big] = m_0\,. \tag{17}$$

Consider now the second r.h.s. term of (4). Following [15][p.106 seq.], since the family $T^s_C$ for $s \geq 0$ is defined on the rest of the process and thus is independent from the request process $\xi = \sum_{r=1}^{N} \delta_{\Theta_r}$ for the tagged document,

$$\mathbb{E}\left[\mathbb{1}\{N > 2\} \sum_{r=2}^{N} \mathbb{1}\Big\{\Theta_r > T_C^{\Theta_{r-1}}\Big\} \,\Big|\, \xi\right] = h(\xi) \tag{18}$$

where $h : \mathcal{M}^{\#}(\mathbb{R}) \to \mathbb{R}^+$ is the measurable function defined by

$$h\left(\sum_{r=1}^{n} \delta_{t_i}\right) = \mathbb{1}\{n > 2\}\, \mathbb{E}\left[\sum_{r=2}^{n} \mathbb{1}\left\{t_r > T_C^{t_{r-1}}\right\}\right].$$

Since $T_C^s - s \overset{\mathrm{d}}{=} T_C$, see Proposition 2, the function $h$ can be rewritten as

$$h\left(\sum_{r=1}^{n} \delta_{t_i}\right) = \mathbb{1}\{n > 2\}\, \mathbb{E}\left[\sum_{r=2}^{n} \mathbb{1}\{t_r - t_{r-1} > T_C\}\right].$$

We use this to compute the expectation of the l.h.s. of eq. (18), which combined with (4) and (17) yields that

$$\mathbb{E}[\mu_C] = m_0 + \mathbb{E}\left[\mathbb{1}\{N > 2\} \sum_{r=2}^{N} \mathbb{1}\{\Theta_r - \Theta_{r-1} > T_C\}\right].$$

Now, since the canonical intensity $\lambda$ and exit time $T_C$ are independent from the request process of the tagged document, Proposition 4 yields that

$$\begin{aligned}
\mathbb{E}[\mu_C] &= m_0 + \mathbb{E}\left[\int_0^\infty \mathrm{d}w\, \mathbb{1}\{w > T_C\} \int_0^\infty \mathrm{d}u\, \lambda(u)\lambda(u+w) e^{-(\Lambda(u+w)-\Lambda(u))}\right] \\
&= m_0 + \mathbb{E}\left[\int_0^\infty \lambda(u) e^{-(\Lambda(u+T_C)-\Lambda(u))}\, \mathrm{d}u - \int_0^\infty \lambda(u) e^{-(\Lambda-\Lambda(u))}\, \mathrm{d}u\right] \\
&= \mathbb{E}\left[\int_0^\infty \lambda(u) e^{-(\Lambda(u+T_C)-\Lambda(u))}\, \mathrm{d}u\right],
\end{aligned}$$

where we use for the last equality that, since $\Lambda(\infty) = \Lambda$ and $\Lambda(0) = 0$,

$$\int_0^\infty \lambda(u) e^{-(\Lambda-\Lambda(u))}\, \mathrm{d}u = \left[e^{-(\Lambda-\Lambda(u))}\right]_0^\infty = 1 - e^{-\Lambda}.$$

This last equation and dominated convergence imply that $\lim_{t\to\infty} \downarrow m(t) = \mathbb{E}\left[1 - e^{-\Lambda}\right]$, which concludes the proof.

8.4. *Proof of Proposition 4.* Recall that, given that the process $\xi$ has $k$ points, the request times $(\Theta_r)_{r=1}^k$ have the distribution of the order statistics of a random variable with density $g(t) = \lambda(t)/\Lambda$ for $t \geq 0$, and thus with c.d.f. $G$ given by $G(t) = \Lambda(t)/\Lambda$ for $t \geq 0$. Let $\overline{G} = 1 - G$ denote the complement of $G$. From order statistics theory, it is known that the holding times $\Theta_r - \Theta_{r-1}$ for $2 \leq r \leq k$ have density $\tilde{g}_{k,r}$ given for $w \geq 0$ by

$$\tilde{g}_{k,r}(w) = \frac{k!}{(r-2)!(k-r)!} \int_0^\infty G^{r-2}(u) g(u) g(u+w) \overline{G}^{k-r}(u+w)\, \mathrm{d}u.$$

