Modeling LEAST RECENTLY USED caches with Shot Noise request processes

SIAM J. APPL. MATH.
Vol. 77, No. 2, pp. 361–383

MODELING LEAST RECENTLY USED CACHES WITH SHOT NOISE REQUEST PROCESSES[*]

EMILIO LEONARDI[†] AND GIOVANNI LUCA TORRISI[‡]

Abstract. In this paper we analyze least recently used (LRU) caches operating under the shot noise requests model (SNM). The SNM was recently proposed in [S. Traverso et al., *ACM Comput. Comm. Rev.*, 43 (2013), pp. 5–12] to better capture the main characteristics of today's video on demand traffic. We investigate the validity of Che's approximation [H. Che, Y. Tung, and Z. Wang, *IEEE J. Selected Areas Commun.*, 20 (2002), pp. 1305–1314] through an asymptotic analysis of the cache eviction time. In particular, we provide a law of large numbers, a large deviation principle, and a central limit theorem for the cache eviction time, as the cache size grows large. Finally, we derive upper and lower bounds for the "hit" probability in tandem networks of caches under Che's approximation.

Key words. caching systems, performance evaluation, asymptotic analysis

AMS subject classifications. Primary, 68M20; Secondary, 60F10, 60F05

DOI. 10.1137/15M1041444

1. Introduction.

The design and analysis of caching systems, a very traditional and widely studied topic in computer science, has recently drawn the attention of the networking research community. This interest revival is mainly due to the important role that caches play today in the distribution of contents over the Internet. Massive content delivery networks, indeed, represent the standard solution adopted by content and network providers to reach large populations of geographically distributed users in an effective way. Such delivery networks allow providers to cache contents close to the users, achieving the twofold goal of reducing network traffic while minimizing the latency suffered by users.

Unfortunately, performance evaluation of caching systems is very hard, as the computational cost to analyze the behavior of a cache is exponential in both the cache size and the number of contents [9]. For this reason, the effort of the research community has mainly focused on the development of accurate and computationally efficient approximate techniques for the analysis of caching systems, under various traffic conditions. Che's approximation [4], proposed for the analysis of least recently used (LRU) caches under the independent reference model (IRM), has emerged as one of the most powerful methods to obtain accurate estimates of the "hit" probability at limited computational costs [1, 15, 24]. The main idea of this technique is to summarize the response of a cache to the requests arriving for any possible content by a single primitive quantity, which is assumed to be deterministic and the same for any content. This approximation simplifies the analysis of caching systems because it allows us to decouple the dynamics of different contents. In [15] Che's approximation for LRU caches under the IRM found a theoretical justification.

[*] Received by the editors September 28, 2015; accepted for publication (in revised form) December 13, 2016; published electronically March 2, 2017. A preliminary version of this work was presented at the IEEE International Conference on Computer Communication (Infocom), 2015.
http://www.siam.org/journals/siap/77-2/M104144.html

[†] Dipartimento di Elettronica e Telecomunicazioni, Politecnico di Torino, Torino 10129, Italy (leonardi@polito.it).

[‡] Istituto per le Applicazioni del Calcolo "Mauro Picone," CNR, Rome I-00133, Italy (torrisi@iac.rm.cnr.it).

Shot noise processes constitute a versatile and mathematical tractable class of stochastic models which has found several applications in electrical engineering and queueing theory [2, 3, 13, 16, 17, 18, 23, 25, 31, 32]. In this paper we extend the mathematical analysis of Che's approximation to LRU caches operating under the shot noise model (SNM) [22, 33]. This model provides a simple, flexible, and accurate description of the temporal locality found, e.g., in video on demand (VoD) traffic, capturing today's traffic characteristics in a more natural and precise way than traditional traffic models. Inspired by the seminal paper [15], we investigate the validity of Che's approximation by means of an asymptotic analysis of the cache eviction time. Specifically, we provide a law of large numbers, large deviations, and a central limit theorem for the cache eviction time, as the cache size grows large. Furthermore, to the best of our knowledge, we give for the first time a nonasymptotic analytical upper bound for the error estimate of the "hit" probability entailed by Che's approximation. We provide also upper and lower bounds for the "hit" probability, under Che's approximation, for a tandem network of caches. Finally, we present some numerical illustrations. Our results show that Che's approximation is a provable, highly accurate, and scalable tool to assess the performance of LRU caching systems under the SNM.

2. System description and motivations. We consider a cache, whose size (or capacity), expressed in number of objects (or contents), is denoted by C. The cache is fed by an exogenous arrival process of objects' requests generated by users. Requests which find the object in the cache are said to produce a "hit," whereas requests that do not find the object in the cache are said to produce a "miss." An important performance index is the "hit" probability, which is the fraction of the requests producing a "hit." The miss stream of a cache, i.e., the process of requests which are not locally satisfied by the cache, is forwarded to either other caches or a common repository containing all the objects, i.e., the entire objects' catalogue. In the literature it is common to neglect all propagation delays.

In this paper we focus on caches implementing the LRU policy: upon the arrival of a request, an object not already stored in the cache is inserted into it. If the cache is full, to make room for a new object the LRU item is evicted, i.e., the object which has not been requested for the longest time is expunged from the cache.

Several models have been proposed to describe the process of requests arriving at a cache. The simplest and still the most widely adopted is certainly the IRM [5], which makes the following two fundamental assumptions: (i) the catalogue consists of a fixed number of objects, which does not change over time; (ii) the process of requests of a given object is modeled by a homogeneous Poisson process. As a consequence, the IRM completely ignores all temporal correlations in the sequence of requests and does not take into account a key feature of real traffic referred to as temporal locality, which means that if an object is requested at a given time, then it is more likely that the same object will be requested again in the near future. It is well-known that the temporal locality has a beneficial effect on the cache performance, as it increases the "hit" probability [5].

Several extensions of the IRM have been proposed to incorporate the temporal locality into the traffic model. Existing generalizations [1, 5, 21] typically assume that the process of requests is time-stationary, usually either a renewal process or a Markov modulated Poisson process. However, these models do not capture the kind of temporal locality encountered in traces related to VoD traffic, which is instead well described by the SNM as shown in [22, 33].

3. Shot noise model and cache analysis. The basic idea of the SNM is to represent the requests' process as the superposition of many independent processes, each one referring to a specific object. The requests' process of a fixed content m is specified by two physical (random) parameters: ξ_m and Z_m. ξ_m represents the time instant at which the content enters the system (i.e., it becomes available to the users); mark Z_m describes some attribute of the content m, which summarizes its main characteristics (content type, volume, etc.).

We assume that the set of times $N \equiv \{\xi_m\}_{m\geq 1}$ at which contents become available to users (i.e., they are introduced in the common repository) is distributed according to a homogeneous Poisson process on $\mathbb{R}$ with intensity $\lambda > 0$. Here, $\{\xi_m\}_{m\geq 1}$ is supposed to be an unordered set of times. We suppose that, after the introduction into the catalogue of the content m, the requests for this content arrive at the cache according to a Cox process $N^{(m)} \equiv \{T_n^{(m)}\}_{n\geq 1}$ on $\mathbb{R}$ whose stochastic intensity $\{\lambda_m(t)\}_{t\in\mathbb{R}}$ is defined by

$$\lambda_m(t) := h(t - \xi_m, Z_m);$$

see, e.g., [7]. We assume that $\{Z_m\}_{m\geq 1}$ is a sequence of independent and identically distributed random variables, independent of $\{\xi_m\}_{m\geq 1}$, with values on some measurable space $(E, \mathcal{E})$. Furthermore, we suppose that $h : \mathbb{R} \times E \to [0, \infty)$ is a measurable nonnegative function such that $h(t, z) = 0$ for any $t < 0$ and $z \in E$. Finally, we suppose that, for any $m \geq 1$, $T_1^{(m)} < T_2^{(m)} < \dots$ almost surely and we assume that the Cox processes $\{N^{(m)}\}_{m\geq 1}$ are independent, given $\{(\xi_m, Z_m)\}_{m\geq 1}$.

3.1. Formal definition of the cache eviction time. We denote by m_0 a tagged content introduced into the catalogue at the deterministic time $\xi_{m_0} = -x$, $x > 0$, and requested at time 0. Moreover, we denote by $\mathbf{X}_{m_0}(t)$, $t > 0$, the number of contents different from m_0 that have been requested in the time interval $[0, t]$, i.e.,

$$\mathbf{X}_{m_0}(t) = \sum_{m\neq m_0} \mathbb{1}\left\{m \text{ requested in } [0,t],\ \xi_m \in (-\infty, t]\right\}.$$

Throughout this paper we shall consider the random variable

$$X_{m_0,x}(t) := \mathbf{X}_{m_0}(t) \,|\, \xi_{m_0} = -x, \quad t, x > 0,$$

which plays an important role in the dynamics of an LRU cache because the cache eviction time may be expressed in terms of $X_{m_0,x}(t)$. Indeed, under the LRU replacement policy, we have that the content m_0 is expunged from the cache (provided it is not requested again after time 0) as soon as the Cth content, different from m_0, is requested. So, under the LRU replacement policy, the so-called cache eviction time for the content m_0 is given by the random variable

$$T_C(m_0, x) := \inf\{t > 0 :\ X_{m_0,x}(t) = C\},$$

where once again we remark that, by construction, $T_C(m_0, x)$ is the time at which the content m_0 is expunged from the cache, provided that no requests for the content m_0 are observed in the time interval $(0, T_C(m_0, x)]$.

3.2. The distribution of $X_{m_0,x}(t)$. Define the quantity

$$g(t) := \int_0^\infty \mathbb{E}\left[1 - \mathrm{e}^{-\int_{u-t}^{u} h(s,Z_1)\,\mathrm{d}s}\right] \mathrm{d}u, \quad t > 0. \tag{1}$$

The following proposition holds.

PROPOSITION 3.1. *If $g(t) < \infty$, then the random variable $X_{m_0,x}(t)$ is Poisson distributed with mean $\lambda g(t)$.*

Note that the condition $g(t) < \infty$ is fairly general: for example, it is satisfied whenever the popularity profile is of multiplicative form, i.e. (with a little abuse of notation),

$$h(t,z) := zh(t), \quad t \in \mathbb{R},\, z \in E \subseteq (a,\infty),\, a > 0, \tag{2}$$

and

$$h \equiv 0 \text{ on } (-\infty,0),\ \int_0^\infty h(t)\,\mathrm{d}t = 1 \text{ and } \mathbb{E}[Z_1] < \infty. \tag{3}$$

Indeed, in such a case, for $t > 0$, we have

$$\begin{aligned} g(t) &= \int_0^\infty \left[1 - \phi_{Z_1}\left(-\int_{u-t}^u h(s)\,\mathrm{d}s\right)\right]\mathrm{d}u \\ &\le \mathbb{E}[Z_1]\int_0^\infty \left(\int_{u-t}^u h(s)\,\mathrm{d}s\right)\mathrm{d}u \le \mathbb{E}[Z_1]t < \infty, \end{aligned} \tag{4}$$

where $\phi_{Z_1}(\theta) := \mathbb{E}[\exp(\theta Z_1)]$, $\theta \in \mathbb{R}$, and we used the elementary inequality $\mathrm{e}^x \ge 1+x$, $x \in \mathbb{R}$.

