A Bound on the Moments of a TTL-based Cache's Miss-Process

Daniel S. Berger*, Florin Ciucu†
*Distributed Computer Systems (DISCO) Lab, University of Kaiserslautern, Germany
†T-Labs/ FG INET, TU Berlin, Germany

Abstract—**Due to the omnipresence of caching in modern computing and networks, characterizing the performance of caches is an important aspect of system performance and scalability analysis. While even analytical models for single caches proved to be challenging, nowadays, many configurations already encompass hierarchies of caches. Time-to-live (TTL) based caches have recently been proposed as a general approach to simplify performance analysis of caching networks. Previous work has used approximation and recursion techniques to characterize the inter-miss process which is essential to study this setting. This work introduces an appropriate mathematical abstraction of the cache model as a martingale representation. We derive explicit bounds on all moments of the inter-miss time of a TTL-based cache and report preliminary simulation results. Our approach does not rely on recursion techniques but is still general enough to capture different previously introduced TTL-based caching models.**

I. Introduction

Characterizing the performance of caches is an important aspect of system performance analysis given the omnipresence of caching in modern computer systems. Caches are deployed to transparently store a subset of frequently accessed data items close to customers and processing units in order to decrease access delay while reducing load to a central instance. Computer networks present particular application cases with today's scalability of domain name and content delivery resolution being critically dependent on caching technology.

Despite its relevance, caching was exposed to be a hard problem from a modeling perspective. The most commonly studied policy, the Least-Recently-Used (LRU) policy, was first analytically studied in [1]. Due to the high computational complexity of this exact result, [2] suggested an approximation and today, approximations for LRU caches are established methodology [3]. Similar cases can be made for other popular policies like the First-In-First-Out (FIFO) and the Random Replacement (RR) policy [4], or the Move-to-Front eviction strategy [5]. With the proposal of integrating caching more tightly with a future Internet's architecture [6], [7] networks of caches need to be analyzed to assess these system design proposals. Characterizing performance metrics in this setting increases modeling difficulty even for small networks [8] and further work on approximations appears to be required [9], [10].

Time-to-live (TTL) based caches have recently gained attraction due to their reported capability to mimic metrics of interest for LRU, FIFO and random cache models [10]. Fofack et al. [10] propose their TTL-based model of networks of caches as a generalization which appears simpler to analyze than networks of LRU or FIFO caches. Additionally, TTL-based caches exhibit great relevance for real-world applications such as the domain name system or content delivery [11]–[13].

TTL-based caching evicts an object from the cache after a corresponding timer has expired. After eviction, the next request constitutes a miss, the object enters the cache again, and the TTL timer is reset. Similar to the *independent reference model* used to analyze LRU, FIFO, or RR caches [1], TTL-based cache models assume inter-query times to be given as a series of identically independent (iid) random variables.

For a constant TTL, Jung et al. [12] find that "hit and miss rates have a complicated distribution" which, however, simplifies to a single point mass when considering the limit of time-intervals $(0, t]$ where $t \to \infty$. Under this assumption they derive limit results on the average hit and miss rates. In [14] this limiting hit rate result is generalized to include delays in download time and consistency measures. Recently, Fofack et al. [10] attended to the general case of TTLs being assumed to be iid random variables. This is important in order to model characteristics of other cache models as their representation may require TTLs to be random variables. The authors give exact results on the hit rate, miss rate, and cache occupancy, and derive a recurrent integral equation for the inter-miss time distribution. In an approximated form this result can be applied to the analysis of caching networks for exponential TTLs. However, delivering an exact equation comes at the cost of complexity which makes both iteration and the case of general arrival and TTL distributions hard to analyze.

Previous work exploited mainly renewal properties of both arrival and miss processes. In this work we describe a new mathematical representation of a TTL-based cache which makes martingale results applicable. Given a martingale representation, we are able to draw results from this mathematical framework, in particular Wald's fundamental identity [15], to develop upper and lower bounds on all moments of the inter-miss process of a TTL-based cache.

The inter-miss process is particularly relevant for a networks-of-caches scenario. For this application, the miss-process of a cache in layer n represents the arrival process of a cache in layer $n + 1$, and thus a characterization of the inter-miss process is requisite to this analysis. Additionally, the moments of the inter-miss process characterize the traffic flow exiting some cache which, for a cache deployment, may be used to bound the amount of traffic reduction obtained by the cache or the exit-flow's standard deviation.

Martingales have been considered before in the analysis of computer networks. For example [16] and [17] adopted a martingale to derive results in queueing theory, or [18] used a martingale in the domain of effective bandwidths. Recently, martingale theory has been applied to the domain of network calculus [19]. We seem to be the first to suggest a martingale for the analysis of cache networks.

The rest of this paper is structured as follows. The next section, Section II, introduces and formalizes both cache models. In Section III we give our main result, the bound on the miss-process's moment generating function. Section IV contains an application example to Poisson arrivals and exponential TTLs and we assess the tightness of our bounds briefly with simulation results. We conclude in Section V.

II. Cache Model

Similar to the independent reference model [1], we consider a given arbitrary data item and query arrivals which are independent and identically (iid) distributed. The simplification of iid requests may be conservative [12] as burst-arrivals to the same item would yield a higher hit rate. This issue applies to the lower bound presented here only, and Jung et al. find that under this assumption, the hit rate is still predicted well [12]. Additionally, our approach does not depend on a specific distribution, and, for example, a long-tailed Weibull distribution which is reported to provide a good fit for TCP session arrivals [20] can be plugged into our bounds.

From a user's perspective, obtaining an answer from a cache may be considered beneficial compared to waiting longer for an answer from a central entity, encouraging further requests. We do not consider any such feedback to the user.