Consequently, for $k \geq 2$ we have

$$\mathbb{E}[F(\Theta_r - \Theta_{r-1}) \,|\, N = k] = \int_0^\infty F(w)\tilde{g}_{k,r}(w)\,\mathrm{d}w$$

and hence

$$\mathbb{E}\left[\mathbb{1}\{N \geq 2\}\sum_{r=2}^{N} F(\Theta_r - \Theta_{r-1})\right] = \sum_{k=2}^{\infty}\sum_{r=2}^{k}\mathbb{E}[F(\Theta_r - \Theta_{r-1}) \,|\, N = k]\, e^{-\Lambda}\frac{\Lambda^k}{k!}$$

$$= \int_0^\infty F(w)\, e^{-\Lambda}\sum_{k=2}^{\infty}\sum_{r=2}^{k}\tilde{g}_{k,r}(w)\,\frac{\Lambda^k}{k!}\,\mathrm{d}w\,. \tag{19}$$

Now, using the Binomial Theorem,

$$\sum_{r=2}^{k}\frac{k!}{(r-2)!(k-r)!}G^{r-2}(u)\overline{G}^{k-r}(u+w) = k(k-1)[G(u)+\overline{G}(u+w)]^{k-2}$$

and thus

$$\sum_{r=2}^{k}\tilde{g}_{k,w}(w)\frac{\Lambda^k}{k!} = \int_0^\infty [G(u)+\overline{G}(u+w)]^{k-2}\frac{\Lambda^k}{(k-2)!}g(u)g(u+w)\,\mathrm{d}u\,,$$

and we conclude that

$$e^{-\Lambda}\sum_{k=2}^{\infty}\sum_{r=2}^{k}\tilde{g}_{k,w}(w)\frac{\Lambda^k}{k!} = \Lambda^2\int_0^\infty e^{-\Lambda(1-G(u)-\overline{G}(u+w))}g(u)g(u+w)\,\mathrm{d}u\,.$$

Since $\Lambda\times(1-G(u)-\overline{G}(u+w)) = \Lambda\times(G(u+w)-G(u)) = \Lambda(u+w)-\Lambda(u)$ and $g(u)g(u+w) = \lambda(u)\lambda(u+w)/\Lambda^2$, Equation (19) together with the latter intermediate results concludes the proof.

8.5. *Proof of Proposition 5.* Since the processes $\lambda_a(\cdot)$ and $\lambda_0(\cdot - a)$ have the same distribution, and $\Lambda_a(0) = 0$ for $a \geq 0$, we may write $\Xi(t)$ as

$$\Xi(t) = \gamma\int_{-\infty}^{0}\mathbb{E}\Big[1 - e^{-(\Lambda(t-a)-\Lambda(-a))}\Big]\,\mathrm{d}a + \gamma\int_0^t\mathbb{E}\Big[1-e^{-\Lambda(t-a)}\Big]\,\mathrm{d}a\,.$$

We denote the first integral by $I_1(t)$ and the second by $I_2(t)$. The change of variables $a \mapsto -a$ yields

$$I_1(t) = \int_{-\infty}^{0}\mathbb{E}\Big[1 - e^{-(\Lambda(t-a)-\Lambda(-a))}\Big]\,\mathrm{d}a = \int_0^\infty\mathbb{E}\Big[1 - e^{-(\Lambda(t+a)-\Lambda(a))}\Big]\,\mathrm{d}a$$

and thus, using $\frac{\mathrm{d}}{\mathrm{d}a}e^{-(\Lambda(t+a)-\Lambda(a))} = -(\lambda(t+a)-\lambda(a))e^{-(\Lambda(t+a)-\Lambda(a))}$,

$$\begin{aligned} I_1'(t) &= \int_0^\infty \mathbb{E}\Big[\lambda(t+a)e^{-(\Lambda(t+a)-\Lambda(a))}\Big]\,\mathrm{d}a \\ &= \mathbb{E}\Big[e^{-\Lambda(t)}-1\Big] + \mathbb{E}\Big[\int_0^\infty \lambda(a)e^{-(\Lambda(t+a)-\Lambda(a))}\,\mathrm{d}a\Big] \\ &= \mathbb{E}\Big[e^{-\Lambda(t)}-1\Big] + m(t)\,. \end{aligned}$$

