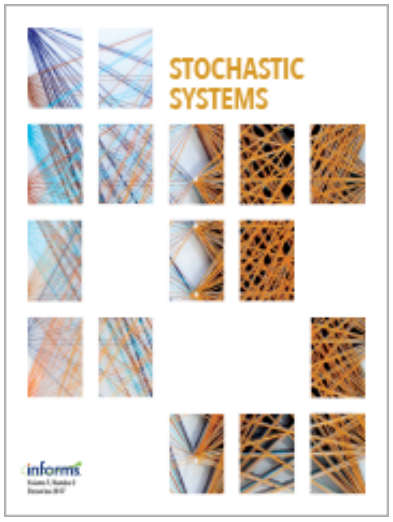


Stochastic Systems

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Optimal Local Storage Policy Based on Stochastic Intensities and Its Large-Scale Behavior

Matias Carrasco, Andres Ferragut, Fernando Paganini

To cite this article:
Matias Carrasco, Andres Ferragut, Fernando Paganini (2026) Optimal Local Storage Policy Based on Stochastic Intensities and Its Large-Scale Behavior. Stochastic Systems

Published online in Articles in Advance 13 May 2026

. https://doi.org/10.1287/stsy.2024.0093

https://pubsonline.informs.org/journal/stsy

STOCHASTIC SYSTEMS
Articles in Advance, pp. 1–30
ISSN 1946-5238 (online)

Optimal Local Storage Policy Based on Stochastic Intensities and Its Large-Scale Behavior

Matias Carrasco,[a] **Andres Ferragut,**[a,*] **Fernando Paganini**[a]

[a] Facultad de Ingenieria, Universidad ORT Uruguay, Montevideo 11100, Uruguay
*Corresponding author
Contact: carrasco_m@ort.edu.uy, https://orcid.org/0009-0001-6843-103X (MC); ferragut@ort.edu.uy, https://orcid.org/0000-0003-0134-5548 (AF); paganini@ort.edu.uy, https://orcid.org/0000-0002-1258-9049 (FP)

Received: November 29, 2024
Revised: August 21, 2025
Accepted: March 31, 2026
Published Online in Articles in Advance: May 13, 2026

https://doi.org/10.1287/stsy.2024.0093

Abstract. In this paper, we analyze the optimal management of local memory systems using the tools of stationary point processes. We provide a rigorous setting of the problem, and we characterize the optimal causal policy that maximizes the stationary hit probability. We then analyze a special case where the request processes come from a scale family and derive a suitable large-scale limit as the catalog size $N \to \infty$ when a fixed fraction c of items can be stored. In this limiting regime, the optimal policy amounts to comparing the stochastic intensity of the process with a fixed *threshold*, which is defined by a quantile of an appropriate limit distribution. We derive asymptotic performance metrics as well as sharp estimates for the prelimit case. Moreover, we establish a connection with optimal timer-based policies for renewal traffic and monotone hazard rates. We also present detailed validation examples of our results, including some closed-form expressions for the miss probability that are compared with simulations. We also use these examples to exhibit the significant superiority of the optimal policy for the case of regular traffic patterns.

Funding: This research was partially supported by the U.S. Air Force Office of Scientific Research [Grant FA9550-23-1-0350].

Keywords: point processes • stochastic intensity • caching systems

1. Introduction

In modern computing systems, a crucial component for improving performance at several layers is the use of a *local memory*, typically referred to as a *cache*. The rationale is that storing frequently used items at a readily available location can improve latency, reducing the cost of retrieval from a more remote location. This strategy is pervasive in computing systems: from local caching of instructions at the processor level to texture caching in graphics processing units, disk caching for quickly retrieving data from hard disks, content caching in web applications, geographically distributed caching in content delivery networks, and cloud storage gateways keeping readily available items stored in the cloud data centers.

A *local memory* consists of a certain amount of memory space that may store a subset of items locally and temporarily; the goal is to appropriately choose from a large catalog the subset of items more likely to be requested next. All of the aforementioned applications can be subsumed into this basic structure. In the case of homogeneous item sizes, the space constraint is reflected in the *number* of objects C that may be locally stored from the catalog of size $N \gg C$.

Classical analyses of local memory systems have focused on modeling the (discrete-time) sequence of item requests. Based on this empirical sequence, the system must determine which items are stored and which must be *evicted* from memory to make room for others. In a stationary regime, one natural strategy would be to store at all times the C items with the highest *popularity* measured by their mean intensity of requests; this simple *static policy* requires, however, popularities to be known. A practical approximation is the *least frequently used* (LFU) eviction policy; here, item request intensities are dynamically estimated from the arrival sequence, and the least popular ones are evicted. Under mild assumptions on stationary arrivals, LFU will eventually converge to the static policy.

A popular alternative is the *least recently used* (LRU) policy, which keeps in memory the C most recently requested items. Upon receiving a request, it will store the item if not already present, and in that case, it will

evict the oldest item in the sequence of requests. This method is better suited to handle highly *correlated* demands, where a requested item is more likely to be fetched again (i.e., bursty request patterns). Smoother variants have been analyzed, and also, combinations of both approaches, such as *Least Frequent Recently Used* (cf. Bilal and Kang 2017), have been proposed. We review the relevant literature below.

A main drawback of the classical analysis is that relevant *time* (continuous-time) information is neglected; by focusing only on the sequence of requests, the models ignore interrequest times that are important characteristics of the request processes. An alternative approach that gives time the center stage is *timer-based* (time-to-live (TTL)) caching policies; here, items are stored upon request and evicted only after a certain amount of time has elapsed because their last appearance in the request stream. This is a common approach in internet-based systems, such as domain name system queries and web applications.

A crucial step toward incorporating time information is the seminal paper by Fofack et al. (2014), where the incoming request stream is assumed to be a stationary point process on the real line. The mean intensities of the underlying request processes for each item capture their relative popularity, whereas by modeling the interrequest times, we can express different types of behavior; for instance, heavy-tailed interrequest times are well adapted to bursty arrival patterns. This approach has led to new insights in the analysis of both replacement and timer-based caching policies, which we also review below.

Within this modeling framework, we ask a natural question. What is the *optimal* memory management policy? By this, we mean the one that maximizes the *hit rate*, which is the frequency of successful retrievals from local memory. In Ferragut et al. (2016, 2018), the optimality question for TTL caching policies was investigated; in particular, it was shown that if requests form a general stationary simple point process, optimality may be characterized by the *hazard rate* function of the interarrival distribution. Later, Panigrahy et al. (2022) analyzed optimal replacement policies (i.e., fixed memory systems), and they identified the optimal local storage policy in terms of the *stochastic intensity*, a generalization of the hazard rate function.

In this paper, our goal is to develop this connection more extensively, bringing to bear the machinery of stationary point processes as in Brémaud (2020) to formally characterize causal memory management policies and the condition for optimality, in particular for general stationary request processes and correlation structures. Subsequently, our aim is to understand the large-scale behavior of the optimal replacement policy for large N under some appropriate assumptions: namely, that request processes are independent and that interarrival times for different items are drawn from a *scale family* parametrized by a distribution of intensities that has an asymptotic limit. We characterize the optimal memory management as a *threshold* policy on the hazard rates observed by the system, and we prove that the threshold has a deterministic limit. We also obtain a closed expression for the optimal *performance* in the large-scale limit as a function of the fundamental parameters of the scale family and the popularity distribution. As a further contribution, we investigate the properties of threshold policies before the asymptotic limit, showing that under monotonicity of the hazard rate function, they become *equivalent* to timer-based policies. As a consequence, we obtain the convergence of the two types of policies (replacement or TTL) to a common optimum in the asymptotic limit, a striking connection between both approaches. Our results are illustrated by a series of parametric examples, and their properties are further exhibited through stochastic simulations.

1.1. Related Work

Being essential to modern computing systems, the performance analysis of caching and local memory management has a long history. The seminal work on replacement-based algorithms with a fixed memory size started in King (1971) and Gelenbe (1973). Despite its apparent simplicity, the analysis of replacement policies is complex, even for a single cache; King (1971) gives an explicit expression of the hit probability of the LRU policy with exponential complexity. In Gelenbe (1973), it is shown that under the so-called *independent reference model* (IRM) assumption, where requests are independent and identically distributed, replacement policies, such as first in, first out, achieve the same hit probability. Exact computation is, however, intractable. Dan and Towsley (1990) provide an approximate computational procedure. Their method has been further extended in Rosensweig et al. (2010) to the case of networks of cache systems. Also, in Gast and Houdt (2015), another approach is given for list-based replacement policies.

A related line of work that includes replacement policies is analyses based on the move-to-front (MTF) rule, which is equivalent to LRU. A first step in this direction is the work by Fill (1996), which addresses the limit cost of the move-to-front rule. By exploiting the connection between MTF and LRU, Jelenković (1999) and Jelenković and Radovanović (2008) provide asymptotic expressions for the hit probability under the IRM assumption and Zipf popularities with parameter β. With a fundamentally different approach, based on Laplace transforms, the same limit result is obtained in Barrera and Fontbona (2010). Our asymptotic limit assumptions are related to

this latter approach. Few results on replacement policies move beyond the IRM assumptions, such as the works by Jelenković and Radovanović (2003, 2004) and Jelenković et al. (2006), which show some insensitivity properties of LRU in a large-scale regime and dependent requests. However, their approach is only valid for light-tailed popularities, and it is related to our result below on optimal performance for the nonuniformly integrable popularity limit. Extensions to a network of caches working cooperatively to optimize performance are proposed in Borst et al. (2010), Ioannidis et al. (2010), and Ioannidis and Yeh (2016).

Because exact analysis of LRU is difficult, the most popular technique for approximating its performance is the so-called "Che approximation" introduced in Che et al. (2002) for web caches. The crucial point is to define a *characteristic time* representing the average permanence time *common* to all files. This assumption is valid in a large-scale regime, and Fricker et al. (2012) perform a second-order analysis to explain why this is a good approximation.

On the other hand, the analysis of policies based on timers is more recent and associated with the growth of internet caches. A first contribution in this line is Jung et al. (2003), with expressions for the steady-state hit probabilities for web pages; this line of work was later extended in Bahat and Makowski (2005) to include update delay. However, a crucial contribution is the introduction of point process theory to capture request correlations and in particular, move beyond the *sequence* of requests to a continuous-time model with general arrival processes. The foundations for this approach were laid down in Fofack et al. (2012, 2014). In Berger et al. (2014), the analysis is extended to a family of TTL policies with the focus of approximating LRU performance in linear and tree networks, and a different policy is proposed in Berger et al. (2015) with the aim of maximizing hit ratios by variance reduction. In Bianchi et al. (2013), the Che approximation is analyzed in a more general setting, and in Martina et al. (2014), it is extended using TTL cache tools to renewal arrivals, showing good accuracy in the case of small caches and Zipf-law popularities and making the connection between replacement and timer-based policies. The fact that timer-based policies decouple the analysis over independent request streams has led to more amenable generalizations to the case of cache networks (Panigrahy et al. 2017, 2020; Dehghan et al. 2019).