In the literature, two different descriptions of TTL-based caching have appeared. The two models differ in when the TTL is reset; informally, they can be outlined as:

- $\mathcal{A}$.) renew the TTL after a miss (and a miss occurs for the first request to the object after the TTL has expired)
- $\mathcal{B}$.) renew the TTL at each request to an object, being either a miss or a hit (and a miss occurs once any request took longer than the current TTL)

The former model, denoted by $\mathcal{A}$.), is used in many real-world applications (e.g. DNS, caching of web objects) and has been assumed in both [12] and [14] to develop hit rate equations. Additionally, this model guarantees weak consistency [13] as for a single cache, any object is at least as recent as the maximal TTL value.

The second model, denoted by $\mathcal{B}$.), was introduced recently [10] due to its favorable mathematical properties: each request constitutes a regeneration point. In particular, any TTL spans at most the time window of one inter-request random variable so that miss probabilities do not depend on a sum of inter-request random variables. According to renewal theory, it is then sufficient to analyze properties of the first request window to characterize the full process.

We formalize the caching models as two stochastic processes. Assume $\{X_i\}_{i\geq 1}$ (the inter-arrivals) and $\{T_i\}_{i\geq 1}$ (the TTLs) to be independent series of real-valued iid random variables with corresponding CDFs $F(z)$ and $G(t)$ and densities $f(z)$, $g(t)$, respectively.

Definition 1 (Caching Models)**.**

$\mathcal{A}$.) TTL renewal after miss

$$Y_n := X_1 + \cdots + X_n \tag{1}$$

$$\tau := \inf\{i : \sum_{k=1}^{i} X_k > T_1\} \tag{2}$$

$\mathcal{B}$.) TTL renewal at each request

$$Y_n := X_1 + \cdots + X_n \tag{3}$$

$$\tau := \inf\{i : X_i > T_i\} \tag{4}$$

Because the overall miss process defines a renewal process in both cases, the stopping time of the first miss characterizes the miss process. In particular, the inter-miss distribution is characterized by the distribution of Y_τ in both cases. The second model, model $\mathcal{B}$.), is simpler because τ does not depend on the sum $\sum X_k$ of random variables.

Observe that τ is a stopping time in both cases if the corresponding filtration includes X_is and T_is, for example, let $\mathcal{F}_t := \sigma((X_1, T_1), \ldots, (X_t, T_t))$.

At its surface, this problem resembles a first hitting time problem of a continuous time random walk (CTRW). However, the fact that we are interested in Y_n (the actual time of the first miss) and not simply τ (the index of the first miss) determines a dependency between the time and the jump altitude of the random walk which prevents the usual decoupling (decoupled CTRW) approach [21]. Likewise, consider the usual conditioning approach $\mathbb{P}(Y_\tau \leq y) = \mathbb{P}(\sum_{i=1}^{\tau} X_i \leq y) = \sum_{t=1}^{\infty} \mathbb{P}(\sum_{i=1}^{\tau} X_i \leq y | \tau = t)\mathbb{P}(\tau = t)$. Although $\mathbb{P}(\tau = t)$ is straightforward to compute, the dependency of X_is and τ prevents any decoupling of the conditional $\mathbb{P}(\sum_{i=1}^{\tau} X_i \leq y | \tau = t)$.

For the reason of decoupling we propose a martingale abstraction to the caching model in the next section. This work is general and our bounds apply to both models; however, for simplicity of demonstration, the application example and simulation results in Section IV are limited to three particular configurations of model $\mathcal{B}$.) .

III. Bounds on the Moment Generating Function

We define a martingale with respect to the cache process Y_n of both models and use results on stopped martingales to derive our bounds in this section. The two models will only differ by their respective stopping times. For this reason, both upper and lower bound are uniform for both models and only when evaluating the bounds for specific parameters $F(z)$ and $G(t)$ will the difference in τ come into play.

Martingales are a class of stochastic processes, such as Markov processes or stationary processes. At some time t in a martingale model, the expected value for time $t+1$ is the same as the present value. In particular, the process's next expected value is still the present value given knowledge about all past events. We omit an introduction to martingales and stopping times as formal knowledge about this theory is not required to follow our argumentation[1]. A martingale model helps to exploit the observation that τ is a stopping time. Additionally, martingale properties may yield high-level insight into the problem.

Definition 2 (A Martingale for Cache Models)**.**

$$Z_n := \frac{e^{\omega Y_n}}{\left(\mathbb{E}\left[e^{\omega X_1}\right]\right)^n} \tag{5}$$

Checking that this definition fulfills the martingale definition is straightforward given that the arrival process is a renewal process and thus in every step the moment generating function (MGF) of the arrivals $\mathbb{E}\left[e^{\omega X_i}\right]$ is equal to $\mathbb{E}\left[e^{\omega X_1}\right]$. We use the common symbol $\phi(\omega)$ to denote the MGF of the inter-arrival process. Then, the martingale definition becomes $Z_n = e^{\omega Y_n} \left(\phi(\omega)\right)^{-n}$.

This martingale[2] is due to Wald [15] and the same work also contributed an identity which is instrumental to analyze Y_τ if τ is a random walk barrier, i.e. the first time Y_n leaves some interval (a, b). For this bounded stopping time, Wald proves that $Z_n = 1$ which is instrumental to approximate the distributions $\mathbb{P}(Y_\tau \leq -a)$ and $\mathbb{P}(Y_\tau \geq b)$ [22], [23]. Our stopping time, however, is more general, and in particular not bounded. We employ standard martingale theory to obtain an appropriate generalization which requires only practical conditions as shown in Section IV.