Now, the change of variables $a \mapsto t-a$ yields

$$I_2(t) = \int_0^t \mathbb{E}\Big[1-e^{-\Lambda(t-a)}\Big]\,\mathrm{d}a = \int_0^t \mathbb{E}\Big[1-e^{-\Lambda(a)}\Big]\,\mathrm{d}a\,,$$

and hence $I_2'(t) = \mathbb{E}\big[1-e^{-\Lambda(t)}\big]$. Thus $\Xi'(t) = \gamma(I_1'(t)+I_2'(t)) = \gamma\, m(t)$ as claimed. We conclude by integrating this, since $\Xi(t)=0$.

8.6. *Proof of Lemma 6.*

(i). This is the optimized exponential Markov inequality which is used for the upper bound in Cramer's large deviations Theorem, see [7, Theorem 2.2.3, Remark (c)].

(ii). Expanding the $k$-th order central moment of $X_C$ in terms of the known moments of $\widehat{T}_C$ yields that

$$\begin{aligned} \mathbb{E}\Big[(X_C-1)^k\Big] &= \sum_{i=0}^k \binom{k}{i}\frac{\mathbb{E}\Big[(\widehat{T}_C)^i\Big]}{C^i}(-1)^{k-i} \\ &= \frac{1}{C^k}\sum_{i=0}^k \binom{k}{i}(-C)^{k-i}\frac{\Gamma(C+i)}{\Gamma(C)} \\ &= \frac{1}{C^k}\phi_k(C)\,, \end{aligned}$$

where $\phi_k$ is a polynomial of degree at most $k$. As shown in [17], the polynomial $\phi_k$ is actually of degree $\lfloor k/2 \rfloor$, which allows us to conclude.

8.7. *Proof of Theorem 7.* Define the function $f_\theta$ by

$$f_\theta(z) = m(M^{-1}(\theta z)) = m(t_{\theta z}).$$

With the scaling $C = \gamma\theta$, Equation (10) can be then written as

$$\mathbb{E}[\mu_C] = \mathbb{E}\left[f_\theta\left(\frac{\widehat{T}_C}{C}\right)\right]. \tag{20}$$

Let again $X_C = \widehat{T}_C/C$ as in Lemma 6, and fix $\eta > 0$. Let us decompose the expectation (20) into $\mathbb{E}[\mu_C] = A_C + B_C$ where

$$A_C = \mathbb{E}\big[f_\theta(X_C)\mathbb{1}_{\{|X_C-1|\geq\eta\}}\big]\,, \quad B_C = \mathbb{E}\big[f_\theta(X_C)\mathbb{1}_{\{|X_C-1|<\eta\}}\big]\,.$$

For $A_C$, recall that the function $m$ is bounded by $\mathbb{E}[\Lambda] < \infty$, and so is $f_\theta$. Then, by Lemma 6 (i), we have

$$A_C \leq \mathbb{E}[\Lambda]\,\mathbb{P}[|X_C - 1| \geq \eta] \leq 2\mathbb{E}[\Lambda]\,e^{-C\cdot\varphi(1+\eta)} = o(1/C)\,.$$

For $B_C$, we write a Taylor expansion of $f_\theta$ at 1 of order two in the form

$$\begin{aligned} f_\theta(X_C) &= f_\theta(1) + f'_\theta(1)\,(X_C - 1) + \frac{f''_\theta(Y_C)}{2}\,(X_C-1)^2 \\ &= h_\theta(X_C) + k_\theta(X_C, Y_C)\,, \end{aligned}$$

where $Y_C$ is a random variable in the random interval $[1, X_C] \cup [X_C, 1]$, and

$$\begin{cases} h_\theta(X_C) = f_\theta(1) + f'_\theta(1)\,(X_C - 1) + \dfrac{f''_\theta(1)}{2}\,(X_C-1)^2\,, \\ k_\theta(X_C, Y_C) = \dfrac{f''_\theta(Y_C) - f''_\theta(1)}{2}\,(X_C-1)^2\,. \end{cases}$$