Proof of Proposition 3.1. For any $t_0 > t > 0$, we define the "restriction" of $\mathbf{X}_{m_0}(t)$ to contents that have been introduced in the model in the time interval $[t-t_0,t]$ by

$$\mathbf{X}_{m_0}^{(t_0)}(t) := \sum_{m \ne m_0} \mathbb{1}\{m \text{ requested in } [0,t],\, \xi_m \in [t-t_0,t]\}.$$

By the Slivnyak–Mecke theorem (see, e.g., Proposition 13.1.VII, p. 281, in [8]), the law of $\{\xi_m\}_{m\ne m_0}$ given the event $\{\xi_{m_0} = -x\}$ coincides with the law of $\{\xi_m\}_{m\ge1}$ and so, for any $\theta \in \mathbb{R}$,

$$\mathbb{E}\left[\mathrm{e}^{\theta \mathbf{X}_{m_0}^{(t_0)}(t)} \mid \xi_{m_0} = -x\right] = \mathbb{E}\left[\mathrm{e}^{\theta \widetilde{\mathbf{X}}_{t_0}(t)}\right], \tag{5}$$

where

$$\widetilde{\mathbf{X}}_{t_0}(t) := \sum_{m\ge1} \mathbb{1}\{m \text{ requested in } [0,t],\, \xi_m \in [t-t_0,t]\}.$$

Letting $N([t-t_0,t])$ denote the number of points $\{\xi_m\}_{m\ge1}$ in the time interval $[t-t_0,t]$ and $N^{(m)}([0,t])$ denote the number of points $\{T_n^{(m)}\}_{n\ge1}$ in the time interval $[0,t]$, we rewrite $\widetilde{\mathbf{X}}_{t_0}(t)$ as

$$\widetilde{\mathbf{X}}_{t_0}(t) = \sum_{m=1}^{N([t-t_0,t])} \mathbb{1}\left\{N^{(m)}([0,t]) \ge 1\right\}.$$

Since, given ξ_m and Z_m, $N^{(m)}$ is a Poisson process with intensity function $h(\cdot - \xi_m, Z_m)$, we have

$$p_t(\xi_m, Z_m) := \mathbb{P}(N^{(m)}([0,t]) \ge 1 \mid \xi_m, Z_m) = 1 - \mathrm{e}^{-\int_0^t h(s-\xi_m,Z_m)\,\mathrm{d}s}.$$

Recalling that, given $\{N([t-t_0,t])=k\}$, the k points of N on $[t-t_0,t]$ are independent and uniformly distributed over $[t-t_0,t]$ (see, e.g., [7]), for any $\theta \in \mathbb{R}$, we have

$$\begin{aligned}\mathbb{E}\left[\mathrm{e}^{\theta \widetilde{\mathbf{X}}_{t_0}(t)} \mid N([t-t_0,t])=k\right] &= \mathbb{E}\left[\prod_{m=1}^{k} \mathrm{e}^{\theta \mathbf{1}\{N^{(m)}([0,t])\geq 1\}} \mid N([t-t_0,t])=k\right]\\ &= \left(1+(\mathrm{e}^{\theta}-1)\frac{1}{t_0}\int_{t-t_0}^{t}\mathbb{E}[p_t(u,Z_1)]\,\mathrm{d}u\right)^{k}.\end{aligned}$$

Therefore, since $N([t-t_0,t])$ is Poisson distributed with mean λt_0, we have

$$\mathbb{E}\left[\mathrm{e}^{\theta \widetilde{\mathbf{X}}_{t_0}(t)}\right] = \exp\left(\lambda(\mathrm{e}^{\theta}-1)\int_{-t_0}^{0}\mathbb{E}\left[p_t(u+t,Z_1)\right]\,\mathrm{d}u\right). \tag{6}$$

The claim follows by (5) and (6), letting t_0 tend to ∞. □

In the context of an LRU cache under the SNM, Che's approximation consists in replacing the cache eviction time $T_C(m_0,x)$ by the deterministic constant

$$t_C(m_0,x) := \inf\left\{t>0:\ \mathbb{E}[X_{m_0,x}(t)]=C\right\}.$$

Note that if $g(t)<\infty$ for any $t>0$, then by Proposition 3.1 we have

$$t_C(m_0,x) = \inf\{t>0:\ \lambda g(t)=C\}.$$

So, if moreover $g:(0,\infty)\to(0,\infty)$ is strictly increasing, we deduce

$$t_C(m_0,x) = g^{-1}(C/\lambda). \tag{7}$$

Since the law of $X_{m_0,x}(t)$ (and therefore of $T_C(m_0,x)$) does not depend on m_0 and x, hereafter we simply write $X(t)$, T_C, and t_C in place of $X_{m_0,x}(t)$, $T_C(m_0,x)$, and $t_C(m_0,x)$.

3.3. Asymptotic analysis of T_C.

In this subsection we investigate the validity of Che's approximation for large values of C. We shall do this by analyzing the behavior of T_C as $C\uparrow\infty$. Intuitively, Che's approximation finds a theoretical justification if we may show that, as $C\uparrow\infty$, $T_C/t_C\to 1$ almost surely. This is indeed achieved in Proposition 3.2. Proposition 3.3 provides asymptotic tail estimates for T_C and Corollary 3.4 gives asymptotic upper and lower bounds for the probability that T_C deviates from its most probable value t_C, as C grows large. Finally, the Gaussian approximation for T_C in Proposition 3.10 allows us to construct asymptotic confidence intervals for T_C; see the short discussion after the statement of Proposition 3.10. Hereafter, we shall consider the function g defined by (1).

3.3.1. Law of large numbers and tail estimates for the cache eviction time.

The following law of large numbers and tail estimates holds.

PROPOSITION 3.2. *If $g:(0,\infty)\to(0,\infty)$ is strictly increasing, $g,g^{-1}:(0,\infty)\to(0,\infty)$ are bijective and continuous (i.e., g is a homeomorphism of $(0,\infty)$) and, for any divergent sequences $\{a_n\}_{n\geq 1}$, $\{b_n\}_{n\geq 1}$ of positive numbers,*

$$\lim_{n\to\infty} g(a_n)/g(b_n)=1 \Rightarrow \lim_{n\to\infty} a_n/b_n = 1, \tag{8}$$

then

$$\lim_{C\to\infty}\frac{T_C}{t_C}=1 \quad \textit{almost surely.} \tag{9}$$

Proposition 3.3. *If $g : (0,\infty) \to (0,\infty)$ is strictly increasing and $g, g^{-1} : (0,\infty) \to (0,\infty)$ are bijective and continuous, then*

$$\lim_{C\to\infty} \frac{1}{C} \log \mathbb{P}(T_C > g^{-1}(Cx_r)) = -I(x_r) \quad \forall\ x_r > 1/\lambda \tag{10}$$

and

$$\lim_{C\to\infty} \frac{1}{C} \log \mathbb{P}(T_C \le g^{-1}(Cx_l)) = -I(x_l) \quad \forall\ x_l \in (0, 1/\lambda), \tag{11}$$

where

$$I(x) := \lambda x - 1 - \log(\lambda x), \quad x > 0 \quad \textit{and} \quad I(0) := +\infty. \tag{12}$$

Corollary 3.4. *Under the assumptions of Proposition* 3.3*, we have that for any $\delta \in (0,1)$ and $\varepsilon > 0$ there exists $C_{\delta,\varepsilon}$ so that for any $C > C_{\delta,\varepsilon}$*

$$\mathrm{e}^{-C(I(g(t_C(1+\delta))/C)+\varepsilon)} \le \mathbb{P}(T_C > t_C(1+\delta)) \le \mathrm{e}^{-C(I(g(t_C(1+\delta))/C)-\varepsilon)} \tag{13}$$

and

$$\mathrm{e}^{-C(I(g(t_C(1-\delta))/C)+\varepsilon)} \le \mathbb{P}(T_C \le t_C(1-\delta)) \le \mathrm{e}^{-C(I(g(t_C(1-\delta))/C)-\varepsilon)}, \tag{14}$$

where the function I is defined by (12)*.*

The proofs of Propositions 3.2 and 3.3 are based on a large deviation principle for the process $\{g(T_C)/C\}_{C\ge 1}$ stated in Lemma 3.5. We recall (see, e.g., [10]) that a nonnegative stochastic process $\{Y(t)\}_{t\ge 0}$ obeys a large deviation principle on $[0,\infty)$ with speed v and rate function J if $v : [0,\infty) \to [0,\infty)$ is a function which increases to infinity and $J : [0,\infty) \to [0,\infty]$ is a lower semicontinuous function such that, for all Borel sets $B \subset [0,\infty)$,

$$-\inf_{x\in B^\circ} J(x) \le \liminf_{t\to\infty} \frac{1}{v(t)} \log \mathbb{P}(Y(t) \in B) \le \limsup_{t\to\infty} \frac{1}{v(t)} \log \mathbb{P}(Y(t) \in B) \le -\inf_{x\in \overline{B}} J(x),$$

where B° denotes the interior of B and $\overline{B}$ denotes the closure of B.

For later purposes, we recall that a rate function J on $[0,\infty)$ has no peaks if (i) there exists $\bar{x} \in (0,\infty)$ such that $J(\bar{x}) = 0$; (ii) J is nonincreasing on $(0,\bar{x})$ and nondecreasing on $(\bar{x},\infty)$.

Lemma 3.5. *Under the assumptions of Proposition* 3.3*, we have that the family of random variables $\{g(T_C)/C\}_{C\ge 1}$ obeys a large deviation principle on $[0,\infty)$ with speed $v(C) := C$ and rate function $J := I$ defined by* (12)*.*

Remark 3.6. For later purposes, we remark that the rate function I defined in (12) is continuous on $(0,\infty)$, $I(1/\lambda) = 0$, and I decreases on $(0, 1/\lambda)$ and increases on $(1/\lambda, \infty)$. So, in particular, I has no peaks.

Remark 3.7. Here, we assume that g defined by (1) is a strictly increasing homeomorphism of $(0,\infty)$, and we give sufficient conditions which guarantee (8).

(i) If g^{-1} is ultimately Lipschitz continuous, i.e.,

there exist $K_1, K_2 > 0$ such that $|g^{-1}(x) - g^{-1}(y)| \le K_1|x-y|$, for any $x, y > K_2$,

then (8) holds. Indeed, let $\varepsilon > 0$ be arbitrarily fixed and $n_\varepsilon^{(1)} \ge 1$ so large that $g(b_n) > K_2 + \varepsilon$ for all $n \ge n_\varepsilon^{(1)}$. By the Lipschitz property of g^{-1} we have

$$\sup_{n\geq n_\varepsilon^{(1)}}|g^{-1}(g(b_n)\pm\varepsilon)-b_n| = \sup_{n\geq n_\varepsilon^{(1)}}|g^{-1}(g(b_n)\pm\varepsilon)-g^{-1}(g(b_n))|\leq K_1\varepsilon.$$

Therefore

$$b_n-K_1\varepsilon<g^{-1}(g(b_n)-\varepsilon)\quad\text{and}\quad g^{-1}(g(b_n)+\varepsilon)<b_n+K_1\varepsilon\quad\forall\ n\geq n_\varepsilon^{(1)}. \tag{15}$$

By assumption $g(a_n)/g(b_n)\to 1$, as $n\to\infty$. Therefore there exists $n_\varepsilon^{(2)}\geq 1$ such that for any $n\geq n_\varepsilon^{(2)}$

$$g^{-1}(g(b_n)-\varepsilon)<a_n<g^{-1}(g(b_n)+\varepsilon). \tag{16}$$

The claim follows combining the inequalities (15) and (16).