Proposition 1 (Optional Stopping Identity)**.**
Assuming $\tau < \infty$ almost surely, and Z_n uniformly integrable, then it holds:

$$\mathbb{E}\left[Z_\tau\right] = \mathbb{E}\left[Z_1\right] = 1 \tag{6}$$

Proof: This is an application of a special form of the optional stopping theorem which can, for example, be found in section 12.5 on optional stopping in [22] ■

The assumption of an almost surely finite stopping time is feasible for most realistic caching configurations. In particular, given space constraints of a cache and at least a positive probability of arrivals, the number of arrivals between misses intuitively stays finite with probability one. The second condition restricts the weight of the tails of Z_n's distribution *uniformly* over the whole family, i.e., independent of n.

Using Proposition 1 it is now possible to derive upper and lower bounds on all moments of the miss process:

Lemma 1 (Lower Bound on Y_τ)**.** *For $p \in (1, \infty)$, and under the assumptions of Proposition 1:*

$$\mathbb{E}\left[e^{\theta Y_\tau}\right] \geq \left(\sum_{k=1}^{\infty} \left(\phi(\theta/p)\right)^{\frac{-kp}{p-1}} \, P(\tau = k)\right)^{1-p} \tag{7}$$

Proof: Apply Hölder's inequality to split the martingale from Definition 2. Given $p, q > 1$ and $p^{-1} + q^{-1} = 1$:

$$1 = \mathbb{E}\left[Z_\tau\right] \leq \left(\mathbb{E}\left[\left|(e^{\omega Y_\tau})^p\right|\right]\right)^{\frac{1}{p}} \left(\mathbb{E}\left[\left|(\phi(\omega)^{-\tau})^q\right|\right]\right)^{\frac{1}{q}} \tag{8}$$

And thus, by raising to the power of p and setting $\omega := \theta/p$.

$$\left(\mathbb{E}\left[(e^{\omega Y_\tau})^p\right]\right)^{\frac{1}{p}} \geq \left(\mathbb{E}\left[(\phi(\omega)^{-\tau})^q\right]\right)^{\frac{-1}{q}} \tag{9}$$

$$\mathbb{E}\left[e^{\theta Y_\tau}\right] \geq \left(\mathbb{E}\left[(\phi(\theta/p)^{-\tau})^q\right]\right)^{\frac{-p}{q}} \tag{10}$$

Finally, we set $q := \frac{p}{p-1}$ to obtain a single optimization parameter fulfilling the Hölder conjugate constraint and replace the expectation by its definition. This is possible because $\Phi(\theta/p)$ is constant and thus independent of τ. ■

Observe that although the identity in Proposition 1 decouples the dependencies of Y_n, the result is still insufficient for straightforward evaluation. In textbook examples [22], [23] the remaining dependency is usually broken by finding a ω such that $\Phi(\omega) = \mathbb{E}\left[e^{\omega X}\right] = 1$. However, already for Poisson arrivals there is no non-trivial ω which fulfills this property. Moreover, even given existence of such an ω, setting ω to a fixed value prevents the optimization of p to obtain a tight bound on $\mathbb{E}\left[e^{\theta Y_\tau}\right]$: θ needs to be a real-valued parameter and was obtained by setting $\omega := \theta/p$.

In our approach, the dependency between the stopping time and the cumulative arrivals remains and we solve this issue by relaxing the problem to stochastic bounds instead of an exact analysis.

Slightly extending the technique from Lemma 1, we derive an upper bound on the moments of the miss process in Lemma 2.

Lemma 2 (Upper Bound on Y_τ)**.** *For $p \in (1, \infty)$, and under the assumptions of Proposition 1:*

$$\mathbb{E}\left[e^{\theta Y_\tau}\right] \leq \left(\sum_{k=1}^{\infty} \left(\phi(\theta p)\right)^{\frac{k}{p-1}} \, P(\tau = k)\right)^{\frac{p-1}{p}} \tag{11}$$

Proof: We use a simple reverse of Hölder's inequality (proof in the appendix, Lemma 3) to derive:

$$1 = \mathbb{E}\left[Z_\tau\right] \geq \left(\mathbb{E}\left[\left|(e^{\omega Y_\tau})^{\frac{1}{p}}\right|\right]\right)^p \left(\mathbb{E}\left[\left|(\phi(\omega))^{\frac{-\tau(-1)}{p-1}}\right|\right]\right)^{-(p-1)} \tag{12}$$

Similar to before, taking the p-th root and setting $\omega := \theta p$ gives a bound for the MGF of Y_τ:

$$\left(\mathbb{E}\left[e^{\frac{\omega}{p} Y_\tau}\right]\right)^p \leq \left(\mathbb{E}\left[\left|(\phi(\omega))^{\frac{\tau}{p-1}}\right|\right]\right)^{(p-1)} \tag{13}$$

$$\mathbb{E}\left[e^{\theta Y_\tau}\right] \leq \left(\mathbb{E}\left[\left|(\phi(\theta p))^{\frac{\tau}{p-1}}\right|\right]\right)^{\frac{p-1}{p}} \tag{14}$$

■

Observe that we need $\mathbb{P}(\tau = k)$ in both lemmas. This probability mass function is actually convenient to derive in many cases due to the cache process's renewal properties. For example for cache model $\mathcal{B}$.), the stopping time distribution is

[1] An interested reader may consider textbooks about stochastic processes for an introduction. As an example, we consider chapter 12.5 in [22] about optional stopping to be particularly relevant.