Then $B_C = D_C + E_C$ where

$$D_C = \mathbb{E}\big[h_\theta(X_C)\mathbb{1}_{\{|X_C-1|<\eta\}}\big]\,, \quad E_C = \mathbb{E}\big[k_\theta(X_C, Y_C)\mathbb{1}_{\{|X_C-1|<\eta\}}\big]\,.$$

We then compute

$$D_C = \mathbb{E}[h_\theta(X_C)] - \mathbb{E}\big[h_\theta(X_C)\mathbb{1}_{\{|X_C-1|\geq\eta\}}\big] \tag{21}$$

where

$$\mathbb{E}[h_\theta(X_C)] = f_\theta(1) + \frac{f''_\theta(1)}{2C}$$

since $\mathbb{E}[X_C - 1] = 0$ and $\mathbb{E}\big[(X_C-1)^2\big] = 1/C$. Besides, to deal with the second term $\mathbb{E}\big[h_\theta(X_C)\mathbb{1}_{\{|X_C-1|\geq\eta\}}\big]$ in the right-hand side of (21), we use the Cauchy-Schwarz inequality to write

$$\big|\mathbb{E}\big[h_\theta(X_C)\mathbb{1}_{\{|X_C-1|\geq\eta\}}\big]\big| \leq \sqrt{\mathbb{E}[h_\theta(X_C)^2]}\sqrt{\mathbb{P}[|X_C-1|\geq\eta]}$$

and note that $\mathbb{E}\big[h_\theta(X_C)^2\big] = O(1)$ for all $C > 1$ by Lemma 6 (ii). Applying Lemma 6 (i) then eventually shows that $\mathbb{E}\big[h_\theta(X_C)\mathbb{1}_{\{|X_C-1|\geq\eta\}}\big]$ is $O(e^{-\frac{C}{2}\cdot\varphi(1+\eta)})$ which is, in particular, $o(1/C)$. At this stage, we therefore conclude from (21) and the latter discussion that

$$D_C = f_\theta(1) + \frac{f''_\theta(1)}{2C} + o\left(\frac{1}{C}\right). \tag{22}$$

Lastly, we show that the term $E_C$ is $o(1/C)$. To this aim, it is sufficient to show that the sequence $W_C = C \cdot k_\theta(X_C, Y_C)$ for $C > 1$ converges in probability to zero and that it is uniformly integrable ([22, Theorem 13.7]).

- To prove the convergence in probability, note that since $X_C \to 1$ a.s. when $C \to \infty$ and $Y_C \in [1, X_C] \cup [X_C, 1]$, then $Y_C \to 1$ a.s. It follows from the continuity of $f''_\theta$ in the interval $(1-\eta, 1+\eta)$ that $f''_\theta(1) - f''_\theta(Y_C) \to 0$ a.s. and, in particular, in probability. On the other hand, since $X_C = \widehat{T}_C/C$ is an average of $C$ i.i.d. random variables with mean 1, the continuous mapping theorem for weak limits implies that $C(X_C - 1)^2$ converges in distribution (the limit distribution is $\chi^2$ with parameter 1 but this specific limit has no importance for the present proof). Finally, since $\mathbb{1}\{|X_C - 1| < \eta\} \to 1$ a.s., Slutsky's theorem ([14, Th. 11.4]) allows us to conclude that

$$W_C = \frac{f''_\theta(1) - f''_\theta(Y_C)}{2} \times C(X_C - 1)^2 \times \mathbb{1}\{|X_C - 1| < \eta\} \to 0$$

in distribution as $C \to \infty$, and thus in probability as well.

- To prove the uniform integrability of $W_C$, it suffices to show that

$$\sup_{C\geq 1} \mathbb{E}\big[W_C^2\big] < \infty \tag{23}$$

(see [22, Theorem 13.3]). Since $f_\theta$ is twice continuously differentiable,

$$\left|\frac{f''_\theta(1) - f''_\theta(Y_C)}{2}\mathbb{1}_{\{|X_C-1|<\eta\}}\right| \leq K$$

for any $C > 1$ and for some constant $K$ depending on $\eta$ only. By Lemma 6, we further have $\mathbb{E}\Big[C^2\,(X_C - 1)^4\Big] = C^2 \times O(C^{-2}) = O(1)$. We finally conclude that $\mathbb{E}\big[W_C^2\big] < K^2 \times O(1) < \infty$, which proves the claimed property (23).