The search for optimal policies with respect to the hit rate has received recent attention. In Ferragut et al. (2016, 2018), the problem of finding the optimal *timer-based* policy is formulated and related to the *hazard rate function* of the interrequest times. Using tools from convex optimization, Ferragut et al. (2016, 2018) show that *decreasing* hazard rates lead to a nontrivial timer-based policy that achieves optimality and characterize its large-scale behavior. However, more regular request traffic with *increasing* hazard rates does not benefit from caching at all, a striking result that highlights the role of traffic regularity in local memory management systems. More recently in Ferragut et al. (2024), a generalization of timer-based policies to consider *prefetching* was introduced, leading to a nontrivial optimal policy for increasing hazard rates. A parallel research line developed in Panigrahy et al. (2022) considers the optimality question for *replacement-based* policies; the structure of the optimal *causal* policy is identified, incorporating for the first time the notion of *stochastic intensity* of point processes, a generalization of the hazard rate function. The current paper builds on these immediate precursors.

1.2. Main Contributions

The first contribution of this paper is a general foundation for the problem of optimal causal policies for the local memory or caching problem. The setup is the same as the one considered in Panigrahy et al. (2022), but we provide here a more complete mathematical treatment within the framework of point processes in the real line following Brémaud (2020). In particular, this approach enables us to properly define predictable policies in terms of the underlying natural filtration of the involved processes, which is a key technical issue to be addressed in the proofs. We also aim at greater generality, determining the optimal causal policy for a general superposition of request processes, including the correlated case, as a function of stochastic intensities of the underlying request processes and their correlation structure. We then specialize this result to independent sources.

The second most significant part of the paper concerns the asymptotic behavior of the optimal policy under a large-scale regime (i.e., when the item catalog size goes to infinity). Assuming that the request processes for different items are independent with stochastic intensities from a common scale family and that their popularities have a limiting distribution, we prove that the optimal policy converges to a *threshold* policy in the stochastic intensities, with a deterministic threshold that follows closed-form expressions. Armed with this result, we characterize the limit of the optimal *miss rate* in large-scale systems through an explicit formula. Therefore, we provide a *universal asymptotic bound* on performance for any practical policy.

A third contribution concerns the behavior of deterministic threshold policies before the asymptotic limit; under renewal traffic and additional monotonicity assumptions of the *hazard rate function* of the interrequest times, it is shown that previously analyzed timer-based policies are, in fact, of the threshold type and in particular, that under renewal assumptions, the *optimal timer policy* for decreasing hazard rates identified in Ferragut et al. (2018) achieves the universal bound. For the dual case of increasing hazard rates, we further analyze the

timer-based prefetching policy introduced in Ferragut et al. (2024) and prove that it also achieves the universal asymptotic bound.

Although the preceding theory is developed for the case of homogeneous item sizes, we also explore the generalization to the heterogeneous case.

In our simulations section, we extensively analyze certain concrete parametric examples that illustrate the theory and serve as a benchmark for classical caching policies in comparison with the optimal. Finally, we provide some remarks on the practical challenges in implementing the optimal policy.

1.3. Organization of the Paper

The paper is organized as follows. In Section 2, we lay out the main tools of point processes and stochastic intensities required to analyze our system. We then define our local memory model, establish the framework for causal policies, and characterize the optimal policy in Section 3. Our main theorem describing the large-scale limit is presented in Section 4, and the ensuing universal performance bound is presented in Section 5. In Section 6, we describe the connection between the optimal policy and timer-based policies. In Section 7, we discuss the extension of the results to heterogeneous item sizes. Simulations and examples are presented in Section 8, and conclusions are given in Section 9.

2. Preliminaries and Notation

Throughout this paper, we will consider *stationary point processes* defined on a common probability space $(\Omega, \mathcal{F}, P)$. We recall now some basic concepts that will be useful in the following and introduce our notation; we refer the reader to Brémaud (2020) for a thorough treatment.

A simple and locally finite point process Φ on the real line is a random and strictly increasing sequence of points $\Phi = \{\tau_k\}_{k\in\mathbb{Z}}$ satisfying $\lim_{k\to\pm\infty}\tau_k = \pm\infty$. More formally, Φ can be cast as a *random counting measure* (i.e., $\Phi = \sum_k \delta_{\tau_k}$), a measurable map from $(\Omega, \mathcal{F}) \to (M^{\#}(\mathbb{R}), \mathcal{M}^{\#}(\mathbb{R}))$. Here, $M^{\#}(\mathbb{R})$ is the space of locally finite measures on $\mathbb{R}$ taking values in $\mathbb{N}\cup\{\infty\}$, and $\mathcal{M}^{\#}(\mathbb{R})$ is the smallest σ-algebra such that for all Borel sets $B \in \mathcal{B}(\mathbb{R})$, $\Phi(B) = \sum_k \mathbf{1}_{\{\tau_k\in B\}}$ is measurable. Moreover, $\Phi(B)$ is assumed to be finite for bounded B, and thus, it is a nonnegative integer-valued random variable. By definition, all points τ_k are different, and thus, $\Phi(\{x\}) \leqslant 1$ P almost surely (a.s.) for all $x \in \mathbb{R}$. In order to label the points, we follow the usual convention (Brémaud 2020), where $\tau_0(\Phi) \leqslant 0$ and $\tau_1(\Phi) > 0$. With this convention, $\tau_0 = \tau_0(\Phi)$ represents the first point before the time origin of the process Φ.[1]

Let $S_t(\Phi)$ denote the shift operator for measures in $\mathbb{R}$ (i.e., $S_t(\Phi)(B) := \Phi(B + t)$ for all Borel sets $B \subset \mathbb{R}$, where $B + t := \{x + t : x \in B\}$). The point process Φ is stationary if $S_t(\Phi)$ has the same distribution as Φ for all $t \in \mathbb{R}$. The mean measure of the point process is $\lambda(B) := \mathrm{E}[\Phi(B)]$. If the process is stationary, then this measure is translation invariant and thus, a multiple of the Lebesgue measure on $\mathbb{R}$ (i.e., $\lambda(B) = \lambda m(B)$). The constant λ is called the (average) *intensity* of the stationary point process. In what follows, we assume $\lambda > 0$ to avoid the trivial case where the process has no points.

For a simple stationary point process Φ, an important measure is the *Palm probability* P^0_Φ. This is a probability measure defined in $(\Omega, \mathcal{F})$ that captures the stochastic behavior of the point process when the observer is *synchronized* with it. In particular, $P^0_\Phi(\tau_0 = 0) = 1$ (i.e., there is P^0_Φ-a.s. a point at the origin). We refer the reader to Brémaud (2020) for a formal definition.

The key relationship between the Palm probability and the stationary probability is the following *inversion formula* valid for any nonnegative real-valued measurable function $f : M^{\#}(\mathbb{R}) \to \mathbb{R}_+$:

$$\mathrm{E}[f(\Phi)] = \lambda \mathrm{E}^0_\Phi\left[\int_0^{\tau_1} f(S_t(\Phi))dt\right]. \tag{1}$$

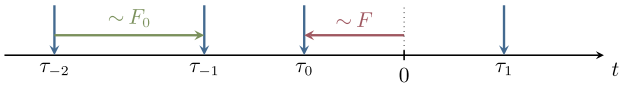
That is, in order to know the average value of a property in the stationary measure, we can integrate over one cycle of the process using the Palm measure and scale by λ.

The *interarrival distribution* of the point process Φ is defined as $F_0(t) := P^0_\Phi(\tau_1 - \tau_0 \leqslant t) = P^0_\Phi(\tau_1 \leqslant t)$. Because the process is simple, F_0 has support in $\mathbb{R}_+$. Moreover, $\mathrm{E}^0_\Phi[\tau_1] = 1/\lambda$, which follows from taking $f \equiv 1$ in (1).

A second important distribution is the *age distribution*, which is the age of the current interval when the process is observed at a point t not synchronized with it. Because the process is stationary, we can take without loss of generality $t = 0$, and thus, the age distribution is just

$$F(t) := P(-\tau_0 \leqslant t) = \lambda \int_0^t (1 - F_0(s))ds, \tag{2}$$

Figure 1. Interarrival and Age Distribution of a Stationary Point Process



a result that also follows from (1) with $f(\Phi) = \mathbf{1}_{\{\tau_0(\Phi) \geqslant -t\}}$. Note that these distributions are different in general because of the bias toward larger intervals when sampling in the steady state. A depiction of this sampling effect is shown in Figure 1.

2.1. Stochastic Intensity

We now introduce the concept of *stochastic intensity*, which is crucial for the analysis in this paper. First, we need the following definition.

Definition 1. A filtration $\{\mathcal{F}_t\}_{t\in\mathbb{R}}$ in $(\Omega, \mathcal{F})$ (i.e., an increasing family of σ-algebras contained in $\mathcal{F}$) is *a history* of the simple locally finite point process Φ if $\Phi((a,b])$ is $\mathcal{F}_t$ measurable for all $a < b \leqslant t$. The *natural history* of Φ is $\mathcal{F}_t = \sigma(\Phi((a,b]) : a < b \leqslant t)$, the smallest filtration satisfying this property.

We define the stochastic intensity of a process for a given history (Brémaud 2020, definition 5.1.1).

Definition 2. Let Φ be a simple locally finite point process, and let $\mathcal{F}_t$ be a history of Φ. If there exists a locally integrable $\mathcal{F}_t$-adapted process $\varphi(t) \geqslant 0$ satisfying

$$\mathrm{E}[\Phi((s,t]) \,|\, \mathcal{F}_s] = \mathrm{E}\left[\int_s^t \varphi(u)du \,\middle|\, \mathcal{F}_s\right] \quad \text{for all } s < t, \tag{3}$$

then $\varphi(t)$ is called an $\mathcal{F}_t$ *stochastic intensity* of Φ.

The process $\varphi(t)$ acts as the local likelihood of a point appearing at time t given past information. This notion will play a key role in our particular application. Also, directly from the definition, one can prove that $\mathrm{E}[\varphi(t)] = \mathrm{E}[\varphi(0)] = \lambda$, the average intensity of the process.

As an example, the stationary Poisson process of intensity λ with its natural history satisfies (3) with $\varphi(t) \equiv \lambda$, a deterministic constant, and thus, the likelihood of a point appearing in time is independent of the past, reflecting the total randomness property of the Poisson process.

We now highlight some key properties of the stochastic intensity that will be useful for our later analysis. The first one is related to predictability; for this, we need the following definition.