[2] The martingale in Definition 2 is often called "Wald's Martingale" and the corresponding identity is "Wald's fundamental identity" [15], [22], [23]

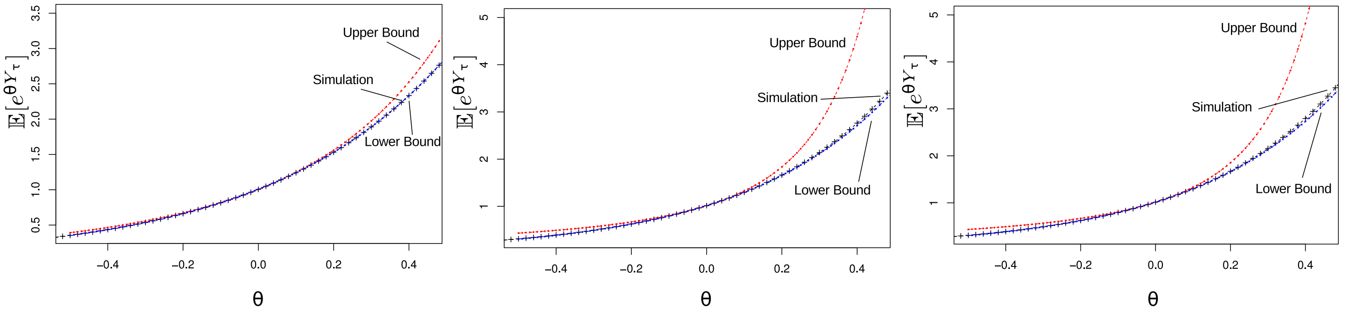

Fig. 1. Simulation vs. upper and lower bounds for different cache configurations. In all cases, we see that the bounds work well for $|\theta|$ near to zero, which is the range most important to derive the miss process' moments. The lower bound is tight for even a larger range of θ. Left: for a somewhat typical case of an arrival rate ($\lambda = 9$) which is considerably larger than the deterministic TTL ($T = 2$), the bounds work particularly well. This is the case for all configurations with $lambda$ considerably greater than T. For corner cases, where the two are close together, the tightness degrades: the central plot shows $\lambda = 2.1$, and the left plots shows $\lambda = 2.01$ with $T = 2$ in both cases.

geometric:

$$\mathbb{P}(\tau = k) = (P(X_1 \leq T_1))^{k-1} P(X_1 > T_1) \tag{15}$$

$$= \left(\int_0^\infty F(t) dG(t)\right)^{k-1} \left(1 - \int_0^\infty F(t) dG(t)\right) \tag{16}$$

Together with Lemmas 1 and 2, this gives us a simple way to access principal properties of the miss process, i.e., its moments, via the corresponding moment generating function (MGF).

IV. Application Cases and Simulation Results

In this section we seek to show the applicability of the results obtained in the previous section. For this reason we focus on the case of Poisson arrivals. More realistic inter-arrival distributions like the Weibull distribution [20] also possess an MGF transform, and thus are candidates which can be considered with our approach. Nevertheless, for the example of Weibull distributed inter-arrivals, the analysis is more involved as the MGF is usually not available in closed form, although some progress has been obtained for special cases of the Weibull MGF [24].

To be able to apply Proposition 1 which is a necessity for the bounds, we need to fulfill the corresponding conditions. Firstly, we shall prove the less common of the two conditions, uniform integrability, for the case of Poisson arrivals.

Proposition 2. *For Poisson arrivals, the caching martingale Z_n is uniformly integrable for an ϵ-environment around zero:*

$$\mathbb{P}(Z_n 1_{\{Z_n > a\}}) \to 0 \text{ as } n \to \infty \tag{17}$$

Proof: We first employ Hölder's inequality for $p, q = 2$ and then Markov's inequality:

$$\mathbb{P}(Z_n 1_{\{Z_n > a\}}) \tag{18}$$

$$\leq \mathbb{E}\left[(Z_n)^2\right]^{\frac{1}{2}} \mathbb{E}\left[(1_{\{Z_n > a\}})^2\right]^{\frac{1}{2}} \tag{19}$$

$$\leq \mathbb{E}\left[(Z_n)^2\right]^{\frac{1}{2}} \frac{\mathbb{E}\left[Z_n\right]}{a} \tag{20}$$

Because $\mathbb{E}\left[Z_N\right] = 1$ and because $\mathbb{E}\left[(Z_n)^2\right] = \mathbb{E}\left[e^{2\omega Y_n} / (\Phi(\omega))^{-2n}\right]$, it is always possible to find $\omega' < \omega$ with $|\omega'| > 0$ so that $\mathbb{E}\left[(Z_n)^2\right]$ is bounded. ■

Instead of the last argument, it is also possible to use the Erlang density to algebraically verify the claim of the proposition. This is omitted for brevity.

Next, we derive bounds for two exemplar caching configurations, using both caching models, as well as both deterministic and random TTLs.

Example 1 (Poisson Arrivals, Det. TTLs, and Model $\mathcal{A}$.))**.** *Assume caching model $\mathcal{A}$.), $X_i \sim exp(\lambda)$, and $T_i = T \in \mathbb{R}^+$. Then $p > 1$ can be optimized on the bounds:*

$$\mathbb{E}\left[e^{\theta Y_\tau}\right] \geq c^p e^{(1-p)(c^{\frac{-p}{p-1}} - 1)\lambda T} \tag{21}$$

$$\mathbb{E}\left[e^{\theta Y_\tau}\right] \leq d^{\frac{1}{p}} e^{\frac{p-1}{p}(d^{\frac{1}{p-1}} - 1)\lambda T} \tag{22}$$

where $c := \frac{\lambda p}{\lambda p - \theta}$, and $d := \frac{\lambda}{\lambda - \theta p}$.

Proof: As a first step, we derive the probability mass function for the stopping time. It is clear that $\mathbb{P}(\tau < 1) = 0$, and $\mathbb{P}(\tau = 1) = \mathbb{P}(X_1 > T_1) = e^{-\lambda T}$. For $\mathbb{P}(\tau > 1)$ we use 1. a conditioning trick known from random walk theory, 2. the Renewal properties of the arrivals, and 3. the property that the sum of $k-1$ exponential random variables is Erlang distributed with shape parameter $k - 1$ and the same rate.

$$\mathbb{P}(\tau = k) = \mathbb{P}(\sum_{i=1}^{k-1} X_i \leq T, \sum_{i=1}^{k} X_i > T) \tag{23}$$

$$= \int_0^T \mathbb{P}(z + X_k > T | \sum_{i=1}^{k-1} X_i = z) \mathbb{P}(\sum_{i=1}^{k-1} X_i = z) dz \tag{24}$$

$$= \int_0^T \left(e^{-\lambda(T-z)}\right) \frac{\lambda^{k-1} z^{k-2} e^{-\lambda z}}{(k-2)!} dz \tag{25}$$

$$= \frac{\lambda^{k-1} e^{-\lambda T}}{(k-2)!} \frac{T^{k-1}}{k-1} = \frac{\lambda^{k-1} e^{-\lambda T} T^{k-1}}{(k-1)!} \tag{26}$$

Observe that this formula is general and includes the case of $\tau = 1$.