(ii) If there exists $\bar{t}>0$ such that g is differentiable on $(\bar{t},\infty)$ and $\inf_{t>\bar{t}}g'(t)>0$, then (8) holds. Indeed, for all $x>g(\bar{t})$ we have

$$0<(g^{-1})'(x)=1/g'(g^{-1}(x))\leq(\inf_{t>\bar{t}}g'(t))^{-1};$$

therefore $(g^{-1})'$ is ultimately bounded and the claim follows by the previous point (i).

Example 3.8. Consider the SNM defined by a multiplicative popularity profile of the form (2) and assume (3). In such a case, g is given by (4) and it clearly satisfies the assumptions of Proposition 3.3. We check that g satisfies (8). Setting $H(t):=\int_0^t h(s)\,\mathrm{d}s$, $t>0$, we have

$$\begin{aligned} g(t) &= \int_0^\infty [1-\phi_{Z_1}(-H(u)+H(u-t))]\,\mathrm{d}u \\ &= \int_0^t [1-\phi_{Z_1}(-H(u))]\,\mathrm{d}u+\int_0^\infty [1-\phi_{Z_1}(-H(u+t)+H(u))]\,\mathrm{d}u. \end{aligned}\tag{17}$$

Since $\phi_{Z_1}(\cdot)$ is differentiable on $(-\infty,0)$, we easily have that $g(\cdot)$ is differentiable on $(0,\infty)$ and

$$\begin{aligned} g'(t) &= 1-\phi_{Z_1}(-H(t))+\int_0^\infty \phi'_{Z_1}(-H(u+t)+H(u))\,h(u+t)\,\mathrm{d}u \\ &\geq 1-\phi_{Z_1}(-H(t)), \end{aligned}\tag{18}$$

where the latter inequality follows by the nonnegativity of the third addend in the right-hand side of (18). The claim follows by Remark 3.7(ii) noticing that if $\bar{t}$ is such that $H(\bar{t})>0$, then

$$\inf_{t>\bar{t}}g'(t)\geq 1-\phi_{Z_1}(-H(\bar{t}))>0.$$

In the particular case when

$$h(t,z):=\frac{z}{L}\mathbb{1}_{[0,L]}(t)\quad\text{for some constant } L>0 \tag{19}$$

we have

$$g(t)=2(t\wedge L)+(t\vee L-t\wedge L)\left(1-\phi_{Z_1}(-(t\wedge L)/L)\right)-2\mathbb{E}\left[\frac{L}{Z_1}\left(1-\mathrm{e}^{-\frac{(t\wedge L)Z_1}{L}}\right)\right], \tag{20}$$

where for $a, b \in \mathbb{R}$ we set $a \wedge b := \min\{a, b\}$ and $a \vee b := \max\{a, b\}$. Indeed, for $t > 0$, we have

$$g(t) = \int_0^\infty \Big(1 - \phi_{Z_1}\Big(-\frac{1}{L}\int_{u-t}^{u} \mathbb{1}_{[0,L]}(s)\,\mathrm{d}s\Big)\Big)\,\mathrm{d}u = \int_0^\infty \Big(1 - \phi_{Z_1}(-\eta(t,u))\Big)\,\mathrm{d}u,$$

where

$$\eta(t,u) := \frac{1}{L}\int_{u-t}^{u} \mathbb{1}_{[0,L]}(s)\,\mathrm{d}s = \frac{(u \wedge L - (u-t)^+)^+}{L}$$

and for $a \in \mathbb{R}$ we set $a^+ := a \vee 0$. We distinguish two cases: $0 < t \leq L$ and $t > L$. If $0 < t \leq L$, then if $u < t$ then $u < L$. So, for $t \in (0, L]$,

$$\begin{aligned} g(t) &= \int_0^t \Big(1 - \phi_{Z_1}\Big(-\frac{u}{L}\Big)\Big)\,\mathrm{d}u + \int_t^{L+t} \Big(1 - \phi_{Z_1}\Big(-\frac{1}{L}(t\mathbb{1}_{(0,L]}(u) \\ &\qquad\qquad + (t-u+L)\mathbb{1}_{(L,L+t]}(u))\Big)\Big)\,\mathrm{d}u \\ &= t - \mathbb{E}\left[\frac{L}{Z_1}\Big(1 - \mathrm{e}^{-\frac{Z_1}{L}t}\Big)\right] + (L-t)\Big(1 - \phi_{Z_1}\Big(-\frac{t}{L}\Big)\Big) \\ &\quad + \int_L^{L+t} \Big(1 - \phi_{Z_1}\Big(-\frac{1}{L}(t-u+L)\Big)\Big)\,\mathrm{d}u \\ &= 2t + (L-t)\Big(1 - \phi_{Z_1}\Big(-\frac{t}{L}\Big)\Big) - 2\mathbb{E}\left[\frac{L}{Z_1}\Big(1 - \mathrm{e}^{-\frac{Z_1}{L}t}\Big)\right]. \end{aligned}$$

If $t > L$, then if $u \geq t$ then $u \geq t > L$. So, for $t > L$,

$$\begin{aligned} g(t) &= \int_0^L \Big(1 - \phi_{Z_1}\Big(-\frac{u}{L}\Big)\Big)\,\mathrm{d}u + (t-L)(1 - \phi_{Z_1}(-1)) \\ &\qquad + \int_t^{L+t} \Big(1 - \phi_{Z_1}\Big(-\frac{t-u+L}{L}\Big)\Big)\,\mathrm{d}u \\ &= L - \mathbb{E}\left[\frac{L}{Z_1}\left(1 - \mathrm{e}^{-Z_1}\right)\right] + (t-L)(1 - \phi_{Z_1}(-1)) + L - \mathbb{E}\left[\frac{L}{Z_1}\left(1 - \mathrm{e}^{-Z_1}\right)\right] \\ &= 2L + (t-L)(1 - \phi_{Z_1}(-1)) - 2\mathbb{E}\left[\frac{L}{Z_1}\left(1 - \mathrm{e}^{-Z_1}\right)\right]. \end{aligned}$$

Proof of Proposition 3.2. It is well-known that

$$\lim_{C\to\infty} \frac{g(T_C)}{C} = 1/\lambda \quad \text{almost surely} \tag{21}$$

if and only if

$$\mathbb{P}\left(\bigcap_{n\geq 1}\bigcup_{C\geq n}\Big|\frac{g(T_C)}{C} - \frac{1}{\lambda}\Big| > \varepsilon\right) = 0 \quad \text{for any } \varepsilon > 0.$$

Therefore (21) follows by the Borel–Cantelli lemma if we check that

$$\sum_{C\geq 1} \mathbb{P}\left(\Big|\frac{g(T_C)}{C} - \frac{1}{\lambda}\Big| > \varepsilon\right) < \infty \quad \text{for any } \varepsilon > 0.$$

Let $\varepsilon \in (0,\lambda^{-1})$ be arbitrarily fixed and take $\delta \in (0, I(\lambda^{-1}-\varepsilon) \wedge I(\lambda^{-1}+\varepsilon))$. By Lemma 3.5 we have that there exists a nonnegative integer C_δ such that for any $C \geq C_\delta$

$$\mathbb{P}\left(\frac{g(T_C)}{C} \geq \lambda^{-1}+\varepsilon\right) \leq \mathrm{e}^{-\left(\inf_{x\geq\lambda^{-1}+\varepsilon} I(x)-\delta\right)C} = \mathrm{e}^{-(I(\lambda^{-1}+\varepsilon)-\delta)C}$$

and

$$\mathbb{P}\left(\frac{g(T_C)}{C} \leq \lambda^{-1}-\varepsilon\right) \leq \mathrm{e}^{-\left(\inf_{x\leq\lambda^{-1}-\varepsilon} I(x)-\delta\right)C} = \mathrm{e}^{-(I(\lambda^{-1}-\varepsilon)-\delta)C},$$

where we used Remark 3.6. Therefore

$$\sum_{C\geq C_\delta} \mathbb{P}\left(\left|\frac{g(T_C)}{C}-\frac{1}{\lambda}\right| > \varepsilon\right) \leq \sum_{C\geq C_\delta} \mathbb{P}\left(\frac{g(T_C)}{C} \geq \lambda^{-1}+\varepsilon\right)$$
$$+\sum_{C\geq C_\delta} \mathbb{P}\left(\frac{g(T_C)}{C} \leq \lambda^{-1}-\varepsilon\right) \leq \sum_{C\geq C_\delta} \mathrm{e}^{-(I(\lambda^{-1}+\varepsilon)-\delta)C} + \sum_{C\geq C_\delta} \mathrm{e}^{-(I(\lambda^{-1}-\varepsilon)-\delta)C} < \infty,$$

which proves (21). Then the claim follows by assumption (8) noticing that by (21) and relation $t_C = g^{-1}(C/\lambda)$ we easily get $g(T_C)/g(t_C) \to 1$ almost surely, as $C \to \infty$. □

Proof of Proposition 3.3. By Remark 3.6 we have that, for any $x_r > 1/\lambda$, $\inf_{y>x_r} I(y) = \inf_{y\geq x_r} I(y) = I(x_r)$, and, for any $x_l \in (0,1/\lambda)$, $\inf_{y<x_l} I(y) = \inf_{y\leq x_l} I(y) = I(x_l)$. Relations (10) and (11) follow by applying the large deviation principle of Lemma 3.5 considering, respectively, the Borel sets $B = (x_r,\infty)$ and $B = (0,x_l)$ (note that g^{-1} is strictly increasing since g is such). □

Proof of Corollary 3.4. The claim easily follows by (7), (10), and (11). □

The proof of Lemma 3.5 uses a result from [12], which we recall for the sake of clarity. We first introduce some notation and terminology. Let $\{Y(t)\}_{t\geq 0}$ be a nonnegative stochastic process whose sample paths are right-continuous, nondecreasing, and such that $\lim_{t\to\infty} Y(t) = \infty$ almost surely. We define the inverse process of $\{Y(t)\}_{t\geq 0}$ as

$$W(z) := \inf\{t \geq 0 :\ Y(t) \geq z\}, \quad z \geq 0.$$

The following theorem holds (see Theorem 1(i) in [12]).

Theorem 3.9. *Let $\{Y(t)\}_{t\geq 0}$ and $\{W(z)\}_{z\geq 0}$ be as above and let $v : (0,\infty) \to (0,\infty)$ be a strictly increasing homeomorphism of $(0,\infty)$. We have that if $\{Y(t)/v(t)\}_{t\geq 0}$ obeys a large deviation principle on $[0,\infty)$ with speed v and rate function I which has no peaks, then $\{v(W(z))/z\}_{z>0}$ obeys a large deviation principle on $[0,\infty)$ with speed $\tilde{v}(z) := z$ and rate function $\tilde{I}(z) := zI(1/z)$, $z > 0$, $\tilde{I}(0) := \lim_{z\to 0^+} zI(1/z)$.*

Proof of Lemma 3.5. Let $\theta \in \mathbb{R}$ be arbitrarily fixed. By Proposition 3.1 $X(t)$ is Poisson distributed with mean $\lambda g(t)$. Therefore

$$\lim_{t\to\infty} \frac{1}{g(t)} \log \mathbb{E}\left[\mathrm{e}^{\theta X(t)}\right] = \lim_{t\to\infty} \frac{1}{g(t)} \log \mathrm{e}^{\lambda g(t)(\mathrm{e}^{\theta}-1)} = \lambda(\mathrm{e}^{\theta}-1) := \Lambda(\theta).$$

So by the Gärtner–Ellis theorem (see, e.g., [10]) the stochastic process $\{X(t)/g(t)\}_{t\geq 1}$ satisfies a large deviation principle on $[0,\infty)$ with speed g and rate function $\Lambda^*(x) :=$

$\sup_{\theta\in\mathbb{R}}(\theta x-\Lambda(\theta))=\lambda-x+x\log(x/\lambda)$, $x>0$, $\Lambda^*(0):=\lambda$. Note that $\{T_C\}_{C\geq 1}$ is the inverse process of $\{X(t)\}_{t\geq 0}$. The claim then follows by Theorem 3.9. Indeed the rate function Λ^* has no peaks since $\Lambda^*(\lambda)=0$ and Λ^* decreases on $(0,\lambda)$ and increases on (λ,∞). □

3.3.2. Normal approximation of the cache eviction time. Hereafter, we denote by $\mathcal{N}(0,1)$ a standard normal random variable and by $\xrightarrow{law}$ the convergence in distribution. Following the ideas in [15] (see Propositions 1 and 3 therein), we derive a central limit theorem for the cache eviction time of the SNM.