Definition 3. Let $\{\mathcal{F}_t\}_{t\in\mathbb{R}}$ be a filtration on $(\Omega, \mathcal{F})$. The *predictable* σ-algebra $\mathcal{P}(\mathcal{F}.)$ associated with $\mathcal{F}_t$ is the σ-algebra on $\mathbb{R} \times \Omega$ generated by the sets of the form

$$(a,b] \times A,\ a < b,\ A \in \mathcal{F}_a.$$

A stochastic process $X(t,\omega)$ taking values on a measurable space $(E, \mathcal{E})$ is $\mathcal{F}_t$ *predictable* if the mapping $(t,\omega) \longmapsto X(t,\omega)$ is $\mathcal{P}(\mathcal{F}.)$ measurable.

The following properties can be found in Brémaud (2020); any real-valued and *left*-continuous stochastic process adapted to $\mathcal{F}_t$ is $\mathcal{F}_t$ predictable. Moreover, the stochastic intensity of a point process Φ if it exists can always be chosen to be a.s. predictable up to a set of Lebesgue measure 0, and thus, it is essentially unique. Also, the following result provides a *smoothing formula* for predictable processes based on the stochastic intensity.

Theorem 1. *Let Φ be a simple point process in $\mathbb{R}$ with history $\mathcal{F}_t$, and let $\varphi(t)$ be an $\mathcal{F}_t$-adapted and a.s.-integrable process. Then, $\varphi(t)$ is an $\mathcal{F}_t$-stochastic intensity of Φ if and only if*

$$\mathrm{E}\left[\int Z(t)\Phi(dt)\right] = \mathrm{E}\left[\int Z(t)\varphi(t)dt\right] \tag{4}$$

holds for all $\mathcal{F}_t$-predictable processes $Z(t)$.

Finally, we quote the following result, known as Papangelou's formula (Brémaud 2020, theorem 7.7.5).

Theorem 2. *Let Φ be a simple point process in $\mathbb{R}$ with history $\mathcal{F}_t$, stochastic intensity $\varphi(t)$, and mean intensity λ. If $Z(t)$ is an $\mathcal{F}_t$-predictable process and $f(z)$ is a real-valued function, then*

$$\mathrm{E}[f(Z(0))\varphi(0)] = \lambda \mathrm{E}_\Phi^0[f(Z(0))]. \tag{5}$$

For the purposes of our analysis, an important random variable is the stochastic intensity observed when sampling the renewal process at a fixed time; for convenience, it is chosen to be $t = 0$. We call it the observed stochastic intensity (OSI) and denote it by $X := \varphi(0)$.

To find the distribution of the OSI, there are two underlying probability measures to consider.

Definition 4. Let $X := \varphi(0)$ be the observed stochastic intensity at time 0. Define

$$G(x) := \mathrm{P}(X \leqslant x) = \mathrm{E}[\mathbf{1}_{\{\varphi(0) \leqslant x\}}]; \qquad G_0(x) := \mathrm{P}^0_\Phi(X \leqslant x) = \mathrm{E}^0_\Phi[\mathbf{1}_{\{\varphi(0) \leqslant x\}}]. \tag{6}$$

Here, $G(x)$ is the distribution of the stochastic intensity at a sampling point not synchronized with the process. G_0 is the distribution of the OSI at the time of a request; because the stochastic intensity can be chosen to be left continuous, G_0 identifies the stochastic intensity distribution prior to this arrival. A useful inequality that will be needed later is the following lemma.

Lemma 1. *The distribution of the OSI prior to arrival events satisfies*

$$G_0(x) \leqslant \frac{x}{\lambda}. \tag{7}$$

Proof. The result follows from the formula of Papangelou; taking $Z(t) = \varphi(t)$ and $f(z) = \mathbf{1}\{z \leqslant x\}$ in (5), we get

$$\lambda G_0(x) = \lambda \mathrm{E}^0_{\Phi_0}[\mathbf{1}_{\{\varphi(0) \leqslant x\}}] = \mathrm{E}[\varphi(0)\mathbf{1}_{\{\varphi(0) \leqslant x\}}] \leqslant \mathrm{E}[x\mathbf{1}_{\{\varphi(0) \leqslant x\}}] \leqslant x. \qquad \square$$

2.2. Renewal Point Processes

As an important special case, consider now that Φ is a stationary *renewal* process, meaning that the interrequest time sequence $\{\tau_{k+1} - \tau_k\}_{k\in\mathbb{Z}}$ includes P^0_Φ independent and identically distributed random variables with common distribution F_0. We assume that F_0 has a density f_0 and recall the definition of the *failure rate* or *hazard rate* function associated with F_0:

$$\eta(t) = \frac{f_0(t)}{1 - F_0(t)}. \tag{8}$$

Then, if $\mathcal{F}_t$ is the natural history of the process, the stochastic intensity of Φ is given by Daley and Vere-Jones (2003, chapter 7):

$$\varphi(t) = \eta(t - \tau^-(t)), \tag{9}$$

where $\tau^-(t) = \sup\{\tau_k : \tau_k < t\}$ is the last point before t. Note that $\tau^-(t)$ is a left-continuous process. In particular, because of the renewal property, the local intensity of the process depends only on the *age* of the current interval and the hazard rate function of the interarrival distribution.

When F_0 is the exponential distribution as in the Poisson process, $\eta(t) \equiv \lambda$, so the stochastic intensity is constant as previously stated. Another interesting parametric example that we will use for illustration is the case of Pareto interarrival times; it is presented below in Section 2.3.

In the renewal case, we can provide more explicit formulas for the distribution of the observed stochastic intensity $X := \varphi(0) = \eta(-\tau^-(0))$. In particular, at a time not synchronized with arrivals, we will have

$$G(x) := \mathrm{P}(X \leqslant x) = \mathrm{P}(\eta(-\tau_0) \leqslant x) = \mathrm{P}(-\tau_0 \in \eta^{-1}([0,x])) = \int_{\eta^{-1}([0,x])} F(dt) \tag{10}$$

because $-\tau_0 \sim F$, the age distribution. We note that the set $\eta^{-1}([0,x])$ will be an interval if the hazard rates are monotone.

If instead, we are evaluating the OSI at an arrival time, we must use the Palm probability P^0_Φ, for which $\tau_0 \equiv 0$ a.s. and $\tau^-(\tau_0) = \tau^-(0) = \tau_{-1}$. Therefore, $X = \eta(-\tau_{-1})$, and its distribution is given by

$$G_0(x) := \mathrm{P}^0_\Phi(X \leqslant x) = \mathrm{P}^0_\Phi(\eta(-\tau_{-1}) \leqslant x) = \int_{\eta^{-1}([0,x])} F_0(dt) \tag{11}$$

because $-\tau_{-1} \sim F_0$ under the Palm probability.

2.3. Pareto Interarrival Times

An interesting parametric example of stationary renewal processes, which is considered in Ferragut et al. (2016) for the same application, is when interarrival times follow a heavy-tailed Pareto distribution. In this case, all of the above magnitudes have explicit expressions, which we now compute.

Example 1 (Renewal Pareto Process). In the above setting, choose

$$F_0(t) = 1 - \left(\frac{1}{1+\gamma t}\right)^{\alpha}, \quad f_0(t) = \frac{\alpha\gamma}{(1+\gamma t)^{\alpha+1}} \quad (t \geqslant 0). \tag{12}$$

Thus, F_0 is a Pareto distribution (starting at zero) with tail parameter $\alpha > 1$. The number γ acts as a scale parameter, and by direct computation, in order to have $\mathrm{E}_{\Phi}^{0}[\tau_1] = 1/\lambda$, it should be chosen such that $\gamma = \frac{\lambda}{\alpha-1}$.

From Equations (2) and (8), we can compute

$$F(t) = 1 - \left(\frac{1}{1+\gamma t}\right)^{\alpha-1}, \quad \eta(t) = \frac{\alpha\gamma}{1+\gamma t}, \quad (t \geqslant 0). \tag{13}$$

For this example, note that the hazard rate function is decreasing for any choice of the parameters. Therefore, at an arrival time τ_k, the stochastic intensity *increases* (the hazard rate resets to $\eta(0)$ following (9)), and a subsequent arrival becomes more likely. This gives rise to *bursty* traffic as depicted in Figure 2.

For this process, we can also compute the distributions of the observed stochastic intensity at nonsynchronized or synchronized times:

$$G(x) = \int_{\eta^{-1}(x)}^{\infty} F(dt) = \int_{\frac{\alpha}{x}-\frac{1}{\gamma}}^{\infty} F(dt) = \begin{cases} \left(\frac{x}{\alpha\gamma}\right)^{\alpha-1}, & 0 \leqslant x \leqslant \alpha\gamma, \\ 1 & x > \alpha\gamma. \end{cases} \tag{14}$$

$$G_0(x) = \int_{\eta^{-1}(x)}^{\infty} F_0(dt) = \int_{\frac{\alpha}{x}-\frac{1}{\gamma}}^{\infty} F_0(dt) = \begin{cases} \left(\frac{x}{\alpha\gamma}\right)^{\alpha}, & 0 \leqslant x \leqslant \alpha\gamma, \\ 1 & x > \alpha\gamma. \end{cases} \tag{15}$$

In Figure 3, we depict the distributions F_0, F, G, and G_0 as well as the hazard rate function η for the case $\alpha = 2$ and $\gamma = 1$ (with average intensity $\lambda = 1$). In this particular case, the nonsynchronized OSI distribution is uniform in $[0, 2]$.

3. Local Memory Systems and Optimal Storage Policy

We start by rigorously defining our model for a *local memory system*. Requests from a *catalog* of N equally sized items are received (we defer the discussion of heterogeneous sizes to Section 7). We model item requests by stationary stochastic point processes Φ_i, $i = 1, \ldots, N$, which are defined on a common probability space and with finite intensity $\lambda_i > 0$, a measure of the popularity of item i. By appropriate labeling, we can choose $\lambda_1 \geqslant \lambda_2 \geqslant \ldots$ (i.e., the objects are ordered by decreasing popularities). Thus, the complete request process is the superposition $\Phi := \sum_{i=1}^{N} \Phi_i$, and its total intensity is $\lambda^N = \sum_{i=1}^{N} \lambda_i$.

The local memory is limited, and thus, it can only keep readily available a subset of size $C < N$ of the items. These items can upon request be served from the local memory with lower cost than retrieving them from a central repository. This formulation is quite general, and it subsumes the typical notion of *caching*, which is useful in many applications.[2] Mathematically, the local memory can keep at any point in time a subset

$$\mathcal{C}(t) = \{i_1, \ldots, i_k\} \subset \{1, \ldots, N\} \text{ with } 0 \leqslant k \leqslant C.$$

We call the process $\mathcal{C} : \Omega \times \mathbb{R} \to \mathcal{P}_C(N)$ the *storage process* of the system, where $\mathcal{P}_C(N)$ denotes the subsets of $\{1, \ldots, N\}$ with size less than C.