Finally gathering $\mathbb{E}[\mu_C] = A_C + B_C = A_C + D_C + E_C$ with $A_C = o(1/C)$, $E_C = o(1/C)$ and $D_C$ expanded in (22), we thus have proved that

$$\mathbb{E}[\mu_C] = f_\theta(1) + \frac{f''_\theta(1)}{2C} + o\left(\frac{1}{C}\right) \tag{24}$$

as $C \to \infty$. To conclude the proof, we now express the function $f_\theta$ and its derivatives at 1 in terms of function $m$ and its derivatives at $t_\theta$. By implicit differentiation,

$$f'_\theta(z) = \frac{m'(t_{\theta z})}{m(t_{\theta z})}\theta, \quad f''_\theta(z) = \frac{\theta^2}{m(t_{\theta z})^2}\left(m''(t_{\theta z}) - \frac{m'(t_{\theta z})^2}{m(t_{\theta z})}\right),$$

and the values of $f'_\theta$ and $f''_\theta$ at $z = 1$ consequently follow. Replacing them into (24), we finally prove the expansion (12), as claimed.

8.8. *Proof of Proposition 10.* Differentiating (7) under the integral sign, with $\lambda(u)$ expressed by (13), readily gives formulas (14) after using the change of variables $u \mapsto u/L$. The validity of these formulas can then be simply proved by showing that these integrals for $m'$ and $m''$ are finite.

Given $t > 0$ and $L$, define $u^* = u^*(t, L) = \inf\{u : f(u) > f(u + t/L)\}$, so that $f(u) \le f(u + t/L)$ for $u \le u^*$ and $f(u) > f(u + t/L)$ for $u > u^*$. The existence of $u^*$ is ensured from the unimodality of $f$, and we have $u^* = 0$ if and only if $f$ is non-increasing. Finally, define $\tilde{u} = \inf\{u : f(u) = \max f\}$ (see Fig. 6 for a schematic view of these definitions).

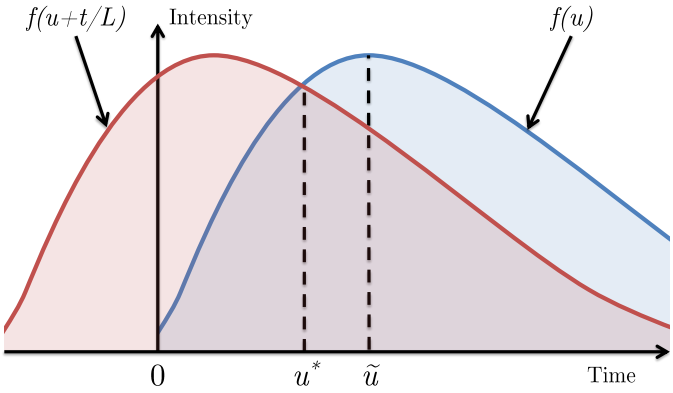



Fig 6: *Schema for unimodal f*

Since $f$ is differentiable and unimodal, it is quasi-concave (see [8, Lemma 2.4.1]), that is, for any $0 \le \eta \le 1$ we have $f(\eta u_1 + (1-\eta)u_2) \ge f(u_1) \wedge f(u_2)$ for $u_1, u_2 \ge 0$. As a consequence, for any $t > 0$, the area under the graph of $f$ in the interval $[u, u + t/L]$ can be bounded below by

$$F(u + t/L) - F(u) \ge \begin{cases} f(u) \cdot t/L, & u \le u^*, \\ f(u + t/L) \cdot t/L, & u > u^*. \end{cases} \tag{25}$$

We now partition the integrals in (14) into their contributions from intervals $[0, u^*]$ and $[u^*, \infty)$, respectively, and bound them separately. For the first derivative $m(t)$, the lower bounds (25) yield

$$|m'(t)| \le \mathbb{E}\left[RL \int_0^{u^*} f(u + t/L) Rf(u) e^{-Rf(u)t}\, \mathrm{d}u\right] + \mathbb{E}\left[RL \int_{u^*}^{\infty} f(u) Rf(u + t/L) e^{-Rf(u+t/L)t}\, \mathrm{d}u\right] \le \frac{2}{et}\mathbb{E}[RL]$$

where the last inequality is justified by the bound $xe^{-ax} \le 1/ae$ for any fixed $a > 0$, and the fact that $\int f = 1$.