PROPOSITION 3.10. *Assume $g:(0,\infty)\to(0,\infty)$ bijective, strictly increasing, and such that there exists a positive function f such that*

$$\lim_{y\to\infty} f(y)\in[0,\infty] \quad \textit{and} \quad \lim_{y\to\infty}\frac{g(y)-g(y+xf(y))}{\sqrt{g(y+xf(y))}}=-\frac{x}{\sqrt{\lambda}}. \tag{22}$$

Then

$$\frac{T_C-t_C}{f(t_C)}\xrightarrow{law}\mathcal{N}(0,1) \quad \textit{as } C\to\infty. \tag{23}$$

Under the assumptions of Proposition 3.10 one can construct asymptotic confidence intervals for T_C. Indeed, if $\nu>0$ is such that $\mathbb{P}(|\mathcal{N}(0,1)|\leq\nu)=\mu\in(0,1)$, then, as $C\to\infty$, $[t_C-\nu f(t_C),t_C+\nu f(t_C)]$ is an asymptotic confidence interval for T_C at the level μ, as the following simple computation shows:

$$\begin{aligned}\mathbb{P}\left(t_C-\nu f(t_C)\leq T_C\leq t_C+\nu f(t_C)\right)&=\mathbb{P}\left(\Big|\frac{T_C-t_C}{f(t_C)}\Big|\leq\nu\right)\\&\simeq\mathbb{P}(|\mathcal{N}(0,1)|\leq\nu)=\mu \quad \text{as } C\to\infty.\end{aligned}$$

Clearly, for a fixed level μ, by using the tables of the Gaussian distribution one finds the value ν which determines the asymptotic confidence interval.

Example 3.11. Consider the SNM defined by a multiplicative popularity profile of the form (2) and assume (3) and

$$\int_0^\infty th(t)\,\mathrm{d}t<\infty.$$

In such a case, g is given by (4) and, as noticed in Example 3.8, g is an increasing homeomorphism of $(0,\infty)$. We shall check later on that (22) holds with

$$f(x):=\sqrt{\frac{x}{\lambda(1-\phi_{Z_1}(-1))}}. \tag{24}$$

Therefore we have the normal approximation (23) and asymptotic confidence intervals for T_C can be constructed as described above. To verify (22) we start noticing that

$$\lim_{t\to\infty}\frac{g(t)}{t(1-\phi_{Z_1}(-1))}=1. \tag{25}$$

Indeed, by l'Hôpital's rule we have

$$\begin{aligned}\lim_{t\to\infty}\frac{1}{t}\int_0^t\Big(1-\phi_{Z_1}(-H(u))\Big)\,\mathrm{d}u&=1-\lim_{t\to\infty}\frac{1}{t}\int_0^t\phi_{Z_1}(-H(u))\,\mathrm{d}u\\&=1-\phi_{Z_1}(-1).\end{aligned}$$

So (25) follows if we check that the second term in (17) is bounded. By the elementary inequality $\mathrm{e}^x \geq 1 + x$, $x \in \mathbb{R}$, we have

$$\begin{aligned} 0 &\leq \int_t^\infty \Big(1 - \phi_{Z_1}(H(u-t) - H(u))\Big)\,\mathrm{d}u \\ &\leq \mathbb{E}[Z_1] \int_t^\infty (H(u) - H(u-t))\,\mathrm{d}u \\ &\leq \mathbb{E}[Z_1] \int_0^\infty (1 - H(u))\,\mathrm{d}u = \mathbb{E}[Z_1] \int_0^\infty u h(u)\,\mathrm{d}u < \infty, \end{aligned}$$

where the latter equality is a consequence of the fact that h is a probability density on $(0, \infty)$. Finally, we check that (25) implies the second limit in (22) (the first limit being obvious by the definition of f). Letting $o(1)$ denote a function which tends to zero as $y \to \infty$, by (24) we have

$$\begin{aligned} \lim_{y\to\infty} \frac{g(y) - g(y + xf(y))}{\sqrt{g(y + xf(y))}} &= \lim_{y\to\infty} \frac{y(1 - \phi_{Z_1}(-1)) - (y + xf(y))(1 - \phi_{Z_1}(-1)) + o(1)}{\sqrt{(y + xf(y))(1 - \phi_{Z_1}(-1)) + o(1)}} \\ &= -\lim_{y\to\infty} \frac{xf(y)(1 - \phi_{Z_1}(-1)) + o(1)}{\sqrt{(y + xf(y))(1 - \phi_{Z_1}(-1)) + o(1)}} \\ &= -\lim_{y\to\infty} \frac{x(1-\phi_{Z_1}(-1))\sqrt{y}+o(1)}{\sqrt{\lambda(1-\phi_{Z_1}(-1))}\sqrt{(y+xf(y))(1-\phi_{Z_1}(-1))+o(1)}} \\ &= -\lim_{y\to\infty} \frac{x(1 - \phi_{Z_1}(-1))\sqrt{y}}{\sqrt{\lambda(1 - \phi_{Z_1}(-1))}\sqrt{(1 - \phi_{Z_1}(-1))y}} = -\frac{x}{\sqrt{\lambda}}. \end{aligned}$$

The proof of Proposition 3.10 uses Lemma 3.12 below, which is of its own interest. We denote by Lip(1) the class of real-valued Lipschitz functions from $\mathbb{R}$ to $\mathbb{R}$ with Lipschitz constant less than or equal to one. Given two real-valued random variables U and U', the Wasserstein distance between the laws of U and U', written $d_W(U, U')$, is defined as

$$d_W(U, U') := \sup_{\varphi \in \mathrm{Lip}(1)} |\mathbb{E}[\varphi(U)] - \mathbb{E}[\varphi(U')]|.$$

We recall that the topology induced by d_W on the class of probability measures over $\mathbb{R}$ is finer than the topology of weak convergence (see, e.g., [26]).

Lemma 3.12. *If $g(t) < \infty$, then*

$$d_W\left(\frac{X(t) - \lambda g(t)}{\sqrt{\lambda g(t)}}, \mathcal{N}(0,1)\right) \leq \frac{1}{\sqrt{\lambda g(t)}}.$$

Remark 3.13. Lemma 3.12 provides a Gaussian approximation for $X(t)$ in the Wasserstein distance. On the other hand, Proposition 1 in [15] provides a Gaussian approximation, in the Kolmogorov distance d_{Kol}, for the corresponding quantity under the IRM. We note that a Gaussian approximation of $X(t)$, in the Kolmogorov distance, under the SNM may be easily obtained by Lemma 3.12 using the relation

$$d_{Kol}(X, \mathcal{N}(0,1)) := \sup_{x\in\mathbb{R}} |\mathbb{P}(X \leq x) - \mathbb{P}(\mathcal{N}(0,1) \leq x)| \leq 2\sqrt{d_W(X, \mathcal{N}(0,1))},$$

where X is a real-valued random variable; see, e.g., [26].

Proof of Proposition 3.10. By the assumptions on g we have $C = \lambda g(t_C)$, $t_C \uparrow \infty$, and $g(t_C) \uparrow \infty$, as $C \uparrow \infty$. For any $x \in \mathbb{R}$,

$$\begin{aligned} \mathbb{P}(T_C - t_C > xf(t_C)) &= \mathbb{P}(X(t_C + xf(t_C)) < C) \\ &= \mathbb{P}\left(\frac{X(t_C + xf(t_C)) - \lambda g(t_C + xf(t_C))}{\sqrt{\lambda g(t_C + xf(t_C))}} < \sqrt{\lambda}\frac{g(t_C) - g(t_C + xf(t_C))}{\sqrt{g(t_C + xf(t_C))}}\right). \end{aligned} \tag{26}$$

By Lemma 3.12 we have

$$\frac{X(t) - \lambda g(t)}{\sqrt{\lambda g(t)}} \xrightarrow{law} \mathcal{N}(0,1) \quad \text{as } t \to \infty.$$

So, letting C tend to infinity in (26) and using (22) we deduce

$$\lim_{C\to\infty} \mathbb{P}\left(\frac{T_C - t_C}{f(t_C)} > x\right) = \mathbb{P}(\mathcal{N}(0,1) \le -x) = \mathbb{P}(\mathcal{N}(0,1) > x). \qquad \square$$

Proof of Lemma 3.12. Define the Borel measure $\mu(\mathrm{d}x) := \lambda \mathrm{d}g(x)$ over $[0,t]$ (note that g increases on $[0,t]$ and so $\mathrm{d}g$ is a Lebesgue–Stieltjes measure) and the function $h(x) := \mathbb{1}_{[0,t]}(x)/\sqrt{\lambda g(t)}$, $x \in [0,t]$. By Corollary 3.4 in [27] and Proposition 3.1, we have

$$\begin{aligned} d_W\left(\frac{X(t) - \lambda g(t)}{\sqrt{\lambda g(t)}}, \mathcal{N}(0,1)\right) &\le \left|1 - \int_{[0,t]} |h(x)|^2 \mu(\mathrm{d}x)\right| + \int_{[0,t]} |h(x)|^3 \mu(\mathrm{d}x) \\ &= \frac{1}{\sqrt{\lambda g(t)}}. \end{aligned} \qquad \square$$

3.4. The "in" and the "hit" probabilities. The results of the previous subsection provide a justification to the Che approximation, as $C \to \infty$. Indeed, the law of large numbers (9) guarantees that, asymptotically in C, t_C is a correct approximation of T_C; the bounds (13) and (14) guarantee that deviations of T_C from its most probable value t_C are, asymptotically in C, exponentially small; the normal approximation (23) allows us to identify the typical asymptotic values of T_C via the construction of asymptotic confidence intervals. In this subsection we provide complementary nonasymptotic analytical upper bounds on the prediction error entailed by Che's approximation of the "hit" probability (see Proposition 3.14). This result allows us to assess the accuracy of the Che approximation in many cases of practical interest; cf. section 5.

3.4.1. The "in" probability. The "in" probability is defined as the probability of finding at time t a tagged content m_0 in the cache, given that $\xi_{m_0} = x$ and $Z_{m_0} = z$. Thus:

(i) Under Che's approximation, the "in" probability is given by

$$\begin{aligned} p_{\text{in,Che}}^{(t-x)}(z, t_C) &:= \mathbb{P}(N^{(m_0)}((t - t_C, t]) \ge 1 \mid (\xi_{m_0}, Z_{m_0}) = (x,z)) \\ &= 1 - \mathrm{e}^{-\int_{t-x-t_C}^{t-x} h(u,z)\,\mathrm{d}u}, \end{aligned} \tag{27}$$

where $N^{(m_0)}(A)$ denotes the number of points $\{T_n^{(m_0)}\}_{n\ge1}$ in $A \subset \mathbb{R}$.