Figure 2. Stochastic Intensity of a Renewal Pareto Process Showing Decreasing Hazard Rates

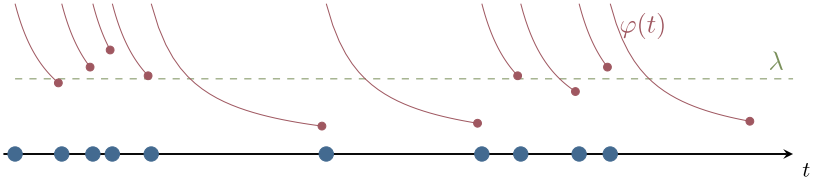
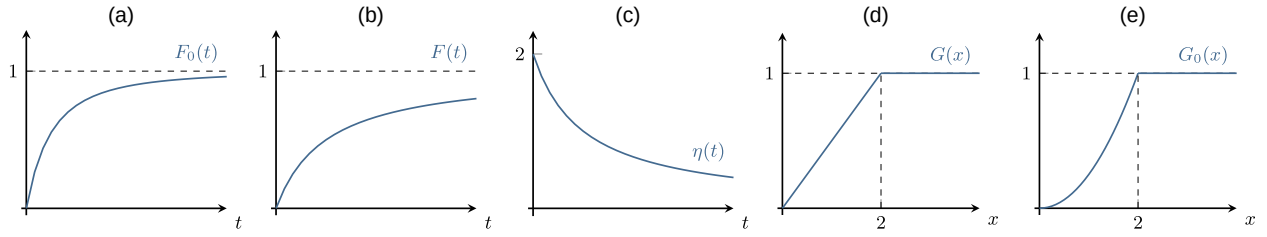


Figure 3. Cumulative Distribution Function (cdf) Shapes for the Pareto Renewal Example with $\alpha = 2, \gamma = 1$, and Therefore, $\lambda = 1$



Notes. (a) Interarrival cdf. (b) Age cdf. (c) Hazard rate. (d) Stationary observed hazard rate cdf. (e) Observed hazard rate upon arrival cdf.

For a given storage process, we can define the *hit-counting process* of the local memory system as

$$\Psi_{\mathcal{C}}(B) = \sum_{i=1}^{N} \int_B \mathbf{1}_{\{i \in \mathcal{C}(t)\}} \Phi_i(dt) \tag{16}$$

(i.e., the thinned process that counts the requests of objects only if they are stored in memory at the request time). The mean intensity $h_{\mathcal{C}}$ of the process $\Psi_{\mathcal{C}}$ is called the *hit rate*, and accordingly, it satisfies $h_{\mathcal{C}} \leqslant \lambda^N$.

The natural objective in this setting is to maximize $h_{\mathcal{C}}$ (i.e., the rate at which requests can be served directly from the local memory) by choosing an appropriate *policy* that defines the storage process $\mathcal{C}(t)$. However, this policy should be *causal* (i.e., it cannot make use of future information). Otherwise, the policy that only stores the next arriving item achieves a maximum rate using only one unit of memory for any N, and the problem is trivial. This is where the predictability notion introduced in Section 2.1 becomes important.

Let us denote by $\mathcal{F}_t^{(i)}$ the natural history of the ith request process Φ_i, and $\varphi_i(t)$ is its stochastic intensity with respect to $\mathcal{F}_t^{(i)}$. Define $\mathcal{F}_t = \sigma(\mathcal{F}_t^{(i)}; i = 1, \dots, N)$ as the aggregated history (i.e., the information about all arrivals up to time t). Note that we are not assuming independence between the request processes. We have the following definition.

Definition 5. Consider a local memory system with request processes $\{\Phi_i : i = 1 \dots N\}$, aggregated natural history $\mathcal{F}_t$, and capacity C. A *causal memory management* policy is a stationary $\mathcal{F}_t$-predictable stochastic process $\mathcal{C}(t)$, with values in the space $P_C(N)$ of all subsets of $\{1, \dots, N\}$ of size at most C (equipped with the discrete σ-algebra).

We would like to characterize the causal policy that maximizes the hit rate $h_{\mathcal{C}}$. We need the following assumption.

Assumption 1. *The aggregated process Φ is simple, and each process Φ_i admits a stochastic intensity $\tilde{\varphi}_i(t)$ with respect to the aggregated natural history $\mathcal{F}_t$ for all $i = 1, \dots, N$.*

Note that the stochastic intensity $\tilde{\varphi}_i(t)$ need not be equal to the individual stochastic intensity $\varphi_i(t)$ because of crossinformation across arrivals. However, the average intensity should still satisfy $E[\tilde{\varphi}_i(0)] = \lambda_i$. The following example illustrates this fact.

Example 2. Consider the following process; objects $1 \dots, N$ are requested alternatively one after another. When a request for item i arrives, an exponential distribution with parameter λ_i is started, and after expiration, the request for object $i + 1$ arrives (with the cyclic convention $N + 1 \longmapsto 1$). Let Φ_i denote the requests for item i and Φ denote the total request process. Then, it is clear that Φ_i is a renewal process with interval length following a hypoexponential (phase-type) distribution (i.e., the sum of independent exponentials with parameters λ_i). In this case, by Equation (9), $\varphi_i(t) = \eta(t - \tau_i^-)$, where η is the hazard rate function of the resulting hypoexponential, and it is the same for all i.

However, when looking at the aggregated history, it is clear that arrivals for item i can only come at rate λ_i after a request from process $i - 1$; therefore,

$$\tilde{\varphi}_i(t) = \lambda_i \mathbf{1}_{\{\tau^-(t) = \tau_{i-1}^-(t)\}}, \quad i = 1, \dots, N,$$

where $\tau^-(t)$ is the last point before t of the aggregated process.

Assumption 1 is also needed to guarantee the existence of $\tilde{\varphi}$; as an extreme example, consider in particular two processes Φ_1 and $\Phi_2 = S_{t_0}(\Phi_1)$ (i.e., a delayed version of the first with $t_0 > 0$ fixed). Then, their stochastic intensities with respect to their own natural histories are just delayed versions of one another: $\varphi_2(t) = \varphi_1(t - t_0)$. However, in the combined history of both processes, the stochastic intensity of process 2 is zero everywhere

except that it degenerates into Dirac δ functions at the t_0-shifted points of process 1. Hence, these processes do not have a stochastic intensity in the sense of Definition 2 (which requires a proper stochastic function $\varphi_i(t)$) with respect to the combined history of both processes. Assumption 1 ensures that these pathological examples with complete correlation are excluded.

We are now in position to compute the stochastic intensity of the hit process.

Lemma 2. *Let $\mathcal{C}(t)$ be a causal policy. The stochastic intensity of its associated hit process $\Psi_{\mathcal{C}}(t)$ defined by (16) with respect to the aggregated history $\mathcal{F}_t$ is*

$$\varphi_{\mathcal{C}}(t) = \sum_{i=1}^{N} \tilde{\varphi}_i(t)\mathbf{1}_{\{i\in\mathcal{C}(t)\}}. \tag{17}$$

Proof. We apply the smoothing Equation (4) in Section 2.1. For an $\mathcal{F}_t$-predictable process $Z(t)$, we write

$$\begin{aligned}\mathrm{E}\left[\int_{\mathbb{R}} Z(t)\Psi_{\mathcal{C}}(dt)\right] &= \sum_{i=1}^{N}\mathrm{E}\left[\int_{\mathbb{R}} Z(t)\mathbf{1}_{\{i\in\mathcal{C}(t)\}}\Phi_i(dt)\right] = \sum_{i=1}^{N}\mathrm{E}\left[\int_{\mathbb{R}} Z(t)\mathbf{1}_{\{i\in\mathcal{C}(t)\}}\tilde{\varphi}_i(t)\,dt\right]\\ &= \mathrm{E}\left[\int_{\mathbb{R}} Z(t)\left(\sum_{i=1}^{N}\mathbf{1}_{\{i\in\mathcal{C}(t)\}}\tilde{\varphi}_i(t)\right)dt\right] = \mathrm{E}\left[\int_{\mathbb{R}} Z(t)\varphi_{\mathcal{C}}(t)\,dt\right].\end{aligned}$$

The second equality follows from the smoothing Equation (4) because by Assumption 1, the stochastic intensity of Φ_i with respect $\mathcal{F}_t$ is $\tilde{\varphi}_i(t)$. The last equality is because of the process $Z(t)\mathbf{1}_{\{i\in\mathcal{C}(t)\}}$ being $\mathcal{F}_t$ predictable. This is the case because $\mathcal{C}(t)$ is causal and thus, $\mathcal{P}(\mathcal{F}_\cdot)$ measurable; therefore, $Z(t)\mathbf{1}_{\{i\in\mathcal{C}(t)\}}$ is also $\mathcal{P}(\mathcal{F}_\cdot)$ measurable for all i.

Because the identity holds for *any* $\mathcal{F}_t$-predictable process $Z(t)$, the converse implication in the smoothing formula establishes (17). □

Define now the policy $\mathcal{C}^*(t)$ as follows. At any point in time, rank the items by decreasing stochastic intensities $\tilde{\varphi}_i(t)$, and store the C highest; ties may be broken arbitrarily. Note that this policy maximizes the sum of the stochastic intensities of the items in memory

$$\mathcal{C}^*(t) := \operatorname*{arg\,max}_{\mathcal{C}\in\mathcal{P}_C(N)} \sum_{i\in\mathcal{C}} \tilde{\varphi}_i(t). \tag{18}$$

Theorem 3 (Optimal Causal Memory Management Policy). *Given a local memory system with capacity C fed by a family of stationary point processes, $\{\Phi_i : i = 1,\dots,N\}$ with aggregated natural history $\mathcal{F}_t$, satisfying Assumption 1. Then, for any causal policy $\mathcal{C}(t)$, its stationary hit rate $h_{\mathcal{C}}$ satisfies*

$$h_{\mathcal{C}} = \mathrm{E}[\varphi_{\mathcal{C}}(0)] \leqslant \mathrm{E}[\varphi_{\mathcal{C}^*}(0)] = h_{\mathcal{C}^*},$$

with $\varphi_{\mathcal{C}}(t)$ as in (17) and $\mathcal{C}^*$ as defined in (18). In addition, the hit process (16) is also simple, and the hit probability

$$H_{\mathcal{C}} := \mathrm{P}^0_\Phi(\Psi_{\mathcal{C}}(\{0\}) = 1)$$

is maximized by the policy $\mathcal{C}^*$.

Proof. Note from Equation (17) and the construction of $\mathcal{C}^*$ that for any realization,

$$\varphi_{\mathcal{C}}(0) = \sum_{i=1}^{N} \tilde{\varphi}_i(0)\mathbf{1}_{\{i\in\mathcal{C}\}} \leqslant \max_{\{i_1,\dots,i_C\}} \sum_{i\in\{i_1,\dots,i_C\}} \tilde{\varphi}_i(0) = \varphi_{\mathcal{C}^*}(0). \tag{19}$$

The first inequality follows by taking expectations on both sides with respect to the joint probability measure P of the arrival processes (nonsynchronized with arrivals).