It holds that $\mathbb{P}(\tau < \infty) = 1$ because

$$\mathbb{P}(\tau < \infty) = \sum_{k=1}^{\infty} \mathbb{P}(\tau = k) = \sum_{k=1}^{\infty} \frac{(\lambda T)^{k-1}}{(k-1)!} e^{-\lambda T} \tag{27}$$

$$= e^{-\lambda T + \lambda T} = 1 \tag{28}$$

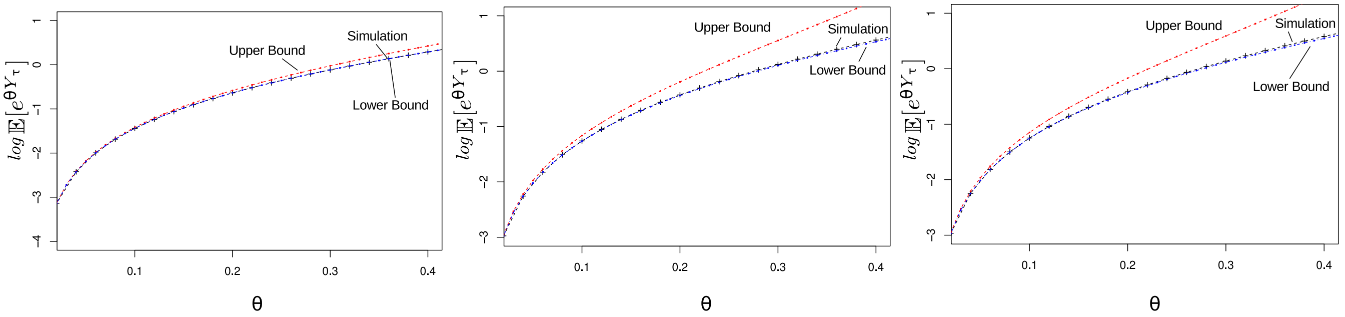

Fig. 2. Simulation vs. upper bounds on a log scale for the arrival rates as considered in Figure 1: $\lambda = 9$ (left), $\lambda = 2.1$ (center), $\lambda = 2.01$ (right). Again, this plots show that for θ close to zero, the bound are tight and capture the simulation's gradient closely.

Together with Proposition 2 this proves that the assumptions of optional stopping hold in this case, and that Lemmas 1 and 2 are applicable.

A closed form of the exponential MGF is known as $\Phi(\omega) = (1-\omega/\lambda)^{-1}$. For the lower bound, $\omega = \theta/p$ by definition. To shorten the derivation, we abbreviate $\Phi(\theta/p)$ by c as introduced in Example 1. Next, we plug the probability mass function of τ into the bounds and algebraically reformulate to the series representation of the exponential function.

$$\mathbb{E}\left[e^{\theta Y_\tau}\right] \geq \left(\sum_{k=1}^{\infty} c^{\frac{-pk}{p-1}} \frac{\lambda^{k-1} e^{-\lambda T} T^{k-1}}{(k-1)!}\right)^{1-p} \tag{29}$$

$$= \left(c^{\frac{-p}{p-1}} e^{-\lambda T} \left(\sum_{k=0}^{\infty} \frac{\left(c^{\frac{-p}{p-1}} \lambda T\right)^k}{k!} - 1\right)\right)^{1-p} \tag{30}$$

$$= (c^{\frac{-p}{p-1}} e^{-\lambda T} e^{c^{\frac{-p}{p-1}} \lambda T})^{1-p} = c^p e^{(1-p)(c^{\frac{-p}{p-1}}-1)\lambda T} \tag{31}$$

The derivation of the upper bound follows the same scheme with changed exponents and a different constant d as now $\omega = \theta p$ as defined in Example 1.

$$\mathbb{E}\left[e^{\theta Y_\tau}\right] \leq \left(\sum_{k=1}^{\infty} d^{\frac{k}{p-1}} \frac{\lambda^{k-1} e^{-\lambda T} T^{k-1}}{(k-1)!}\right)^{\frac{p-1}{p}} \tag{32}$$

$$= \left((d^{\frac{1}{p-1}} e^{(d^{\frac{1}{p-1}}-1)\lambda T}\right)^{\frac{p-1}{p}} \tag{33}$$

■

Example 2 (Poisson Arrivals, Exp. TTLs, and Model $\mathcal{B}$.))**.** *Assume caching model $\mathcal{B}$.), $X_i \sim exp(\lambda)$, and $T_i \sim exp(\mu)$. Then p can be optimized on the bounds:*

$$\mathbb{E}\left[e^{\theta Y_\tau}\right] \geq \left(\frac{\mu}{\lambda} \frac{c_1}{1-c_1}\right)^{1-p} \tag{34}$$

$$\mathbb{E}\left[e^{\theta Y_\tau}\right] \leq \left(\frac{\mu}{\lambda} \frac{c_2}{1-c_2}\right)^{\frac{p-1}{p}} \tag{35}$$

given that the following constraints are fulfilled:

- $c_1 := \left(\frac{p\lambda}{p\lambda-\theta}\right)^{\frac{-p}{p-1}} \frac{\lambda}{\lambda+\mu} < 1$
- $c_2 := \left(\frac{\lambda}{\lambda-\theta p}\right)^{\frac{1}{p-1}} \frac{\lambda}{\lambda+\mu} < 1$
- *$1 < p$, and $\theta < \lambda$.*

Proof: Assume $X_i \sim exp(\lambda)$, and $T_i \sim exp(\mu)$. The probability mass function of the stopping time can be derived as previously suggested in Equations (15) and (16):

$$\mathbb{P}(\tau = k) = \left(\int_0^\infty (1-e^{-\lambda x})\mu e^{-\mu x} dx\right)^{k-1} \tag{36}$$

$$\int_0^\infty e^{-\mu x} \lambda e^{-\lambda x} dx \tag{37}$$

This is easy to solve, as both integrals translate into the definition of the MGF of exponential random variables.

$$\mathbb{P}(\tau = k) = \left(1 - \frac{\mu}{\lambda+\mu}\right)^{k-1} \frac{\mu}{\lambda+\mu} \tag{38}$$

$$= \frac{\mu}{\lambda}\left(1 - \frac{\mu}{\lambda+\mu}\right)^k \tag{39}$$

As in the previous example, it holds that $\mathbb{P}(\tau < \infty) = \sum_{k=1}^{\infty} \mathbb{P}(\tau = k) = \sum_{k=1}^{\infty} \frac{\mu}{\lambda}\left(\frac{\lambda}{\lambda+\mu}\right)^k = 1$ and we can use Proposition 2 to show that the requirements of the bounds are fulfilled. Given the equation for $\mathbb{P}(\tau = k)$, both upper and lower bound are immediate because $\phi(\omega) = \frac{\lambda}{\lambda-\omega}$:

$$\mathbb{E}\left[e^{\theta Y_\tau}\right] \geq \left(\frac{\mu}{\lambda}\sum_{k=1}^{\infty}\left(\left(\frac{p\lambda}{p\lambda-\theta}\right)^{\frac{-p}{p-1}} \frac{\lambda}{\lambda+\mu}\right)^k\right)^{1-p} \tag{40}$$

$$\mathbb{E}\left[e^{\theta Y_\tau}\right] \leq \left(\frac{\mu}{\lambda}\sum_{k=1}^{\infty}\left(\left(\frac{\lambda}{\lambda-\theta p}\right)^{\frac{1}{p-1}} \frac{\lambda}{\lambda+\mu}\right)^k\right)^{\frac{p-1}{p}} \tag{41}$$

We denote the constants of the two geometric series by c_1 for the lower bound and c_2 for the upper bound, as defined in Example 2. Both constants must be strictly less than 1. Furthermore, $1 < p$, and $\theta p < \lambda$, and thus $\theta < \lambda$. ■

A. Simulation Results on the MGF

Comparing simulation results to the bounds on a moment generating function may not appear straightforward as the interpretation of the values of an MGF $\mathbb{E}\left[e^{\theta Y_\tau}\right]$ for various parameters θ is not obvious. However, we know that in order

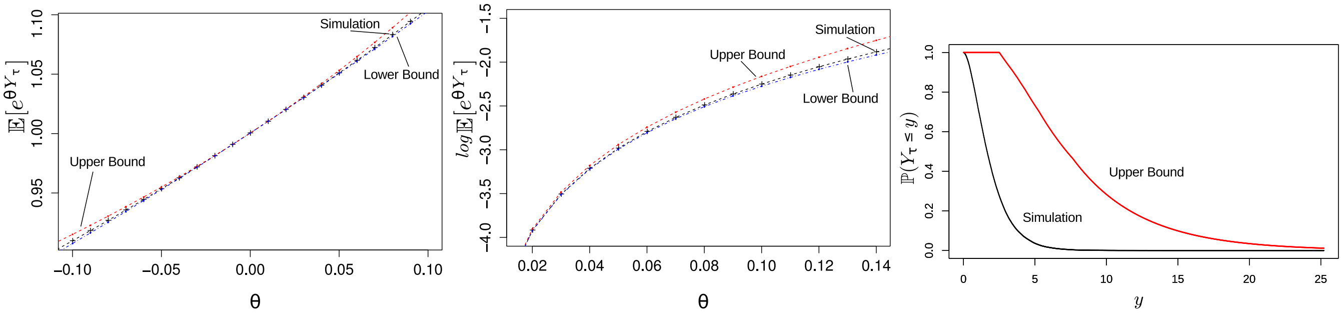

Fig. 3. Left: simulation vs upper and lower bound for both exponential arrivals and timers, $\lambda = 2.1$ and $\mu = 2$. Here, the lower bound's tightness is worse than for deterministic TTLs, but still both bounds capture the process's dynamics closely. For other configurations the tightness behaves comparable to the case of deterministic TTLs. Center: log-scale plot of the same configuration. Right: the Chernoff bound on the CCDF of Y_τ for a particularly bad configuration: the upper bound is impractically loose.

to obtain the i-th moments from an MGF, we take the i-th derivation and evaluate for $\theta = 0$. Thus, a bound needs to capture the gradient of the MGF at $\theta = 0$, in the best case for the first few orders of derivations.

By simulation we found the bounds to be sensitive to the ratio λ/μ. For this reason we compare different configurations of λ and μ. For small ratios between λ and μ, the bounds are only tight for a small environment of θ around zero. This environment expands for greater ratios, and, for what we consider a realistic configuration, $\lambda = 9$ and $\mu = 2$, the range of θ where the bounds appears tight increases visibly. Corresponding plots of an empirical MGF obtained from simulation are compared to our bounds in Figure 1 (for a deterministic TTL, Example 1) and in Figure 3 (for exponential TTLs, Example 2).