For the second derivative $m''(t)$, we introduce the integrals

$$A_1(t) = \mathbb{E}\left[RL \int_0^{\infty} R^2 f(u) f(u + t/L)^2 e^{-RL(F(u+t/L)-F(u))\,\mathrm{d}u}\right],$$
$$A_2(t) = \mathbb{E}\left[R \int_0^{\infty} Rf(u) f'(u + t/L) e^{-RL(F(u+t/L)-F(u))\,\mathrm{d}u}\right]$$

so that $|m''(t)| \le |A_1(t)| + |A_2(t)|$. For $A_1(t)$, we have

$$|A_1(t)| \le \mathbb{E}\left[RL \int_0^{u^*} f(u + t/L)^2 f(u) R^2 e^{-Rf(u)t}\, \mathrm{d}u\right] + \mathbb{E}\left[RL \int_{u^*}^{\infty} f(u) R^2 f(u + t/L)^2 e^{-Rf(u+t/L)t}\, \mathrm{d}u\right] \le \mathbb{E}\left[\frac{4RL}{e^2 t^2 f(0)} \int f^2\right] + \mathbb{E}\left[\frac{RL}{et}\right] \le \frac{1}{et}\left(1 + \frac{4}{f(0)et}\int f^2\right)\mathbb{E}[RL] < \infty$$

where the last inequality follows from the bounds $xe^{-ax} \le 1/ae$, $x^2 e^{-ax} \le 4/a^2e^2$ for any fixed $a > 0$, and the fact that $0 < f(0) \le f(u) \le f(u + t/L)$ for $u \in [0, u^*]$. Regarding $A_2(t)$, we have

$$|A_2(t)| \le \mathbb{E}\left[R \int_0^{u^*} Rf(u) |f'(u + t/L)| e^{-Rf(u)t}\, \mathrm{d}u\right] + \mathbb{E}\left[R \int_{u^*}^{\infty} Rf(u) |f'(u + t/L)| e^{-Rf(u+t/L)t}\, \mathrm{d}u\right] = B_1(t) + B_2(t).$$

Using again $xe^{-ax} \le 1/ae$, we have

$$B_1(t) \le \frac{\mathbb{E}[R]}{et} \int |f'| < \infty.$$

Finally, to deal with $B_2(t)$ we note that $f'(u+t/L) \leq 0$ for $u \in [u^*, \infty)$ and thus $|f'(u+t/L)| = -f'(u+t/L)$. We then use integration by parts to obtain

$$\begin{aligned} B_2(t) &= -\frac{1}{t}\mathbb{E}\left[R\left(\left[-e^{-Rf(u+t/L)t}f(u)\right]_{u=u^*}^{\infty} + \int_{u^*}^{\infty} f'(u)e^{-Rf(u+t/L)t}\,\mathrm{d}u\right)\right] \\ &= -\frac{1}{t}\mathbb{E}\left[Rf(u^*)e^{-Rf(u^*+t/L)}\right] \\ &\quad -\frac{1}{t}\mathbb{E}\left[R\int_{u^*}^{\tilde{u}} f'(u)e^{-Rf(u+t/L)t}\,\mathrm{d}u\right] - \frac{1}{t}\mathbb{E}\left[R\int_{\tilde{u}}^{\infty} f'(u)e^{-Rf(u+t/L)t}\,\mathrm{d}u\right]. \end{aligned}$$

The first term in the latter expression is trivially negative; the second is also negative since $f$ is non-decreasing in $[0, \tilde{u})$. As a consequence both terms can be ignored to obtain

$$B_2(t) \leq \frac{1}{t}\mathbb{E}\left[R\int_{\tilde{u}}^{\infty} |f'(u)|e^{-Rf(u+t/L)t}\,\mathrm{d}u\right] \leq \frac{\mathbb{E}[R]}{t}\int |f'| < \infty$$

thus concluding the proof.

**Acknowledgements.** The authors wish to thank Bruno Kauffmann at Orange Labs for his deep insight and fruitful discussions along with Byron Schmuland for pointing out reference [17] in this question posed at the *Mathematics StackExchange* website.