To derive the second statement, we begin by observing that by definition (Equation (16)), $\Psi_{\mathcal{C}}(B) \leqslant \Phi(B)$ for any $B \in \mathcal{B}(\mathbb{R})$, so $\Psi_{\mathcal{C}}$ must also be simple. Therefore, it suffices to show that for any causal policy, $h_{\mathcal{C}} = \lambda^N H_{\mathcal{C}}$ (i.e., the hit rate is the total rate "thinned" by the hit probability). This is a natural property but nontrivial because the thinning is not independent of the arrival process.

From the superposition properties of stationary point processes (Brémaud 2020, example 7.2.11), under Assumption 1, the aggregated process is simple, and for any event A, we must have

$$\mathrm{P}^0_\Phi(A) = \sum_{i=1}^{N} \frac{\lambda_i}{\lambda^N}\mathrm{P}^0_{\Phi_i}(A). \tag{20}$$

To interpret the above, note that given a point in Φ occurring at time 0, λ_i/λ^N is the probability that this point comes from process i. Now, apply (20) to the event $A = \{\omega : \Psi_{\mathcal{C}}(\{0\}) = 1\}$:

$$H_{\mathcal{C}} = \mathrm{P}^0_{\Phi}(\Psi_{\mathcal{C}}(\{0\}) = 1) = \sum_{i=1}^{N} \frac{\lambda_i}{\lambda^N} \mathrm{P}^0_{\Phi_i}(\Psi_{\mathcal{C}}(\{0\}) = 1).$$

We now compute $\mathrm{P}^0_{\Phi_i}(\Psi_{\mathcal{C}}(\{0\}) = 1)$ as

$$\mathrm{P}^0_{\Phi_i}(\Psi_{\mathcal{C}}(\{0\}) = 1) = \mathrm{P}^0_{\Phi_i}\left(\bigcup_i \{\Phi_i(\{0\}) = 1, i \in \mathcal{C}(0)\}\right) = \mathrm{P}^0_{\Phi_i}(i \in \mathcal{C}(0))$$

because under $\mathrm{P}^0_{\Phi_i}$, $\mathrm{P}^0_{\Phi_i}(\{\Phi_i(\{0\}) = 1\}) = 1$ and $\mathrm{P}^0_{\Phi_i}(\{\Phi_j(\{0\}) = 1\}) = 0$ for $j \neq i$ because of Φ being simple. We conclude that

$$H_{\mathcal{C}} = \sum_{i=1}^{N} \frac{\lambda_i}{\lambda^N} \mathrm{P}^0_{\Phi_i}(i \in \mathcal{C}(0)). \tag{21}$$

Now, return to (17) and its expectation at $t = 0$:

$$h_{\mathcal{C}} = \mathrm{E}\left[\sum_{i=1}^{N} \tilde{\varphi}_i(0) \mathbf{1}_{\{i \in \mathcal{C}(0)\}}\right] = \sum_{i=1}^{N} \mathrm{E}[\tilde{\varphi}_i(0) \mathbf{1}_{\{i \in \mathcal{C}(0)\}}].$$

Because $\mathbf{1}_{\{i \in \mathcal{C}(t)\}}$ is F_t predictable for any causal policy, it follows from Papangelou's formula (Equation (5)) that

$$\mathrm{E}[\tilde{\varphi}_i(0) \mathbf{1}_{\{i \in \mathcal{C}(0)\}}] = \lambda_i \mathrm{E}^0_{\Phi_i}[\mathbf{1}_{\{i \in \mathcal{C}(0)\}}] = \lambda_i \mathrm{P}^0_{\Phi_i}(i \in \mathcal{C}(0)).$$

Summation over i gives together with (21), $h_{\mathcal{C}} = \lambda^N H_{\mathcal{C}}$ as claimed, and the result follows. □

Thus, Theorem 3 gives a clean formulation of what the optimal policy is: just keep track of the most likely arrivals in the system given the complete past history of the arrival stream. In particular, this generalizes Panigrahy et al. (2022) to general (possibly correlated) superpositions that provided that they satisfy Assumption 1.

3.1. The Case of Independent Request Streams

In the preceding analysis, computing the optimal policy requires knowledge of the stochastic intensities $\{\tilde{\varphi}_i(t) : i = 1, \ldots, N\}$ with respect to the aggregated history. This may require the knowledge of the underlying correlations. This is greatly simplified if the request processes are independent.

Consider now a local memory system with capacity C fed by *N-independent* stationary request processes Φ_i with average intensities $\lambda_i > 0$ as before (in decreasing order). Again, let $\mathcal{F}^{(i)}_t$ be the natural history of the ith process, and let $\mathcal{F}_t$ be the aggregated history. Again, let P^0_{Φ} denote the Palm probability of the superposition process, and let $\mathrm{P}^0_{\Phi_i}$ be the Palm probability of the ith process.

The superposition of independent simple processes is simple (Baccelli and Brémaud 2013, property 1.1.1), and besides Equation (20), we also have (Brémaud 2020, example 7.2.8)

$$\mathrm{P}^0_{\Phi_i}(\Phi_1 \in \Gamma_1, \ldots, \Phi_N \in \Gamma_N) = \mathrm{P}^0_{\Phi_i}(\Phi_i \in \Gamma_i) \prod_{j \neq i} \mathrm{P}(\Phi_j \in \Gamma_j) \tag{22}$$

for any $\Gamma_i \in \mathcal{M}^{\#}(\mathbb{R})$ and $i = 1, \ldots, N$.

The interpretation of Equation (22) is the following; in order to compute Palm probabilities given that the point comes from process i, we must use the Palm probability for process i and the stationary probability for any other $j \neq i$, and the processes remain independent.

This, in turn, allows us to prove the following property.

Lemma 3. *If $\Phi = \sum_i \Phi_i$ is the superposition process, then $\mathcal{F}_t$ is a history of Φ. Moreover, $\varphi_i(t)$ is a stochastic intensity for process i with respect to the enlarged history $\mathcal{F}_t$, and the total stochastic intensity of Φ is simply $\varphi(t) = \sum_{i=1}^{N} \varphi_i(t)$.*

Proof. First, $\Phi((a, b])$ is $\mathcal{F}_t$ measurable for all $a < b \leqslant t$ because the sum function is measurable and $\mathcal{F}^{(i)}_t \subset \mathcal{F}_t$. Second, $\varphi_i(t)$ is the stochastic intensity of Φ_i with respect to the shared history $\mathcal{F}_t$ because the conditional expectation given $\mathcal{F}^{(i)}_t$ coincides with the conditional expectation given $\mathcal{F}_t$ for any random variable independent of $\mathcal{F}^{(j)}_t$ with $j \neq i$. This fact is, in turn, a consequence of the independence of the filtrations $\{\mathcal{F}^{(i)}_t\}_{i=1}^{N}$ and that $\mathcal{F}_t$ is generated by events of the form $\bigcap_{i=1}^{N} A_i$ with $A_i \in \mathcal{F}^{(i)}_t$, $i = 1, \ldots, N$. Finally, that $\varphi(t)$ is the stochastic intensity of $\Phi(t)$ with respect to $\mathcal{F}_t$ follows immediately by linearity of conditional expectation. □

The above lemma guarantees that the stochastic intensity of process i is not altered by superposition with other independent processes; there is no crossinformation between the processes. In particular, Assumption 1 is satisfied, and therefore, we can replace $\tilde{\varphi}_i(t)$ by $\varphi_i(t)$ in Theorem 3 to obtain Theorem 4.

Theorem 4 (Optimal Causal Memory Management Policy, Independent Case). *Given a local memory system with capacity C fed by independent stationary point processes, $\{\Phi_i : i = 1, \dots, N\}$ with individual stochastic intensities $\varphi_i(t)$. Then, the optimal causal policy $\mathcal{C}^*$ corresponds to the following.*

- *At any point in time, rank the items in decreasing order of their individual stochastic intensities $\varphi_i(t)$.*
- *Store in memory the first C objects in the ranking.*

Any other causal policy $\mathcal{C}$ will satisfy $h_{\mathcal{C}} \leqslant h_{\mathcal{C}^*}$ and $H_{\mathcal{C}} \leqslant H_{\mathcal{C}^*}$.

The above theorem characterizes the structure of the optimal *causal* policy for any superposition of *independent* request processes (i.e., where no additional information is available about the future other than the natural history of the requests and where there is no crossinformation between them). If the processes are *correlated* among them, this statement will not be true in general, and we will have to keep track of the shared stochastic intensity with respect to the common history $\tilde{\varphi}_i(t)$.

We now analyze some examples where the above policy can be further characterized. The simplest is the Poisson case.

Example 3 (Poisson Arrivals). In case all processes Φ_i are Poisson with $\lambda_1 \geqslant \dots \geqslant \lambda_N$, $\varphi_i(t) \equiv \lambda_i$. By Theorem 3, the optimal policy $\mathcal{C}^*$ is the *static policy* that stores the C objects with higher (average) intensities at all times. This is, of course, connected to the total independence property of Poisson processes, and it is also well known in the caching literature (see, e.g., Garetto et al. 2016) under the name independent reference model.

Example 4 (Renewal Arrivals). In the more general class of renewal arrival processes, by Equation (9), $\varphi_i(t) = \eta_i(t - \tau_i^-(t))$. In this case, the optimal policy can be recast in the following way.

- Rank all contents in decreasing order of the *current interval hazard rates*, $\eta_i(t - \tau_i^-(t))$.
- Store in memory the first C objects in the ranking.

As we can see, the role of the hazard rates is crucial as already identified in Ferragut et al. (2016) for timer-based caching. Different monotonicity assumptions on these hazard rates lead to completely different optimal policies as we shall see in Section 6.

An issue with the optimal policy is that its main performance metric, the hit rate, cannot be computed exactly except in some special cases. We now derive an asymptotic result that characterizes the optimal policy for large-scale systems and allows us to calculate asymptotic performance limits.

4. Large-Scale Analysis of the Optimal Causal Policy

In this section, we will present results concerning the asymptotic behavior of the optimal policy in a large-scale regime, where both the system load and the memory size scale appropriately to infinity. For this purpose, we need to introduce more structure into the problem; in particular, we will assume that requests for different items come from independent processes from a common *scale family* in the following sense.

Consider a base process with unit intensity Φ_0, natural history $\mathcal{F}_t^{(0)}$, and stochastic intensity $\varphi_0(t)$ with respect to $\mathcal{F}_t^{(0)}$. In particular, $E[\varphi_0(t)] = 1$. For our asymptotic analysis, we specify our scale family as satisfying the following.