We show the bounds on logarithmic scale in Figure 2 and find that for deterministic TTL the upper bound matches the MGF's slope at $\theta = 0$ in all cases. In particular, the lower bound is tight in all cases, while the upper bound can get loose for θ far from zero. For exponential TTLs, and even for unfavorable conditions (small ratio of λ and μ), the upper bound is tighter than for deterministic TTLs, but the lower bound deteriorates slightly (cf. Figure 3).

B. Simulation Results on First and Second Moment

In this section, we show how to derive an upper bound on the first and second moment for two configurations of exponential timers in the caching model $\mathcal{B}$.) as this derivation is shorter than others and the bounds do not match as well as for other configurations. For example, the mean for deterministic TTLs matched our simulation results with three digits in all cases.

Thus, assume the bounds from Example 2 and let us first consider the configuration $\lambda = 9$ and $\mu = 2$. For simplicity we omit any optimization of the parameter p in the former case, fix $p = 2$, and obtain from Lemma 2:

$$\mathbb{E}\left[e^{\theta Y_\tau}\right] \le \left(\frac{2}{9}\,\frac{1}{\frac{1}{\left(\frac{9}{9-2\theta}\,\frac{9}{11}\right)} - 1}\right)^{\frac{1}{2}} = 3\sqrt{\frac{1}{9-11\theta}} \tag{42}$$

Next we obtain the first derivative and evaluate at $\theta = 0$ to obtain the first moment of Y_τ:

$$\frac{d}{d\theta}\left(3\sqrt{\frac{1}{9-11\theta}}\right) = \frac{33}{2}\left(\frac{1}{9-11\theta}\right)^{3/2} \tag{43}$$

$$\mathbb{E}\left[Y_\tau\right] \le \frac{33}{2}\left(\frac{1}{9}\right)^{3/2} = \frac{11}{18} \tag{44}$$

From corresponding simulations we obtain $\mathbb{E}\left[Y_\tau\right] \approx 0.6104813$ and this compares to the upper bound of $\frac{11}{18} = 0.6\bar{1}$ with a relative error of less than one percent.

In the same way, we compute the first moment for what we previously found to be a worst-case corner case: $\lambda = 2.1$ and $\mu = 2$. This time we fix a better p which is the smallest p such that $c_2 < 1$, cf. Fig. 3. We obtain:

$$\frac{d}{d\theta}\left(\frac{2}{2.1}\,\frac{1}{\frac{1}{\left(\left(\frac{2.1}{2.1-1.01\theta}\right)^{\frac{1}{0.01}}\,\frac{2.1}{2.1+2}\right)} - 1}\right)^{\frac{0.01}{1.01}}(0) = 0.97619 \tag{45}$$

which still compares to a simulation result of $\mathbb{E}\left[Y_\tau\right] \approx 0.9761672$ with a relative error of less than one percent.

However, the tightness of our bound on the second moment is already very loose. Comparing the upper bound

$$\frac{d}{d\theta}\frac{33}{2}\left(\frac{1}{9-11\theta}\right)^{\frac{3}{2}} = \frac{1089}{4}\left(\frac{1}{9-11\theta}\right)^{\frac{5}{2}} \tag{46}$$

$$sd(Y_\tau) = \sqrt{\frac{1089}{4}\left(\frac{1}{9}\right)^{\frac{5}{2}}} = \sqrt{\frac{121}{108}} \tag{47}$$

to the simulation result gives a relative error of more than fifty percent.

C. Simulation Result on a Classical Chernoff Bound

In principle it is possible to use classical inequalities to bound the distribution of the process Y_τ. The upper bound from Lemma 2 can be plugged into a Chernoff bound [25] to obtain:

$$P(Y_\tau > y) \le \inf_{0<\theta}\left(\mathbb{E}\left[e^{\theta Y_\tau}\right]e^{-\theta y}\right) \tag{48}$$

Then, under previous assumptions and similar constraints on p and θ it holds:

$$P(Y_\tau > y) \leq \inf_{0<\theta} \left(\left(\frac{\mu}{\lambda} \frac{1}{\frac{1}{\left(\left(\frac{\lambda}{\lambda - \theta p} \right)^{\frac{1}{p-1}} \left(\frac{\lambda}{\lambda + \mu} \right) \right)} - 1} \right)^{\frac{p-1}{p}} e^{-\theta y} \right) \tag{49}$$

However, simulations show that this bound is impractical even for simple cases such as Poisson arrivals. In Figure 3 we show a particularly bad result for a small ratio between λ and μ. This result constitutes a clear limitation of this approach.

V. Conclusion

In this work we propose a martingale approach to the modeling of TTL-based caching systems. We are able to derive tight bounds on the first moment of the inter-miss process. Our approach is sufficiently general to capture moments of even more general caching models as the model $\mathcal{B}$.) introduced in Section II because it requires only humble renewal-property assumptions. However, we observe the results presented to have severe limitations, too. Considering the case of cache networks, even a simple line of cache evades our analysis as we are not able to bound the miss process itself (but only its lower-order moments). In particular, using stochastic bounds on the miss probability distribution such as a Chernoff bound leads to loose results which prevent repeated application.

To address the limitations of this work, we consider further abstractions from the process to be most promising. Defining different martingales which do not exhibit the dependency as in our martingale definition may yield tighter bounds or finally exact results on this problem.

References

[1] W. F. King, "Analysis of demand paging algorithms," in *IFIP Congress (1)*, 1971, pp. 485–490.

[2] A. Dan and D. Towsley, *An approximate analysis of the LRU and FIFO buffer replacement schemes*. ACM, 1990, vol. 18, no. 1.

[3] C. Fricker, P. Robert, and J. Roberts, "A versatile and accurate approximation for LRU cache performance," in *Proceedings of the 24th International Teletraffic Congress*, ser. ITC '12. International Teletraffic Congress, 2012, pp. 8:1–8:8.