## References.

[1] Ahlgren, B., Dannewitz, C., Imbrenda, C., Kutscher, D., and Ohlman, B. (2012). A Survey of Information-Centric Networking. *Communications Magazine, IEEE* **50**, 7, 26–36.

[2] Baccelli, F. and Brémaud, P. (2013). *Elements of queueing theory: Palm Martingale calculus and stochastic recurrences*. Vol. **26**. Springer Science & Business Media.

[3] Che, H., Tung, Y., and Wang, Z. (2002). Hierarchical Web Caching Systems: Modeling, Design and Experimental Results. *Selected Areas in Communications, IEEE Journal on* **20**, 7, 1305–1314.

[4] Cisco Systems, Inc. (2015). Cisco Visual Networking Index: Forecast and Methodology, 2014–2019.

[5] Daley, D. J. and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes*, 2nd ed. Vol. **1**. Springer.

[6] Daley, D. J. and Vere-Jones, D. (2008). *An Introduction to the Theory of Point Processes*, 2nd ed. Vol. **2**. Springer.

[7] Dembo, A. and Zeitouni, O. (2009). *Large Deviations Techniques and Applications*. Vol. **38**. Springer Science & Business Media.

[8] dos Santos Gromicho, J. A. (2013). *Quasiconvex optimization and location theory*. Vol. **9**. Springer Science & Business Media.

[9] Fill, J. A. and Holst, L. (1996). On the Distribution of Search Cost for the Move-to-Front Rule. *Random Structures & Algorithms* **8**, 3, 179–186.

[10] Fofack, N. C., Nain, P., Neglia, G., and Towsley, D. (2012). Analysis of TTL-based Cache Networks. In *6th International Conference on Performance Evaluation Methodologies and Tools (VALUETOOLS)*. IEEE, 1–10.
[11] Fofack, N. C., Towsley, D., Badov, M., Dehghan, M., and Goeckel, D. L. (2014). An approximate analysis of heterogeneous and general cache networks. Tech. rep., Inria.
[12] Folland, G. B. (1999). *Real Analysis: Modern Techniques and their Applications*, 2nd ed. John Wiley & Sons.
[13] Fricker, C., Robert, P., and Roberts, J. (2012). A Versatile and Accurate Approximation for LRU Cache Performance. In *Proceedings of the 24th International Teletraffic Congress*. International Teletraffic Congress, 8.
[14] Gut, A. (2006). *Probability: A Graduate Course*. Springer Science & Business Media.
[15] Kallenberg, O. (2006). *Foundations of Modern Probability*. Springer Science & Business Media.
[16] Leonardi, E. and Torrisi, G. L. (2015). Least Recently Used caches under the Shot Noise Model. In *IEEE INFOCOM 2015*.
[17] Mathematical Association of America. (2011). Problems and Solutions. *The American Mathematical Monthly* **118**, 3, pp. 275–282.
[18] Miller, P. D. (2006). *Applied Asymptotic Analysis*. Vol. **75**. American Mathematical Soc.
[19] Olmos, F., Kauffmann, B., Simonian, A., and Carlinet, Y. (2014). Catalog Dynamics: Impact of Content Publishing and Perishing on the Performance of a LRU cache. In *26th International Teletraffic Congress (ITC)*. IEEE, 1–9.
[20] Traverso, S., Ahmed, M., Garetto, M., Giaccone, P., Leonardi, E., and Niccolini, S. (2013). Temporal locality in today's content caching: why it matters and how to model it. *ACM SIGCOMM Computer Communication Review* **43**, 5, 5–12.
[21] Whitt, W. (2002). *Stochastic-process limits: an introduction to stochastic-process limits and their application to queues*. Springer Science & Business Media.
[22] Williams, D. (1991). *Probability with Martingales*. Cambridge University Press.

Orange Labs
Department OLN / NMP / TRM
38 - 40 rue du Général Leclerc
92794 Issy-Les-Moulineaux, France
E-mail: luisfelipe.olmosmarchant@orange.com
E-mail: alain.simonian@orange.com

Centre de Mathématiques Appliquées
École Polytechnique, CNRS
Université Paris Saclay
Route de Saclay
91128 Palaiseau, France
E-mail: carl.graham@polytechnique.edu