Assumption 2. *The request processes Φ_i for items $i \in \{1, \dots, N\}$ are independent with stochastic intensities*

$$\varphi_i(t) \overset{d}{=} \lambda_i \varphi_0(\lambda_i t),$$

where λ_i is the process average request rate. Without loss of generality, we will assume that these rates are decreasing (i.e., $\lambda_1 \geqslant \dots \geqslant \lambda_N > 0$); equivalently, items are ordered in decreasing popularity.

A construction that generates processes Φ_i satisfying this assumption is to take independent copies of Φ_0 and set $\Phi_i(B) \overset{d}{=} \Phi_0(\lambda_i B)$ (i.e., rescale the time axis).

It is direct from Definition 6 and Assumption 2 that if $X_i = \varphi_i(0)$ is the observed stochastic intensity for process i, then the following scaling relationships hold:

$$G^{(i)}(x) := \mathrm{P}(X_i \leqslant x) = \mathrm{P}(\varphi_0(0) \leqslant x/\lambda_i) = G(x/\lambda_i); \tag{23a}$$

$$G_0^{(i)}(x) := \mathrm{P}_{\Phi_i}^0(X_i \leqslant x) = \mathrm{P}_{\Phi_0}^0(\varphi_0(0) \leqslant x/\lambda_i) = G_0(x/\lambda_i). \tag{23b}$$

Example 5 (Scaled Renewal Process). A simple example where all of the above assumptions hold is when the base process is renewal with some base distribution $F_0(t)$ with a mean of one. In this case, the interarrival distribution of the ith process is

$$F_0^{(i)}(t) = F_0(\lambda_i t).$$

We can also express the age distribution $F^{(i)}(t)$ and hazard rate function $\eta^{(i)}(t)$ for the ith request process in terms of the base case $F(t)$, $\eta(t)$:

$$F^{(i)}(t) = \lambda_i \int_0^t (1 - F_0(\lambda_i s))ds = \int_0^{\lambda_i t} (1 - F_0(u))du = F(\lambda_i t);$$
$$\eta^{(i)}(t) = \frac{f_0^{(i)}(t)}{1 - F_0^{(i)}(t)} = \frac{\lambda_i f_0(\lambda_i t)}{1 - F_0(\lambda_i t)} = \lambda_i \eta(\lambda_i t).$$

4.1. Threshold Characterization of the Optimal Policy

Under Assumption 2, the optimal policy of Theorem 4 can be recast as follows; at any given time, we have a sample of random variables X_i, $i = 1, \dots, N$, each representing the observed stochastic intensity (for the renewal case, hazard rate since the last request) of the ith request process. Because the processes are independent, the X_i's are independent but *nonidentically* distributed; in fact, $X_i \sim G^{(i)}$.

According to Theorem 4, an item will be stored in memory if and only if its OSI is one of the C highest. An alternative way to cast this optimal policy is to consider the *empirical distribution* of the observed stochastic intensities defined by

$$\hat{G}_N(x) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\{X_i \leqslant x\}}; \tag{24}$$

then, object i is locally stored if and only if $\hat{G}_N(X_i) \geqslant \frac{N-C}{N} = 1 - \frac{C}{N}$.

Introducing the *quantile function* (inverse of the empirical distribution)

$$\hat{Q}_N(p) := \inf\{x : \hat{G}_N(x) \geqslant p\}, \quad p \in [0, 1], \tag{25}$$

the condition for storing item i in memory at a certain time may be expressed as

$$X_i \geqslant \hat{\theta}_N := \hat{Q}_N\left(1 - \frac{C}{N}\right). \tag{26}$$

The random variable $\hat{\theta}_N$ acts as a *threshold* in the OSI that determines the items to optimally store in local memory. Equivalently, $\hat{\theta}_N = Y_{N-C}$, where $\{Y_i : i = 1, \dots, N\}$ are the order statistics of the sample $\{X_i, i = 1, \dots, N\}$.

We will pursue asymptotic results for a system with a very large catalog size N. If we can find a suitable limit for the empirical distribution $\hat{G}_N(x)$ as $N \to \infty$ and if the memory size scales linearly with N such that $\frac{C}{N} \to_N c$, then the quantile $\hat{\theta}_N$ should approach a limit (i.e., the large-scale behavior should resemble a policy with a fixed threshold in OSI).

Because the X_i's are not identically distributed, we cannot use classical Glivenko–Cantelli arguments for empirical distribution convergence. Nevertheless, we will show that a suitable limit arises under Assumption 2 if the system load as defined by the request intensities also scales appropriately with N.

4.2. Scaling of the Request Intensities

We will construct a sequence of systems indexed by N, the number of arrival streams or in other words, items in its catalog. Denote by $\{\lambda_i^{(N)}\}_{i=1}^N$ the arrival rates of the system of size N, with the above convention that $\lambda_1^{(N)} \geqslant \cdots \geqslant \lambda_N^{(N)} > 0$. We may think of these points as a discrete measure on the axis $\lambda \in \mathbb{R}_+$ of possible process intensities. Normalizing this measure to total unit mass, we may write its distribution function

$$L_N(\lambda) := \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\{\lambda_i^{(N)} \leqslant \lambda\}}. \tag{27}$$

The total arrival rate for system of size N can then be expressed as

$$\lambda^N = \sum_i \lambda_i^{(N)} = N \int_0^\infty \lambda L_N(d\lambda).$$

Our limit theorems will assume that this family of discrete distributions has a limit with $N \to \infty$.

Assumption 3. *There exists a fixed distribution L with no atoms at $\lambda = 0$ such that $L_N \Rightarrow L$ when $N \to \infty$. Here, $\Rightarrow$ denotes usual weak convergence of probability distributions.*

To gain some intuition on the above condition, we turn to the following important example.

Example 6 (Scaling for Zipf Popularities). A widely used model of popularity among different items is the so-called Zipf distribution with the request rates $\lambda_i \propto i^{-\beta}$, where $\beta \geqslant 0$ is the tail parameter of the Zipf law. Note that $\beta = 0$ corresponds to a uniform distribution, whereas as for large β, relative popularities decay very fast.

In order to model this in our setting, we can use the following arrival rates for the Nth system:

$$\lambda_i^{(N)} = \left(\frac{N}{i}\right)^{\beta}$$

with $\beta \geqslant 0$. Under this scaling, the least popular object has intensity 1 for all N. As N grows, larger intensities are included in the mix. Now, for any $\lambda > 1$, we have

$$1 - L_N(\lambda) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\left\{\left(\frac{N}{i}\right)^{\beta} > \lambda\right\}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\left\{i < \frac{N}{\lambda^{1/\beta}}\right\}} = \frac{1}{N}\left\lfloor \frac{N}{\lambda^{1/\beta}} \right\rfloor \underset{N\to\infty}{\to} \lambda^{-1/\beta}.$$

Because the above convergence is point wise and the limit is continuous, we have $L_N(\lambda) \Rightarrow L(\lambda)$ given by

$$L(\lambda) = 1 - \lambda^{-1/\beta} \quad \text{for } \lambda \geqslant 1. \tag{28}$$

In the limit, the popularities follow a continuous distribution over $[1, \infty)$: namely, a standard Pareto distribution with tail parameter $1/\beta$.

If $\beta \geqslant 1$ (i.e., the popularities are light tailed), some objects are extremely more popular than others; in this case, L does not have a finite mean. If instead, $0 < \beta < 1$, where popularities are heavy tailed and thus, more homogeneous, L has finite mean $1/(1-\beta)$. For $\beta = 0$, the system degenerates into every object having the same popularity, and thus, L is the step function at $\lambda = 1$.

It is also worth observing that the total arrival rate of the Nth system satisfies

$$\lambda^N = \sum_{i=1}^{N} \lambda_i^{(N)} = N^{\beta} \sum_{i=1}^{N} \frac{1}{i^{\beta}} =: N^{\beta} S_N(\beta),$$

where $S_N(\beta)$ is the generalized harmonic series partial sum. Using the well-known equivalents for this series, we have that

$$\lambda^N = \begin{cases} O(N^{\beta}) & \text{if } \beta > 1, \\ O(N \log N) & \text{if } \beta = 1, \\ O(N) & \text{if } \beta < 1. \end{cases}$$

In particular, with our scaling, the total arrival rate $\lambda^N \to \infty$ as $N \to \infty$, albeit at different rates depending on the tail parameter β.

4.3. Asymptotic Behavior of the Optimal Causal Policy

We now return to our family of systems indexed by N with independent request streams coming from a common scale family with base distribution F_0 (Assumption 2) and incorporate the scaling in Assumption 3 on the arrival rates $\{\lambda_i^{(N)}\}_{i=1}^N$. Namely, their empirical distribution $L_N(\lambda)$ in (27) has a weak limit $L(\lambda)$.

If we look at each system at a fixed time $t = 0$, we obtain a sample of the current observed stochastic intensities $\{X_i^{(N)} : i = 1, \ldots, N\}$. Considered collectively that for all $N \geqslant 1$, these random variables constitute a *triangular array*; without loss of generality, we may assume that they are all defined in a common probability space $(\Omega, \mathcal{F}, \mathrm{P})$.

For each N, we can define the random function $\hat{G}_N$ by Equation (24). The main result below concerns the asymptotic behavior of these empirical stochastic intensity distributions.

Theorem 5. *Consider a family of local memory systems indexed by N with request processes satisfying Assumption 2 and with intensities $\{\lambda_i^{(N)}\}_{i=1}^N$ satisfying Assumption 3. Then,*

$$\hat{G}_N \underset{N\to\infty}{\Rightarrow} G_{\infty} \quad \mathrm{P}\text{ a.s.},$$

where the function G_∞ is given by

$$G_\infty(x) := \int_0^\infty G(x/\lambda)L(d\lambda). \tag{29}$$

Moreover, assume that the memory size of the Nth system satisfies $\frac{C_N}{N} \to_N c$ with $0 < c \leqslant 1$.

Then, if $1-c$ is a continuity point of the quantile function $Q_\infty = G_\infty^{-1}$, the random threshold $\hat{\theta}_N$ defined by Equation (26) *converges* P *almost surely to $\theta^* = Q_\infty(1-c)$.