(ii) Without relying on Che's approximation, the conditional "in" probability is given by

$$\begin{aligned} p_{\text{in}}^{(t-x)}(z, T_C) &:= \mathbb{P}(N^{(m_0)}((t - T_C, t]) \ge 1 \,|\, (\xi_{m_0}, Z_{m_0}) = (x,z), T_C) \\ &= p_{\text{in,Che}}^{(t-x)}(z, T_C). \end{aligned}$$

3.4.2. The "hit" probability. The "hit" probability is defined as the ratio between the average rate at which "hits" of a tagged content occur and the average rate at which requests of the tagged content are observed. Thus,

(iii) under Che's approximation, the "hit" probability is given by

$$\begin{aligned} p_{\mathrm{hit,Che}}(t_C) : &= \frac{\mathbb{E}[h(t-\xi_{m_0}, Z_{m_0}) p^{(t-\xi_{m_0})}_{\mathrm{hit,Che}}(Z_{m_0}, t_C)]}{\mathbb{E}[h(t-\xi_{m_0}, Z_{m_0})]} \\ &= \frac{\mathbb{E}[h(t-\xi_{m_0}, Z_{m_0}) p^{(t-\xi_{m_0})}_{\mathrm{in,Che}}(Z_{m_0}, t_C)]}{\mathbb{E}[h(t-\xi_{m_0}, Z_{m_0})]} \end{aligned} \tag{28}$$

with the convention $0/0 = 0$. The equality (28) is a consequence of the fact that, under Che's approximation, the probability (denoted by $p^{(t-x)}_{\mathrm{hit,Che}}(z, t_C)$) that the tagged content m_0, introduced into the catalogue at time $\xi_{m_0} = x$ and with mark $Z_{m_0} = z$, is found in the cache by an arriving request at time t is equal to $p^{(t-x)}_{\mathrm{in,Che}}(z, t_C)$. Indeed

$$\begin{aligned} p^{(t-x)}_{\mathrm{hit,Che}}(z, t_C) : &= \mathbb{P}\Big(\sum_{T_n^{(m_0)} \in N^{(m_0)} \setminus \{t\}} \mathbb{1}_{(t-t_C, t]}(T_n^{(m_0)}) \geq 1 \,\Big|\, t \in N^{(m_0)}, \, (\xi_{m_0}, Z_{m_0}) \\ &= (x, z) \Big) \\ &= \mathbb{P}\Big(N^{(m_0)}((t-t_C, t]) \geq 1) \Big| (\xi_{m_0}, Z_{m_0}) = (x, z) \Big) \\ &= p^{(t-x)}_{\mathrm{in,Che}}(z, t_C), \end{aligned} \tag{29}$$

where the equality (29) is a consequence of the Slivnyak–Mecke theorem (see, e.g., Proposition 13.1.VII, p. 281, in [8]).

Note that the probability $p_{\mathrm{hit,Che}}(t_C)$ does not depend on m_0 and t. Indeed, for an arbitrary $s < t$ we have

$$\begin{aligned} &\frac{\mathbb{E}\left[h(t-\xi_{m_0}, Z_{m_0}) p^{(t-\xi_{m_0})}_{\mathrm{in,Che}}(Z_{m_0}, t_C) \mathbb{1}\{s < \xi_{m_0} < t\} \right]}{\mathbb{E}\left[h(t-\xi_{m_0}, Z_{m_0}) \mathbb{1}\{s < \xi_{m_0} < t\} \right]} \\ &\quad = \frac{(t-s)^{-1} \int_s^t \mathbb{E}\left[h(t-u, Z_1) p^{(t-u)}_{\mathrm{in,Che}}(Z_1, t_C) \right] \mathrm{d}u}{(t-s)^{-1} \int_s^t \mathbb{E}\left[h(t-u, Z_1) \right] \mathrm{d}u} \\ &\quad = \frac{\int_s^t \mathbb{E}\left[h(t-u, Z_1) p^{(t-u)}_{\mathrm{in,Che}}(Z_1, t_C) \right] \mathrm{d}u}{\int_s^t \mathbb{E}\left[h(t-u, Z_1) \right] \mathrm{d}u}, \end{aligned}$$

and so letting s tend to $-\infty$ we deduce

$$p_{\mathrm{hit,Che}}(t_C) = \frac{\int_0^\infty \mathbb{E}\left[h(u, Z_1) p^{(u)}_{\mathrm{in,Che}}(Z_1, t_C) \right] \mathrm{d}u}{\int_0^\infty \mathbb{E}\left[h(u, Z_1) \right] \mathrm{d}u}. \tag{30}$$

(iv) Without relying on Che's approximation, the conditional "hit" probability is given by

$$p_{\mathrm{hit}}(T_C) := \frac{\mathbb{E}\left[h(t-\xi_{m_0}, Z_{m_0}) p^{(t-\xi_{m_0})}_{\mathrm{in,Che}}(Z_{m_0}, T_C) \,|\, T_C \right]}{\mathbb{E}\left[h(t-\xi_{m_0}, Z_{m_0}) \right]}.$$

Arguing as above, one may easily check that

$$p_{\text{hit}}(T_C) = \frac{\int_0^\infty \mathbb{E}\left[h(u, Z_{m_0}) p^{(u)}_{\text{in,Che}}(Z_{m_0}, T_C) \mid T_C\right] \mathrm{d}u}{\int_0^\infty \mathbb{E}\left[h(u, Z_1)\right] \mathrm{d}u}.$$

Being Z_{m_0} and T_C independent, the (unconditional) "hit" probability is given by

$$p_{\text{hit}} = \int_{[0,\infty)} p_{\text{hit,Che}}(\theta)\, \mathbb{P}_{T_C}(\mathrm{d}\theta),$$

where $\mathbb{P}_{T_C}$ denotes the law of T_C and $p_{\text{hit,Che}}(\theta)$ is defined as $p_{\text{hit,Che}}(t_C)$ with θ in place of t_C.

3.4.3. Error estimate. By using the above relations and classical estimates for the tail of a Poisson distribution, we can evaluate the error committed by approximating p_{hit} with $p_{\text{hit,Che}}(t_C)$. The following proposition holds.

PROPOSITION 3.14. *If $g : (0,\infty) \to (0,\infty)$ is strictly increasing, then, for any $\delta \in (0,1)$ and $C > 0$, we have*

$$\begin{aligned}|p_{\text{hit}} - p_{\text{hit,Che}}(t_C)| \le\ & \exp(-\lambda g(t_C(1-\delta))R(C/\lambda g(t_C(1-\delta))))\\ & + \exp(-\lambda g(t_C(1+\delta))R(C/\lambda g(t_C(1+\delta))))\\ & + \max_{\theta\in\{t_C(1-\delta),t_C(1+\delta)\}} |p_{\text{hit,Che}}(\theta) - p_{\text{hit,Che}}(t_C)|,\end{aligned}$$

where $R(x) := 1 - x + x\log x$, $x > 0$.

This proposition allows an assessment of the accuracy of Che's approximation in different scenarios. As shown by the numerical simulations in [22] (see also section 5), in most cases by exploiting Proposition 3.14 we can show that Che's approximation leads to surprisingly accurate predictions of caching performance.

Proof of Proposition 3.14. We preliminary note that, for any $\delta \in (0,1)$ and $C > 0$, we have

$$\lambda g(t_C(1-\delta)) \le C \le \lambda g(t_C(1+\delta)). \tag{31}$$

Indeed, since g is strictly increasing (31) is equivalent to $t_C(1-\delta) \le g^{-1}(C/\lambda) \le t_C(1+\delta)$, which holds since $t_C = g^{-1}(C/\lambda)$. Note that, due to (27), $p_{\text{hit,Che}}(\cdot)$ is a nondecreasing function. So, for all $\delta \in (0,1)$, we have

$$\begin{aligned}|p_{\text{hit}} - p_{\text{hit,Che}}(t_C)| &\le \int_{[0,t_C(1-\delta)]} (p_{\text{hit,Che}}(t_C) - p_{\text{hit,Che}}(\theta))\, \mathbb{P}_{T_C}(\mathrm{d}\theta)\\ &\quad + \int_{(t_C(1+\delta),\infty)} (p_{\text{hit,Che}}(\theta) - p_{\text{hit,Che}}(t_C))\, \mathbb{P}_{T_C}(\mathrm{d}\theta)\\ &\quad + \int_{(t_C(1-\delta),t_C(1+\delta)]} |p_{\text{hit,Che}}(\theta) - p_{\text{hit,Che}}(t_C)|\, \mathbb{P}_{T_C}(\mathrm{d}\theta)\\ &\le \mathbb{P}(T_C \le t_C(1-\delta)) + \mathbb{P}(T_C > t_C(1+\delta))\\ &\quad + \max_{\theta\in\{t_C(1-\delta),t_C(1+\delta)\}} |p_{\text{hit,Che}}(\theta) - p_{\text{hit,Che}}(t_C)|.\end{aligned}$$

The claim follows noticing that by the definition of T_C, the inequality (31) and the properties of the Poisson distribution (see, e.g., Lemma 1.2 in [28], formulas (1.10) and (1.11)) we have

$$\mathbb{P}(T_C \le t_C(1-\delta)) = \mathbb{P}(X(t_C(1-\delta)) > C) \le \exp(-\lambda g(t_C(1-\delta))R(C/\lambda g(t_C(1-\delta))))$$

and

$$\begin{aligned}\mathbb{P}(T_C > t_C(1+\delta)) = \mathbb{P}(X(t_C(1{+}\delta))&\\ \le C) \le \exp(-\lambda g(t_C(1{+}\delta))R(C/\lambda g(t_C(1{+}\delta)))).&\end{aligned} \qquad \square$$

3.5. Extension to the case of contents with variable sizes. At a first glance, dealing with contents of variable sizes may appear significantly more challenging. Indeed, before inserting in the cache a new content, enough memory must be freed by selecting a proper set of objects to expunge. The content to be stored, then, needs to be partitioned into small portions (fragments) that fit into the nonadjacent areas of memory, each one corresponding to a different fragment of the expunged contents. Unfortunately an excessive fragmentation of the contents can significantly reduce the bandwidth performance (speed) of the cache and therefore must be prevented by executing complex memory management operations such as periodic defragmentation. A simple method to cache contents of variable sizes, referred to in the following as *chunkization*, consists in breaking each content into an integer number of pieces with a fixed size, called *chunks*, which are treated as independent objects by the caching system. By properly dimensioning the size of the chunk it is possible to achieve an optimal trade-off between memory efficiency and bandwidth performance. Indeed, by enlarging the size of the chunk, memory efficiency decreases (for the effect of the last chunk size rounding), while the cache speed increases since the size of fragments (which are memorized in consecutive memory locations) increases. In this way the degradation of the cache due to content fragmentation is kept under control, without the necessity of executing complex memory management operations. This is the main reason why chunkization has become an almost universally adopted technique in caching systems supporting the distribution of contents over the Internet [19, 29, 11].

In this subsection we briefly discuss how our approach can be extended to evaluate the effectiveness of Che's approximation for an LRU cache which stores contents of variable sizes through chunkization.