[4] N. Tsukada, R. Hirade, and N. Miyoshi, "Fluid limit analysis of FIFO and RR caching for independent reference model," *Performance Evaluation*, 2012.

[5] P. R. Jelenkovic, "Asymptotic approximation of the move-to-front search cost distribution and least-recently used caching fault probabilities," *The Annals of Applied Probability*, vol. 9, no. 2, pp. 430–464, 1999.

[6] V. Jacobson, D. K. Smetters, J. D. Thornton, M. F. Plass, N. H. Briggs, and R. L. Braynard, "Networking named content," in *Proceedings of the 5th International Conference on Emerging networking experiments and technologies*. ACM, 2009, pp. 1–12.

[7] T. Koponen, M. Chawla, B.-G. Chun, A. Ermolinskiy, K. H. Kim, S. Shenker, and I. Stoica, "A data-oriented (and beyond) network architecture," in *ACM SIGCOMM Computer Communication Review*, vol. 37, no. 4. ACM, 2007, pp. 181–192.

[8] H. Che, Y. Tung, and Z. Wang, "Hierarchical web caching systems: Modeling, design and experimental results," *Selected Areas in Communications, IEEE Journal on*, vol. 20, no. 7, pp. 1305–1314, 2002.

[9] E. J. Rosensweig, J. Kurose, and D. Towsley, "Approximate models for general cache networks," in *INFOCOM, 2010 Proceedings IEEE*. IEEE, 2010, pp. 1–9.

[10] N. C. Fofack, P. Nain, G. Neglia, and D. Towsley, "Analysis of TTL-based cache networks," in *Performance Evaluation Methodologies and Tools (VALUETOOLS), 2012 6th International Conference on*. IEEE, 2012, pp. 1–10.

[11] S. Saroiu, K. P. Gummadi, R. J. Dunn, S. D. Gribble, and H. M. Levy, "An analysis of internet content delivery systems," *ACM SIGOPS Operating Systems Review*, vol. 36, no. SI, pp. 315–327, 2002.

[12] J. Jung, A. W. Berger, and H. Balakrishnan, "Modeling TTL-based internet caches," in *INFOCOM 2003. 22nd Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies*, vol. 1. IEEE, 2003, pp. 417–426.

[13] E. Cohen and H. Kaplan, "The time-to-live based consistency mechanism," in *Web Content Delivery*. Springer, 2005, pp. 45–71.

[14] O. Bahat and A. M. Makowski, "Measuring consistency in TTL-based caches," *Performance Evaluation*, vol. 62, no. 1, pp. 439–455, 2005.

[15] A. Wald, "On cumulative sums of random variables," *The Annals of Mathematical Statistics*, vol. 15, no. 3, pp. 283–296, 1944.

[16] F. Baccelli and A. M. Makowski, "Martingale relations for the M/GI/1 queue with markov modulated poisson input," *Stochastic processes and their applications*, vol. 38, no. 1, pp. 99–133, 1991.

[17] N. Duffield, "Exponential bounds for queues with Markovian arrivals," *Queueing Systems*, vol. 17, no. 3-4, pp. 413–430, 1994.

[18] C.-S. Chang and J. A. Thomas, "Effective bandwidth in high-speed digital networks," *Selected Areas in Communications, IEEE Journal on*, vol. 13, no. 6, pp. 1091–1100, 1995.

[19] F. Ciucu, F. Poloczek, and J. Schmitt, "Sharp bounds in stochastic network calculus," in *ACM SIGMETRICS / International Conference on Measurement and Modeling of Computer Systems, SIGMETRICS '13, Pittsburgh, PA, USA, June 17-21*, 2013, pp. 367–368.

[20] A. Feldmann, "Characteristics of TCP connection arrivals," in *Self-Similar Network Traffic and Performance Evaluation*, K. Park and W. Willinger, Eds. John Wiley & Sons, Inc., 2002, ch. 15, p. 367–299.

[21] E. W. Montroll and G. H. Weiss, "Random walks on lattices. ii," *Journal of Mathematical Physics*, vol. 6, p. 167, 1965.

[22] G. R. Grimmett and D. R. Stirzaker, *Probability and random processes*. Oxford University Press, USA, 2001.

[23] S. Karlin and H. Taylor, "A first course in stochastic processes," 1975.

[24] J. Cheng, C. Tellambura, and N. C. Beaulieu, "Performance of digital linear modulations on weibull slow-fading channels," *Communications, IEEE Transactions on*, vol. 52, no. 8, pp. 1265–1268, 2004.

[25] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *The Annals of Mathematical Statistics*, vol. 23, no. 4, pp. 493–507, 1952.

Appendix

Lemma 3.

$$\int |fg| d\mu \geq \left(\int |f|^{\frac{1}{p}} d\mu \right)^p \left(\int |g|^{\frac{-1}{p-1}} d\mu \right)^{-(p-1)} \tag{50}$$

Proof: Set $q = \frac{p}{p-1}$.

With Hölder's inequality we obtain

$$\int |fg|^{\frac{1}{p}} |g|^{\frac{-1}{p}} d\mu \leq \left(\int \left(|fg|^{\frac{1}{p}} \right)^p d\mu \right)^{\frac{1}{p}} \tag{51}$$

$$\left(\int \left(|g|^{\frac{-1}{p}} \right)^{\frac{p}{p-1}} d\mu \right)^{\frac{p-1}{p}} \tag{52}$$

$$= \left(\int |fg| d\mu \right)^{\frac{1}{p}} \left(\int |g|^{\frac{-1}{p-1}} d\mu \right)^{\frac{p-1}{p}} \tag{53}$$

Raising this equation to the power of p gives:

$$\left(\int |f|^{\frac{1}{p}} d\mu \right)^p \leq \int |fg| d\mu \left(\int |g|^{\frac{-1}{p-1}} d\mu \right)^{p-1} \tag{54}$$

■