Proof. The proof begins by computing the expected value of the random function $\hat{G}_N$:

$$\overline{G}_N(x) := \mathrm{E}[\hat{G}_N(x)] = \mathrm{E}\left[\frac{1}{N}\sum_{i=1}^N \mathbf{1}_{\{X_i^{(N)} \leqslant x\}}\right] = \frac{1}{N}\sum_{i=1}^N G^{(i)}(x) = \frac{1}{N}\sum_{i=1}^N G\left(\frac{x}{\lambda_i^{(N)}}\right), \tag{30}$$

where we have applied the scaling in (23a). $\overline{G}_N$ is a (deterministic) distribution function in $x \geqslant 0$. Using the definition of L_N, we can rewrite the above as

$$\overline{G}_N(x) = \int_0^\infty G(x/\lambda)L_N(d\lambda).$$

We first show that as distribution functions, $\overline{G}_N \Rightarrow G_\infty$ as $N \to \infty$. To do so, it is convenient to interpret (30) as follows; consider a pair (X, Λ_N) of independent random variables in $\mathbb{R}_+$ with respective distribution functions $G(x) = P(X \leqslant x)$ and $L_N(\lambda) = P(\Lambda_N \leqslant \lambda)$. Then, $\overline{G}_N(x)$ is the distribution of the product $Z_N = X\Lambda_N$. Indeed,

$$P(Z_N \leqslant x) = \sum_{n=1}^N P(X\Lambda_N \leqslant x | \Lambda_N = \lambda_i^{(N)})P(\Lambda_N = \lambda_i^{(N)}) = \frac{1}{N}\sum_{n=1}^N P(X \leqslant x/\lambda_i^{(N)}) = \overline{G}_N(x).$$

Now, consider the limit in the distribution of the pair (X, Λ_N). Because of their independence and because $L_N \Rightarrow L$, the limit corresponds to the distribution of (X, Λ), a vector of independent random variables with marginal distributions $G(x)$ and $L(\lambda)$. By continuity of the map $(x, \lambda) \longmapsto x\lambda$, we conclude that $Z_N = X\Lambda_N \xrightarrow{d} Z = X\Lambda$. It remains to compute the distribution function of the latter product:

$$P(Z \leqslant x) = \int_{\mathbb{R}_+^2} \mathbf{1}_{\{\xi\lambda \leqslant x\}} L(d\lambda)G(d\xi) = \int_0^\infty L(d\lambda)\int_0^\infty \mathbf{1}_{\{\xi \leqslant x/\lambda\}} G(d\xi) = \int_0^\infty G(x/\lambda)L(d\lambda).$$

We conclude that $\overline{G}_N \Rightarrow G_\infty$ as $N \to \infty$. Equivalently, we have point-wise convergence $\overline{G}_N(x) \to G_\infty(x)$ at any continuity point $x \geqslant 0$ of G_∞.

To relate the mean function $\overline{G}_N(x)$ to the stochastic one $\hat{G}_N(x)$, we now resort to Shorack (1979, theorem 2.1), a generalization of the classical Glivenko–Cantelli theorem for empirical distributions for random variables that are not identically distributed. The theorem states that in the probability space $(\Omega, \mathcal{F}, \mathrm{P})$ where the triangular array $\{X_i^{(N)} : i = 1, \ldots, N, N \geqslant 1\}$ is defined, we have

$$\left\|\hat{G}_N - \overline{G}_N\right\|_\infty \underset{N\to\infty}{\longrightarrow} 0 \quad \mathrm{P\ a.s.} \tag{31}$$

In particular, with P probability 1, $|\hat{G}_N(x) - \overline{G}_N(x)| \to 0$ as $N \to \infty$ for all $x \geqslant 0$. Combining this with the previously obtained point-wise convergence of $\overline{G}_N(x)$, we conclude that P almost surely $\hat{G}_N(x) \to G_\infty(x)$ at any continuity point $x \geqslant 0$ of G_∞, and thus, $\hat{G}_N \Rightarrow G_\infty$.

Finally, convergence of $\hat{\theta}_N$ now follows from the fact that convergence in distribution implies convergence of quantiles; specifically (van der Vaart 1998, lemma 21.2), $\hat{G}_N \Rightarrow G_\infty$ is equivalent to $\hat{Q}_N(p) \to Q_\infty(p)$ for all continuity points $p \in [0,1]$ of $Q_\infty = G_\infty^{-1}$. See also Proposition A.2 in Appendix A.1. □

We have thus established that the optimal causal local memory policy converges in the large-scale limit to a deterministic threshold policy in the observed stochastic intensities. Specifically, when memory scales as a fraction c of the catalog, a large-scale system should store at any given time the items whose current OSI exceeds a threshold θ^* chosen as the $1-c$ quantile of the limit distribution G_∞ in (29).

5. Asymptotic Optimal Performance

In this section, we analyze the *performance* achieved by the optimal policy in the large-scale limit. As discussed in Section 3, performance in local memory systems is measured by the *hit probability* H_C or equivalently, the *hit rate* $h_C = \lambda^N H_C$.

We will find it more convenient to derive expressions for the complementary *miss probability* $M_{\mathcal{C}} = 1 - H_{\mathcal{C}}$ and the *miss rate* $m_{\mathcal{C}} = \lambda^N M_{\mathcal{C}} = \lambda^N - h_{\mathcal{C}}$. Referring back to Equation (21), we can write the following expressions for these quantities in a system of size N.[3]

$$M_{\mathcal{C}}^{(N)} = \sum_{i=1}^{N} \frac{\lambda_i^{(N)}}{\lambda^N} \mathrm{P}_{\Phi_i}^0(i \notin \mathcal{C}(0)). \tag{32a}$$

$$m_{\mathcal{C}}^{(N)} = \sum_{i=1}^{N} \lambda_i^{(N)} \mathrm{P}_{\Phi_i}^0(i \notin \mathcal{C}(0)). \tag{32b}$$

The key challenge in the evaluation of the above formulas is to compute $\mathrm{P}_{\Phi_i}^0(i \notin \mathcal{C}(0))$ (i.e., the probability that an incoming arrival does not find its file in local memory). In the optimal policy $\mathcal{C}^*$ for fixed N, the determining condition is whether the observed stochastic intensity $X_i^{(N)}$ of the requested item finds itself among the C largest. To correctly evaluate this probability, however, we must recognize an asymmetry. The competing $X_j^{(N)}$, $j \neq i$ are being evaluated at a time *not* synchronized with their arrivals and thus, follow independent distributions $G(\cdot/\lambda_j^{(N)})$; for the item in question, we must use the distribution $G_0(\cdot/\lambda_i^{(N)})$ that applies to synchronized sampling of the stochastic intensity.

For this reason, to analyze the comparison, it is convenient to define the empirical distribution

$$\hat{G}_N^{(-i)}(x) = \frac{1}{N-1} \sum_{j \neq i} \mathbf{1}_{\{X_j^{(N)} \leqslant x\}} \tag{33}$$

of the stochastic intensities of nonrequested items and express the *miss* condition as follows:

$$\begin{aligned} i \notin \mathcal{C}(0) &\iff \sum_{j \neq i} \mathbf{1}_{\{X_j^{(N)} > X_i^{(N)}\}} \geqslant C \\ &\iff \sum_{j \neq i} \mathbf{1}_{\{X_j^{(N)} \leqslant X_i^{(N)}\}} \leqslant N - 1 - C. \\ &\iff \hat{G}_N^{(-i)}\left(X_i^{(N)}\right) \leqslant 1 - \frac{C}{N-1} =: p_N. \end{aligned}$$

The above equivalence assumes that there are no ties in the comparison of the OSIs; to simplify the analysis to follow, we will make this a standing assumption in this section.

Assumption 4. *For each* $N \geqslant 1$ *and* $1 \leqslant i \neq j \leqslant N$, $\mathrm{P}_{\Phi}^0(X_i^{(N)} = X_j^{(N)}) = 0$.

Returning to (32b), we may now express the performance criterion of the optimal policy as

$$m_{\mathcal{C}^*}^{(N)} = \sum_{i=1}^{N} \lambda_i^{(N)} \mathrm{P}_{\Phi_i}^0(\hat{G}_N^{(-i)}(X_i^{(N)}) \leqslant p_N) = \sum_{i=1}^{N} \lambda_i^{(N)} \mathrm{P}_{\Phi_i}^0(X_i^{(N)} \leqslant \hat{Q}_N^{(-i)}(p_N)), \tag{34}$$

where $\hat{Q}_N^{(-i)}$ is the inverse (quantile) function of $\hat{G}_N^{(-i)}$.

We now outline informally the essence of the analysis that follows. Because the empirical distribution $\hat{G}_N^{(-i)}$ corresponds to $N-1$ intensities, which are sampled at a nonsynchronized point, its asymptotic behavior under the assumed scaling should follow the conclusions of Theorem 5 (i.e., converge to the distribution G_∞). Under $C/N \to c$, we have $p_N \to 1 - c$, so we should have $\hat{Q}_N^{(-i)}(p_N) \to Q_\infty(1-c) = \theta^*$ as $N \to \infty$.

This leads us to consider the approximate formula

$$m_{\mathcal{C}^*}^{(N)} \approx \sum_{i=1}^{N} \lambda_i^{(N)} \mathrm{P}_{\Phi_i}^0(X_i^{(N)} \leqslant \theta^*) = \sum_{i=1}^{N} \lambda_i^{(N)} G_0(\theta^*/\lambda_i^{(N)});$$

in the last equality, we have invoked the distribution of $X_i^{(N)}$ *synchronized* with the arrival under the Palm probability $\mathrm{P}_{\Phi_i}^0$. Of course, the approximation is not a rigorous step. We have taken the limit in the quantile but not in the rest of Equation (34); nevertheless, it will help us arrive at the right conjecture.

Invoking the distribution of intensities in (27), we may express the approximation as

$$\frac{m_{\mathcal{C}^*}^{(N)}}{N} \approx \frac{1}{N}\sum_{i=1}^{N} \lambda_i^{(N)} G_0(\theta^*/\lambda_i^{(N)}) = \int_0^\infty \lambda G_0(\theta^*/\lambda) L_N(d\lambda). \tag{35}$$

Under Assumption 3, $L_N \Rightarrow L$. The integrand function $\lambda G_0(\theta^*/\lambda)$ is bounded by θ^*, invoking (7); if it is continuous, the right-hand side will converge to the corresponding integral with the limit distribution $L(\lambda)$. This formula for the asymptotic optimal miss rate is the main conjecture that we will prove in Theorem 6 below, making rigorous the preceding approximate reasoning.

We will require some additional technical conditions.

- For continuity of $\lambda G_0(\theta^*/\lambda)$, it will suffice to guarantee it at the atoms of the measure $L(d\lambda)$ (i.e., the discontinuities of $L(\lambda)$); denote this set by D_L, and recall by Assumption 3 that $0 \notin D_L$. Analogously, denote by D_{G_0} the set of discontinuities of G_0, and denote by $D_L \cdot D_{G_0} := \{\lambda x : \lambda \in D_L, x \in D_{G_0}\}$; we will require for our limit theorem that $\theta^* \notin D_L \cdot D_{G_0}$.
- To carry out the limit result for $m_{\mathcal{C}^*}$ in (34), we will treat separately two cases in regard to the *uniform integrability* of the family of popularity distributions $\{L_N\}_{N\geqslant 1}$. The family is uniformly integrable if

$$\forall \epsilon > 0, \exists K \geqslant 0 \text{ such that } \sup_{N\geqslant 1} \frac{1}{N} \sum_{\lambda_i^{(N)} \geqslant K} \lambda_i^{(N)} \leqslant \epsilon.$$

Uniform integrability is a standard assumption (see, e.g., Billingsley 1999, section 1.3) that connects convergence in distribution with convergence of the first moment. In particular, if $L_N \Rightarrow L$ and the $\{L_N\}$'s are uniformly integrable, then

$$\int_0^\infty \lambda L_N(d\lambda) \xrightarrow[N\to\infty]{} \int_0^\infty \lambda L(d\lambda) < \infty. \tag{36}$$

A partial converse also holds; if $L_N \Rightarrow L$, with first moments satisfying (36), the family must be uniformly integrable.