We still assume that the LRU cache operates under the SNM: requests of different chunks corresponding to the same content m are perfectly synchronized, and the process of requests for each chunk of content m is a Cox process with stochastic intensity $\lambda_m(t)$. We denote by A_m the number of chunks in which the content m is partitioned and assume that $\{A_m\}_{m\geq 1}$ is a sequence of independent and identically distributed random variables with values on $\{1, 2, \ldots\}$, independent of $\{\xi_m\}_{m\geq 1}$ and $\{Z_m\}_{m\geq 1}$. The number of chunks (corresponding to contents different from m_0) requested in the time interval $[0, t]$ is given by

$$\mathbf{X}_{m_0}(t) := \sum_{m\neq m_0} A_m \mathbb{1}\left\{m \text{ requested in } [0,t],\, \xi_m \in (-\infty, t]\right\}.$$

Setting $X_{m_0,x}(t) := (\mathbf{X}_{m_0}(t) \,|\, \xi_{m_0} = -x) + A_{m_0} - 1$, with $x > 0$, we define the cache eviction time as

$$\begin{aligned} T_C(m_0, x) : &= \inf\{t \geq 0 : \ X_{m_0,x}(t) = C\} \\ &= \inf\{t \geq 0 : \ (\mathbf{X}_{m_0}(t) \,|\, \xi_{m_0} = -x) = C - A_{m_0} + 1\}, \end{aligned}$$

where we express the caching storage capacity C in number of chunks. The definition of $T_C(m_0, x)$ reflects the fact that we consider a content to be expunged (i.e., unavailable at the cache) when its first chunk is expunged by the cache.

Let g be the function defined by (1). If $g(t) < \infty$ and the A's are light-tail, i.e.,

$$\exists \text{ a right neighborhood of zero, say } \mathcal{N}_+, \text{ such that } \phi_{A_1}(\theta) := \mathbb{E}[\mathrm{e}^{\theta A_1}] < \infty \ \forall \ \theta \in \mathcal{N}_+, \tag{32}$$

then, arguing as in the proof of Proposition 3.1, one has that $\mathbf{X}_{m_0}(t) \,|\, \xi_{m_0} = -x$ follows the same law of $\sum_{i=1}^{S} A_i$, where S is a Poisson distributed random variable with mean $\lambda g(t)$ and S is independent of $\{A_m\}_{m\geq 1}$. Note that the laws of $X_{m_0,x}(t)$ and $T_C(m_0, x)$ do not depend on x, but they depend on m_0. However, for ease of notation, hereafter we do not explicitly indicate this dependence, writing $X(t)$ and T_C in place of $X_{m_0,x}(t)$ and $T_C(m_0, x)$. In this context, Che's approximation of T_C is

$$t_C := \inf \{t \geq 0 : \ \mathbb{E}[X_{m_0,x}(t)] = C\} = \inf \{t \geq 0 : \ \lambda g(t) = (C + 1 - \mathbb{E}[A_1])/\mathbb{E}[A_1]\} .$$

Under the same assumptions as Proposition 3.3 and condition (32), we have that the family of random variables $\{g(T_C)/C\}_{C\geq 1}$ obeys a large deviation principle on $[0, \infty)$ with speed $v(C) := C$ and rate function $I(x) := x\Lambda^*(1/x)$, $x > 0$, $I(0) := \lim_{x\to 0^+} x\Lambda^*(1/x)$, where

$$\Lambda^*(x) := \sup_{\theta\in\mathbb{R}} (\theta x - \lambda(\mathbb{E}[\mathrm{e}^{\theta A_1}] - 1)).$$

Since the derivation of this large deviation principle is not immediate, we sketch the proof. Arguing as in the proof of Lemma 3.5 one has that the process $\{\tilde{X}(t)/g(t)\}_{t\geq 1}$, where $\tilde{X}(t) := \mathbf{X}_{m_0}(t) \,|\, \xi_{m_0} = -x$, obeys a large deviation principle on $[0, \infty)$ with speed g and rate function $I_1(u) := \Lambda^*(u)$. On the other hand, by using the definition of large deviation principle it is readily checked that the process $\{(A_{m_0} - 1)/g(t)\}_{t\geq 1}$ obeys a large deviation principle on $[0, \infty)$ with speed g and rate function $I_2(v) := +\infty 1\!\!1\{v > 0\}$, with the convention $\infty \cdot 0 = 0$. By the independence of the processes $\{\tilde{X}(t)/g(t)\}_{t\geq 1}$ and $\{(A_{m_0} - 1)/g(t)\}_{t\geq 1}$ and the contraction principle (i.e., by Exercise 4.2.7 on p. 129 and Theorem 4.2.1 on p. 126 in [10]) one has that the process $\{X(t)/g(t)\}_{t\geq 1}$ obeys a large deviation principle on $[0, \infty)$ with speed g and rate function

$$\inf\{I_1(u) + I_2(v) : \ u + v = x\} = I_1(x) = \Lambda^*(x).$$

The claimed large deviation principle for the family of random variables $\{g(T_C)/C\}_{C\geq 1}$ follows by applying Theorem 3.9 as in the proof of Lemma 3.5 (note that $\{T_C\}$ is the inverse process of $\{X(t)\}$).

By this large deviation principle one can obtain the law of large numbers (9), the tail estimates (10), (11), and the deviation bounds (13), (14); cf. the proofs of Propositions 3.2 and 3.3 and Corollary 3.4, respectively.

Finally, we note that the proofs of Propositions 3.10 and 3.14 may be easily adapted in order to obtain a normal approximation of the cache eviction time and an estimate of the error committed by approximating the corresponding "hit" probability with its expression under Che's approximation; we omit the details.

4. Networks of caches: The case of two caches is series.

The analysis of networks of caches is a difficult task; indeed an exact characterization of the miss stream of an LRU cache is in general prohibitive. Under the IRM a standard and rather crude approach proposed in the literature (see, e.g., [30]) consists in (i) approximating the miss stream of a content at a cache with a homogeneous Poisson process whose rate matches the miss stream rate; (ii) assuming the state of caches to be independent. However, significant errors may be experienced. An alternative approach, that has been recently proposed for feed-forward networks of LRU caches (such as networks with linear topologies or trees) consists in approximating the real miss stream with that of a cache operating under Che's approximation (see [1] and [14]). This approach has been experimentally shown to be potentially fairly accurate, but, unfortunately, at the same time, it is computationally highly expensive [14]. Recently a more efficient procedure has been proposed in [1], where further approximations are considered to simplify the computation of the "hit" probability. However, in this latter case, the accuracy of the estimate is in part sacrificed.

Here we show how the approach of [14] can be adapted to the SNM. Our study reveals that the exact computation of the "hit" probability, under Che's approximation, for a simple tandem network of caches (i.e., a network of two LRU caches in series) is computationally hard (see Remark 4.2). This is mainly due to the effect of the complex dependencies between the states of the two caches.

Since the analysis at the first cache can be carried on as in the previous section, here we focus on the second cache. Note that an arriving request for content m_0 can produce a "hit" at the second cache only if it misses the content m_0 at the first cache. So, under Che's approximation, the "hit" probability for content m_0, introduced into the catalogue at time $\xi_{m_0}=x$ and with mark $Z_{m_0}=z$, is given by

$$\begin{aligned} &p^{(t-x)}_{\mathrm{hit,Che,II}}(z,t_{C_1},t_{C_2}) := \mathbb{P}\left(\sum_{T_n^{(m_0)}\in N^{(m_0)}\setminus\{t\}} \mathbb{1}_{(t-t_{C_1},t]}(T_n^{(m_0)})=0,\right.\\ &\quad\left.\sum_n \mathbb{1}_{(t-t_{C_2},t)}(T_n^{(m_0)})\mathbb{1}\{T_n^{(m_0)}-T_{n-1}^{(m_0)}>t_{C_1}\}\ge 1 \,\middle|\, t\in N^{(m_0)}, (\xi_{m_0},Z_{m_0})=(x,z)\right), \end{aligned}\tag{33}$$

where t_{C_i} denotes the cache eviction time at the cache $i\in\{1,2\}$ under Che's approximation.

Hereafter, the symbol $\sum_{i_1<i_2}^{0,k-1}$ denotes the sum over all the couples $(i_1,i_2)\in\{0,\dots,k-1\}^2$ such that $i_1<i_2$. The following proposition holds.

PROPOSITION 4.1. *We have that*

$$p^{(t-x)}_{\mathrm{hit,Che,II}}(z,t_{C_1},t_{C_2})=0 \quad \textit{if } t_{C_2}\le t_{C_1}\tag{34}$$

and

$$\mathfrak{L}\le p^{(t-x)}_{\mathrm{hit,Che,II}}(z,t_{C_1},t_{C_2})\le \mathfrak{U} \quad \textit{if, for some integer } k\ge 1,\ kt_{C_1}<t_{C_2}\le (k+1)t_{C_1}.\tag{35}$$

Here

$$\begin{aligned} \mathfrak{L} &:= \mathrm{e}^{-\int_{t-x-t_{C_1}}^{t} h(s,z)\,\mathrm{d}s}\left(\int_{t-x-t_{C_2}}^{t-x-t_{C_1}} h(\tau,z)\mathrm{e}^{-\int_{\tau-t_{C_1}}^{\tau} h(s,z)\,\mathrm{d}s}\mathrm{d}\tau\right.\\ &\quad\left.-\sum_{i_1<i_2}^{0,k-1}\int_{b_{i_1}-x}^{b_{i_1+1}-x} h(\tau,z)\mathrm{e}^{-\int_{\tau-t_{C_1}}^{\tau} h(s,z)\,\mathrm{d}s}\mathrm{d}\tau\int_{b_{i_2}-x}^{b_{i_2+1}-x} h(\tau,z)\mathrm{e}^{-\int_{\tau-t_{C_1}}^{\tau} h(s,z)\,\mathrm{d}s}\mathrm{d}\tau\right), \end{aligned}$$

$$\mathfrak{U} := \mathrm{e}^{-\int_{t-x-t_{C_1}}^{t} h(s,z)\,\mathrm{d}s}\int_{t-x-t_{C_2}}^{t-x-t_{C_1}} h(\tau,z)\mathrm{e}^{-\int_{\tau-t_{C_1}}^{\tau} h(s,z)\,\mathrm{d}s}\mathrm{d}\tau$$

and

$$b_i := \frac{(k-i)(t-t_{C_2}) + i(t-t_{C_1})}{k}, \quad i = 0, \dots, k.$$

Proof. By (33) and the Slivnyak–Mecke theorem (see, e.g., Proposition 13.1.VII, p. 281, in [8]), we have

$$\begin{aligned}&p^{(t-x)}_{\mathrm{hit,Che,II}}(z, t_{C_1}, t_{C_2})\\ &\quad= \mathbb{P}\Big(N^{(m_0)}((t-t_{C_1}, t]) = 0,\\ &\qquad\qquad \sum_n \mathbb{1}_{(t-t_{C_2}, t-t_{C_1}]}(T_n^{(m_0)})\mathbb{1}\{T_n^{(m_0)} - T_{n-1}^{(m_0)} > t_{C_1}\} \geq 1 \,\Big|\, (\xi_{m_0}, Z_{m_0}) = (x,z)\Big),\end{aligned}$$

and this quantity is equal to zero if $t_{C_1} \geq t_{C_2}$, which proves (34). Otherwise, there exists an integer $k \geq 1$ such that $kt_{C_1} < t_{C_2} \leq (k+1)t_{C_1}$. We consider the partition of the set $(t-t_{C_2}, t-t_{C_1}]$ formed by the intervals $I_i := (b_i, b_{i+1}]$, $0 \leq i \leq k-1$, where the b_i's are defined in the statement, and we set $n_i^* := \min\{n :\ T_n^{(m_0)} > b_i\}$. Since, by construction $b_{i+1} - b_i \leq t_{C_1}$, for any $0 \leq i \leq k-1$, provided that $T_n^{(m_0)} \in I_{\bar{i}}$, for some $\bar{i} = 0, \dots, k$, and $T_n^{(m_0)} - T_{n-1}^{(m_0)} > t_{C_1}$, then necessarily $T_{n-1}^{(m_0)} \leq b_{\bar{i}}$. Therefore, setting $A := \{N^{(m_0)}((t-t_{C_1}, t]) = 0\}$ and

$$B_i := \{T^{(m_0)}_{n_i^*} \in I_i,\ N^{(m_0)}((T^{(m_0)}_{n_i^*} - t_{C_1}, b_i)) = 0\}, \quad i = 0, \dots, k-1, \tag{36}$$

we deduce

$$\begin{aligned}p^{(t-x)}_{\mathrm{hit,Che,II}}(z, t_{C_1}, t_{C_2}) &= \mathbb{P}\Big(A \cap \Big(\bigcup_{i=0}^{k-1} B_i\Big) \,\Big|\, (\xi_{m_0}, Z_{m_0}) = (x,z)\Big)\\ &= \mathbb{P}\Big(A \,\Big|\, (\xi_{m_0}, Z_{m_0}) = (x,z)\Big)\mathbb{P}\Big(\bigcup_{i=0}^{k-1} B_i \,\Big|\, (\xi_{m_0}, Z_{m_0}) = (x,z)\Big)\\ &= \mathrm{e}^{-\int_{t-x-t_{C_1}}^{t} h(s,z)\,\mathrm{d}s}\,\mathbb{P}\Big(\bigcup_{i=0}^{k-1} B_i \,\Big|\, (\xi_{m_0}, Z_{m_0}) = (x,z)\Big). \qquad (37)\end{aligned}$$

The Bonferroni inequality and the union bound yield

$$\begin{aligned}\sum_{i=0}^{k-1}\mathbb{P}\Big(B_i \,\Big|\, (\xi_{m_0}, Z_{m_0}) = (x,z)\Big) - \sum_{i_1<i_2}^{0,k-1}\mathbb{P}\Big(B_{i_1} \cap B_{i_2} \,|\, (\xi_{m_0}, Z_{m_0}) = (x,z)\Big)&\\ \leq \mathbb{P}\Big(\bigcup_{i=0}^{k-1} B_i \,\Big|\, (\xi_{m_0}, Z_{m_0}) = (x,z)\Big)&\\ \leq \sum_{i=0}^{k-1}\mathbb{P}\Big(B_i \,\Big|\, (\xi_{m_0}, Z_{m_0}) = (x,z)\Big).& \qquad (38)\end{aligned}$$

For any $i_1 < i_2$, $i_1, i_2 \in \{0, \dots, k-1\}$, we have

$$\begin{aligned}&\mathbb{P}\Big(B_{i_1} \cap B_{i_2} \,\Big|\, (\xi_{m_0}, Z_{m_0}) = (x,z)\Big)\\ &\quad= \Big(\mathbb{P}\Big(B_{i_2}, T_{n^*_{i_2}} - T_{n^*_{i_1}} > t_{C_1} \,\Big|\, B_{i_1},\, (\xi_{m_0}, Z_{m_0}) = (x,z)\Big)\\ &\qquad + \mathbb{P}\Big(B_{i_2}, T_{n^*_{i_2}} - T_{n^*_{i_1}} \leq t_{C_1} \,\Big|\, B_{i_1},\, (\xi_{m_0}, Z_{m_0}) = (x,z)\Big)\Big)\mathbb{P}\Big(B_{i_1} \,\Big|(\xi_{m_0}, Z_{m_0}){=}(x,z)\Big)\\ &\quad= \mathbb{P}\Big(B_{i_2}, T_{n^*_{i_2}} - T_{n^*_{i_1}} > t_{C_1} \,\Big|\, B_{i_1},\, (\xi_{m_0}, Z_{m_0}) = (x,z)\Big)\mathbb{P}\Big(B_{i_1} \,\Big|\, (\xi_{m_0}, Z_{m_0}) = (x,z)\Big),\end{aligned}$$

where the latter equality follows noticing that, given B_{i_1}, $\{B_{i_2}, T_{n^*_{i_2}} - T_{n^*_{i_1}} \leq t_{C_1}\} = \emptyset$. By the independence of the increments of the Poisson process we deduce

$$\begin{aligned}\mathbb{P}\left(B_{i_2}, T_{n^*_{i_2}} - T_{n^*_{i_1}} > t_{C_1} \,\middle|\, B_{i_1}, (\xi_{m_0}, Z_{m_0}) = (x,z)\right)\\ = \mathbb{P}\left(B_{i_2}, T_{n^*_{i_2}} - T_{n^*_{i_1}} > t_{C_1} \,\middle|\, (\xi_{m_0}, Z_{m_0}) = (x,z)\right),\end{aligned}$$

and so

$$\begin{aligned}\mathbb{P}\left(B_{i_1} \cap B_{i_2} \,\middle|\, (\xi_{m_0}, Z_{m_0}) = (x,z)\right) \leq \mathbb{P}\left(B_{i_1} \,\middle|\, (\xi_{m_0}, Z_{m_0}) = (x,z)\right)\\ \mathbb{P}\left(B_{i_2} \,\middle|\, (\xi_{m_0}, Z_{m_0}) = (x,z)\right).\end{aligned}$$

Consequently, by (37) and (38) we have the following bounds on the "hit" probability:

$$\begin{aligned}&\mathrm{e}^{-\int_{t-x_{m_0}-t_{C_1}}^{t} h(s,z)\,\mathrm{d}s} \left(\sum_{i=0}^{k-1} \mathbb{P}\left(B_i \,\middle|\, (\xi_{m_0}, Z_{m_0}) = (x,z)\right)\right.\\ &\qquad \left. - \sum_{i_1<i_2}^{0,k-1} \mathbb{P}\left(B_{i_1} \mid (\xi_{m_0}, Z_{m_0}) = (x,z)\right) \mathbb{P}\left(B_{i_2} \mid (\xi_{m_0}, Z_{m_0}) = (x,z)\right)\right)\\ &\leq p^{(t-x)}_{\mathrm{hit,Che,II}}(z, t_{C_1}, t_{C_2})\\ &\leq \mathrm{e}^{-\int_{t-x-t_{C_1}}^{t} h(s,z)\,\mathrm{d}s} \sum_{i=0}^{k-1} \mathbb{P}\Big(B_i \,\Big|\, (\xi_{m_0}, Z_{m_0}) = (x,z)\Big).\end{aligned} \tag{39}$$

Relation (35) follows by (39) and the following computation:

$$\begin{aligned}&\mathbb{P}\Big(B_i \,\Big|\, (\xi_{m_0}, Z_{m_0}) = (x,z)\Big)\\ &= \int_{b_i}^{b_{i+1}} \mathbb{P}(N^{(m_0)}((\tau - t_{C_1}, b_i)) = 0 \mid (\xi_{m_0}, Z_{m_0}) = (x,z))\, \mathbb{P}_{T^{(m_0)}_{n_i{}^*} \mid (\xi_{m_0}, Z_{m_0}) = (x,z)}(\mathrm{d}\tau)\\ &= \int_{b_i}^{b_{i+1}} \mathrm{e}^{-\int_{\tau-t_{C_1}}^{b_i} h(s-x,z)\,\mathrm{d}s}\, \mathbb{P}_{T^{(m_0)}_{n_i{}^*} \mid (\xi_{m_0}, Z_{m_0}) = (x,z)}(\mathrm{d}\tau)\\ &= \int_{b_i}^{b_{i+1}} \mathrm{e}^{-\int_{\tau-t_{C_1}}^{b_i} h(s-x,z)\,\mathrm{d}s}\, h(\tau - x, z) \mathrm{e}^{-\int_{b_i}^{\tau} h(s-x,z)\,\mathrm{d}s} \mathrm{d}\tau\\ &= \int_{b_i-x}^{b_{i+1}-x} h(\tau, z) \mathrm{e}^{-\int_{\tau-t_{C_1}}^{\tau} h(s,z)\,\mathrm{d}s} \mathrm{d}\tau. \qquad \square\end{aligned}$$

Remark 4.2. Proposition 4.1 provides the exact value of the "hit" probability when $t_{C_2} \leq 2t_{C_1}$. As is clear from the proof, one might exactly compute the "hit" probability even when $t_{C_2} > 2t_{C_1}$ by applying the inclusion-exclusion formula. However, the resulting computational cost would be very high since one has to compute the probability of any intersection of the events B_i defined by (36). In conclusion, we can say that any computationally efficient approach to the performance analysis of a tandem network of caches must resort to some extra approximations (in addition to Che's approximation), which affect inevitably the accuracy of the method.

Remark 4.3. In principle, the approach proposed in this section can be generalized to networks of caches with a feed-forward structure, i.e., roughly speaking, to networks of caches in which the caches are "ordered"(in some sense) and any content request follows a path that traverses caches in "increasing order." Typical examples are networks of caches with a tree structure, where a request for a generic content follows a path in the network which starts from a cache placed on a leaf (belonging conventionally to level 1 of the tree), it is directed toward the cache placed at the root (belonging conventionally to the level K of the tree) and stops as soon as a "hit"is produced. Note that an arriving request for the content m_0 can produce a "hit" at a cache located at the kth level only if it has been missed at caches located at the ith level for any $i \leq k-1$. So, applying similar arguments as in (33), one may obtain a formal expression for the probability that there is a "hit" at a cache located at the kth level of the tree. However, the numerical evaluation of this probability becomes more and more prohibitive as the level grows, i.e., as k increases. As for the case of tandem networks of caches, any computationally efficient approach must rely on some additional approximations (in addition to Che's approximation) which reduce the accuracy of the method. Recently, several approximations have been proposed [1, 14, 24, 30, 34]. The accuracy of such approximations varies significantly from scenario to scenario and can be evaluated experimentally by comparing analytical predictions against Monte Carlo simulations [1, 14, 24, 30, 34].

5. Numerical illustrations. As shown by several recent experimental works, many video contents (such as YouTube contents) exhibits few typical normalized temporal popularity profiles, each profile corresponding to a large class of contents with similar characteristics (e.g., contents in the same YouTube category) [6]. Hence, restricting the analysis to a single class m of contents, we may assume that (i) Z_m represents the demand volume, i.e., the total number of requests it typically attracts; (ii) all contents of the class exhibit the same normalized popularity profile. This justifies the choice of an SNM with an multiplicative popularity profile such as (2).

Recall that for this model the function g is given by (4). Assuming (3), by (27) and (30) (with θ in place of t_C) we easily have

$$p_{\mathrm{hit,Che}}(\theta) = 1 - (\mathbb{E}[Z_1])^{-1} \int_0^\infty h(u) \phi'_{Z_1} \left(- \int_{u-\theta}^{u} h(s)\, \mathrm{d}s \right) \mathrm{d}u, \tag{40}$$

where ϕ'_{Z_1} is the first order derivative of ϕ_{Z_1}. Relations (4) and (40) provide a computationally efficient tool to estimate the "hit" probability, under Che's approximation, of LRU caches under the SNM. Indeed, we may estimate t_C by numerically inverting (4) and using the relation $t_C = g^{-1}(C/\lambda)$. Replacing θ in (40) with such estimate of t_C, we finally have an estimate of the "hit" probability under Che's approximation.

We now assess the accuracy of the Che approximation for the evaluation of the "hit" probability by describing some numerical results. We suppose that the arrival rate of new contents λ is equal to 100,000 units per day; we assume that the demand volume Z_1 follows a Pareto distribution with probability density $f_{Z_1}(z) = \alpha a^\alpha / z^{1+\alpha}, z \geq a > 0$, $\alpha > 1$, and mean $\mathbb{E}[Z_1] = \frac{\alpha a}{\alpha - 1} = 3$ (we refer the reader to [15] and [22] for a practical justification on the choice of a Pareto distribution); we consider a multiplicative popularity profile of the form (2) with $h(t) := \frac{1}{L} \mathbb{1}\{0 \leq t \leq L\}$, where the parameter L has to be interpreted as the content life-span.

Figures 1 and 2 report the "hit" probability, as predicted by Che's approximation, vs the cache size for different values of the exponent α and the content life span L, respectively. For each estimate, the figures show also the interval in which the exact

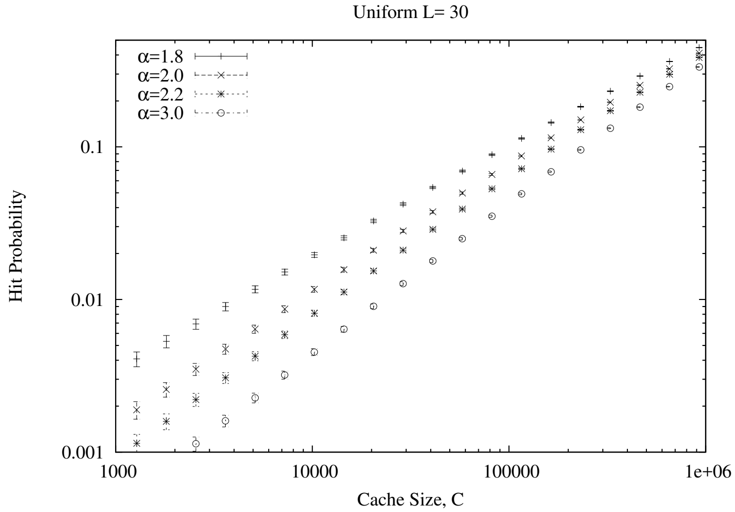

FIG. 1. *p_{hit} versus cache size for different values of the exponent $\alpha > 1$ and content life span $L = 30$.*

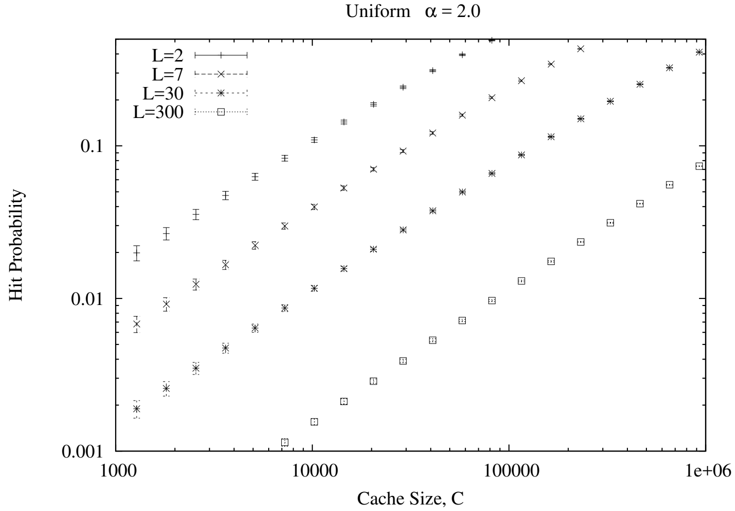

FIG. 2. *p_{hit} versus cache size for different values of the content life span L and exponent $\alpha = 2$.*

value of the "hit" probability falls as given by Proposition 3.14. All computations have been carried out while guaranteeing relative numerical errors smaller than 10^{-2}. Some selected results are additionally reported in Table 1. Note that in all cases of practical relevance (i.e., for values of the "hit" probability exceeding 10^{-2}) Che's approximation leads to negligible errors. The surprisingly good degree of accuracy entailed by Che's approximation, which has been already experimentally (i.e., against simulations) observed by several authors [15, 24], is now confirmed even for the SNM.

Further numerical results providing useful insights on the cache performance can be found in [22].

Table 1
Numerical values for the Che approximation of the "hit" probability ($p_{hit,Che}$) and for the lower ($\underline{p_{hit}}$) and the upper ($\overline{p_{hit}}$) bounds of the true "hit" probability.

α	L	C	$p_{\text{hit,Che}}$	$\underline{p_{\text{hit}}}$	$\overline{p_{\text{hit}}}$	C	$p_{\text{hit,Che}}$	$\underline{p_{\text{hit}}}$	$\overline{p_{\text{hit}}}$
1.8	30	10240	0.019596	0.018880	0.020313	163840	0.144328	0.143126	0.145529
2.0	2	10240	0.109252	0.105353	0.113151	163840	0.671657	0.669498	0.673815
2.0	7	10240	0.039790	0.038232	0.041348	163840	0.343061	0.340319	0.345802
2.0	30	10240	0.011657	0.011158	0.012156	163840	0.114597	0.113516	0.115677
2.0	300	10240	0.001555	0.001480	0.001629	163840	0.017497	0.017305	0.017688
2.2	30	10240	0.008125	0.007747	0.008504	163840	0.096641	0.095651	0.097630
3.0	30	10240	0.004524	0.004293	0.004755	163840	0.068667	0.067871	0.069464

REFERENCES

[1] G. Bianchi et al., *Check before storing: What is the performance price of content integrity verification in LRU caching?* ACM Comput. Comm. Rev., 43 (2013), pp. 59–67.

[2] L. Bondesson, *Shot noise processes and shot noise distributions*, In Encyclopedia of Statistical Sciences, Wiley, New York, 1988, pp. 448–452.

[3] C. Bordenave and G. L. Torrisi, *Monte Carlo methods for sensitivity analysis of Poisson driven stochastic systems, and applications*. Adv. Appl. Probab., 40 (2008), pp. 293–320.

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

[5] E. Coffman and P. Denning, *Operating Systems Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1974.

[6] R. Crane and D. Sornette, *Robust dynamic classes revealed by measuring the response function of a social system*. Proc. Natl. Acad. Sci., 105 (2008), pp. 15649–15653.

[7] D. J. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes, Vol* I, Springer, New York, 2003.

[8] D. J. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes, Vol.* II, Springer, New York, 2008.

[9] A. Dan and D. Towsley, *An approximate analysis of the LRU and FIFO buffer replacement schemes*. SIGMETRICS Perform. Eval. Rev., 18 (1990), pp. 143–152.

[10] A. Dembo and O. Zeitouni, *Large Deviation Techniques and Applications*, Springer, New York, 1998.

[11] P. M. Deshpande, K. Ramasamy, A. Shukla, and J. F. Naughton, *Caching multidimensional queries using chunks*, in Proceedings of SIGMOD, 98, ACM, 1998.

[12] N. G. Duffield and W. Whitt, *Large deviations of inverse processes with nonlinear scalings*. Ann. Appl. Probab., 8 (1998), pp. 995–1026.

[13] K. Duffy and G. L. Torrisi, *Sample path large deviations of Poisson shot noise with heavy tail semi-exponential distributions*, J. Appl. Probab. 48 (2011), pp. 688–698.

[14] N. C. Fofack et al., *Performance evaluation of hierarchical TTL-based cache networks*, Comput. Networks, 65 (2014), pp. 212–231.

[15] C. Fricker, P. Robert, and J. Roberts, *A versatile and accurate approximation for LRU cache performance*, in Proceedings of ITC, Krakow, Poland, 2012, pp. 1–8.

[16] A. Ganesh, C. Macci, and G. L. Torrisi, *Sample path large deviations principles for Poisson shot noise processes, and applications*, Electron. J. Probab., 10 (2005), pp. 1026–1043.

[17] A. Ganesh and G. L. Torrisi, *A class of risk processes with delayed claims: Ruin probability estimates under heavy tail conditions*, J. Appl. Probab., 43 (2006), pp. 916–926.

[18] A. Ganesh, C. Macci, and G. L. Torrisi, *A class of risk processes with reserve-dependent premium rate: Sample path large deviations and importance sampling*, Queueing Syst., 55 (2007), pp. 83–94.

[19] V. Jacobson et al., *Networking named content*, in Proceedings of CoNEXT, Rome, Italy, ACM, 2009.

[20] W. Jiang et al., *Orchestrating massively distributed CDNs*, in Proceedings of CoNEXT, Nice, France, ACM, 2012.

[21] K. Kylakoski and J. Virtamo, *Cache replacement algorithms for the renewal arrival model*, in Proceedings of the 14th Nordic Teletraffic Seminar, Copenhagen, Denmark, 1998, pp. 139–148.

[22] E. Leonardi and G. L. Torrisi, *Least recently used caches under the shot noise model*, in Proceedings of INFOCOM, Hong Kong, IEEE, 2015, pp. 2281–2289.

[23] S. B. Lowen and M. C. Teich, *Power-law shot noise*, IEEE Trans. Inform. Theory, 36 (1990), pp. 1302–1318.

[24] V. Martina et al., *A unified approach to the performance analysis of caching systems*, in Proceedings of INFOCOM, Toronto, IEEE, 2014, pp. 2040–2048.

[25] J. Moller and G. L. Torrisi, *Generalised shot noise Cox processes*, Adv. Appl. Probab., 37 (2005), pp. 47–74.

[26] I. Nourdin and G. Peccati, *Normal Approximation with Malliavin Calculus*, Cambridge University Press, Cambridge, UK, 2012.

[27] G. Peccati, J. L. Solé, M. S. Taqqu, and F. Utzet, *Stein's method and normal approximation of Poisson functionals*, Ann. Probab., 38 (2010), pp. 443–478.

[28] M. Penrose, *Random Geometric Graphs*, Oxford University Press, New York, 2004.

[29] I. Psaras, W. K. Chai, and G. Pavlou, *Probabilistic methods in network caching for information-centric networks*, presented at the ICN Workshop on Information-Centric Networking, 2012.

[30] E. J. Rosensweig et al., *Approximate models for general cache networks*, in Proceedings of INFOCOM, San Diego, IEEE, 2010, pp. 1–9.

[31] G. Stabile and G. L. Torrisi, *Large deviations of Poisson shot noise processes, under heavy tail semi-exponential conditions*, Statist. Probab. Lett., 80 (2010), pp. 1200–1209.

[32] G. L. Torrisi, *Simulating the ruin probability of risk processes with delay in claim settlement*, Stoch. Proc. Appl., 112 (2004), pp. 225–244.

[33] S. Traverso et al., *Temporal locality in today's content caching: Why it matters and how to model it*, ACM Comput. Comm. Rev., 43 (2013), pp. 5–12.

[34] S. Traverso et al., *Unravelling the impact of temporal and geographical locality in content caching systems*, IEEE Trans. Multimedia, 17 (2015), pp. 1839–1854.