Example 7. The Zipf family considered in Example 6 has continuous limit L and thus, satisfies all necessary continuity assumptions. Furthermore, the family L_N is uniformly integrable for $\beta < 1$; for $\beta \geqslant 1$, the limit distribution has infinite mean.

We are now ready to state our main performance result.

Theorem 6 (Asymptotic Miss Rate—Uniformly Integrable Case). *Suppose $\{L_N\}_{N\geqslant 1}$ is uniformly integrable. Let $c \in (0,1]$ be such that $1-c$ is a continuity point of Q_∞ and that $\theta^* := Q_\infty(1-c) \notin D_L \cdot D_{G_0}$. If $C/N \to c$ as $N \to \infty$, then*

$$\lim_{N\to\infty} \frac{m_{\mathcal{C}^*}^{(N)}}{N} = \int_0^\infty \lambda G_0\left(\frac{\theta^*}{\lambda}\right) L(d\lambda). \tag{37}$$

The proof is given in Appendix A.

The result states that for a uniformly integrable scaling of popularity distributions, under some regularity assumptions, the miss rate of the optimal policy scales linearly with N with a proportionality constant given by the integral in (37), where the distribution function G_0 of the OSI under synchronous sampling appears explicitly. On the other hand, the nonsynchronous distribution G of the OSI also influences the formula because it determines the distribution G_∞ in (29), whose quantile is the asymptotic threshold θ^*.

Corollary 1 (Asymptotic Miss Probability—Uniformly Integrable Case). *Under the same hypothesis as Theorem 6, the optimal miss probability satisfies*

$$\lim_{N\to\infty} M_{\mathcal{C}^*}^{(N)} = \frac{\int_0^\infty \lambda G_0\left(\frac{\theta^*}{\lambda}\right) L(d\lambda)}{\int_0^\infty \lambda L(d\lambda)}. \tag{38}$$

Proof. Note that

$$M_{\mathcal{C}^*}^{(N)} = \frac{m_{\mathcal{C}^*}^{(N)}}{\lambda^N} = \frac{m_{\mathcal{C}^*}^{(N)}/N}{\lambda^N/N};$$

the limit of the numerator is given in Theorem 6. For the denominator, we use (36) because $\frac{\lambda^N}{N} = \int_0^\infty \lambda L_N(d\lambda)$, and we have uniform integrability. □

Note regarding Equation (38) that the numerator is bounded by θ^*; indeed, we have $\lambda G_0(\theta^*/\lambda) \leqslant \theta^*$ from Lemma 1, and L has unit mass. As the first moment of this distribution (in the denominator) becomes larger, the miss probability is smaller. This suggests that for L with infinite first moment, the miss probability will be zero. This is indeed true, but we must provide a separate argument because uniform integrability will not hold in this case. Our strategy will be to show that a suboptimal policy (the static one with $\mathcal{C}^s = \{1,\dots,C\}$) already achieves vanishing asymptotic miss probability.

Theorem 7 (Asymptotic Miss Probability—Nonintegrable Case). *Suppose that* $\int_0^\infty \lambda L(d\lambda) = \infty$. *Assume that* $C/N \to c \in (0,1]$ *as* $N \to \infty$. *Then*,

$$M_{\mathcal{C}^*}^{(N)} \leqslant M_{\mathcal{C}^s}^{(N)} \underset{N\to\infty}{\longrightarrow} 0. \tag{39}$$

Proof. The first inequality is immediate by optimality of $\mathcal{C}^*$. So, we focus on the static policy, where misses are certain for $i > C$; therefore,

$$M_{\mathcal{C}^s}^{(N)} = \sum_{i=C+1}^{N} \frac{\lambda_i^{(N)}}{\lambda^N} = \frac{\frac{1}{N}\sum_{i=C+1}^{N} \lambda_i^{(N)}}{\lambda^N/N} = \frac{\int_0^{\alpha_N} \lambda L_N(d\lambda)}{\int_0^\infty \lambda L_N(d\lambda)}, \tag{40}$$

where $\alpha_N = L_N^{-1}(p_N)$ is the p_Nth quantile of L_N and $p_N = 1 - \frac{C}{N}$.

Because $L_N \Rightarrow L$ and $p_N \to 1 - c < 1$, we have that $A := \sup_N \alpha_N < \infty$. Then,

$$\limsup_{N\to\infty} \int_0^{\alpha_N} \lambda L_N(d\lambda) \leqslant \lim_{N\to\infty} \int_0^{A} \lambda L_N(d\lambda) = \int_0^{A} \lambda L(d\lambda) < \infty. \tag{41}$$

For the denominator, we have (Billingsley 1999, theorem 3.4)

$$\liminf_{N\to\infty} \int_0^{\infty} \lambda L_N(d\lambda) \geqslant \int_0^{\infty} \lambda L(d\lambda) = \infty. \tag{42}$$

Together, (41) and (42) imply (39). □

As a comment on the above results, we note the following. When the distribution of item intensities is *not* uniformly integrable as, for example, in the Zipf case with $\beta \geqslant 1$, items with the largest intensities dominate the rest; thus, it is natural that the static policy would achieve asymptotic optimality and perfect performance regardless of the interarrival distribution. Note, however, that such convergence may be slow; we refer to Section 8 for examples.

In the more interesting case of uniform integrability with less disparate popularities, *any* causal memory management policy will incur some performance penalty; the minimum miss probability achieved by the optimal policy can be computed explicitly from the interarrival distribution characteristics and the storage fraction c.

6. Threshold Policies and Their Timer-Based Counterparts

In Section 4, we found that the optimal causal policy could be expressed in terms of a threshold on the observed stochastic intensity. This threshold is stochastically varying but converges in our large-scale regime to a deterministic constant.

This motivates us to look at policies that are *defined* by a deterministic threshold (i.e., where one keeps in memory items with stochastic intensity larger than θ). An immediate observation is that in such policies, the memory constraint C would not be satisfied in a strong way at all times. Rather, we must replace it with the soft constraint where the *average memory occupation* is C. We now make this formal.

6.1. Threshold Policies and Memory Occupation

Consider a local memory system with independent request processes $\{\Phi_i : i = 1\dots N\}$, aggregated natural history $\mathcal{F}_t$, and stochastic intensities $\varphi_i(t)$. Consider the following threshold policy:

$$\mathcal{C}_\theta(t) = \{i : \varphi_i(t) > \theta\} \subset \mathcal{P}(N) \tag{43}$$

(i.e., the store in memory of all of the items that have current stochastic intensity above the threshold θ).

Because the Φ_i's are stationary, the intensities $\varphi_i(t)$ are also stationary, and thus, we have a stationary and causal policy. The memory occupation of this policy is now the stationary random process:

$$U_\theta(t) = \#\mathcal{C}_\theta(t) = \sum_{i=1}^{N} \mathbf{1}_{\{\varphi_i(t)>\theta\}}. \tag{44}$$

Its average can be computed as (using $t = 0$ as a sampling point)

$$\mathrm{E}[U_\theta(0)] = \mathrm{E}\left[\sum_{i=1}^{N} \mathbf{1}_{\{\varphi_i(0)>\theta\}}\right] = \sum_{i=1}^{N} \mathrm{P}(\varphi_i(0) > \theta). \tag{45}$$

Note that $\mathrm{E}[U_\theta(0)]$ is decreasing from N to zero as θ goes from zero to ∞. In order to have a fair comparison against a fixed memory constraint, define the threshold θ_C as

$$\theta_C := \inf\left\{\theta : \sum_{i=1}^{N} \mathrm{P}(\varphi_i(0) > \theta) \leqslant C\right\}. \tag{46}$$

Thus, θ_C is the $1 - C/N$ quantile of the average of the distributions of the stochastic intensities observed at time 0. From the right continuity of the distribution functions, it is easy to check that $\mathrm{E}[U_{\theta_C}(0)] \leqslant C$ with equality if $1 - C/N$ is a continuity point of the quantile function. We shall assume this henceforth to simplify exposition.

Because the arrival processes are independent, the random variables $\varphi_i(0)$ are independent, and thus, we have the following proposition.

Proposition 1. *Assume that the above quantile is exact in the sense that* $E[U_{\theta_C}(0)] = C$. *Take* $C = cN$ *with* $0 < c \leqslant 1$. *Then, for any* $\varepsilon > 0$,

$$\mathrm{P}\left(\left|\frac{U_{\theta_C}(0)}{N} - c\right| > \varepsilon\right) \leqslant 2e^{-2N\varepsilon^2},$$

and therefore, $U_{\theta_C}(0)/N \xrightarrow{\mathrm{P}} c$ *as* $N \to \infty$.

Proof. The proof follows from the Hoeffding inequality applied to the independent Bernoulli random variables $\{\mathbf{1}_{\{\varphi_i(0)>\theta\}}\}$. In fact, for any N and $\varepsilon > 0$,

$$\mathrm{P}\left(\left|\frac{U_{\theta_C}(0)}{N} - c\right| > \varepsilon\right) = \mathrm{P}\left(\left|\sum_{i=1}^{N} \mathbf{1}_{\{\varphi_i(0)>\theta_C\}} - \mathrm{E}\left[\sum_{i=1}^{N} \mathbf{1}_{\{\varphi_i(0)>\theta_C\}}\right]\right| > N\varepsilon\right) \leqslant 2e^{-2N\varepsilon^2}. \quad \square$$

From Proposition 1, we conclude that under independent request processes, the memory occupation of the threshold policy will not deviate much from a fixed memory policy for large N. We will use this to our advantage in Section 8.

6.2. Asymptotic Optimality of Threshold Policies

We focus now on the situation studied in Sections 4 and 5, where by Assumption 2, the request processes come from a common scale family. Furthermore, the traffic intensities satisfy Assumption 3.

In that case, $\varphi_i(0) = X_i$, which is the observed stochastic intensity with distribution $G^{(i)}(x)$. Therefore, θ_C is the solution to the equation

$$\sum_{i=1}^{N} [1 - G^{(i)}(\theta_C)] = C.$$

Invoking Equation (23a) for the scale family and rearranging terms, we have

$$\overline{G}_N(\theta_C) = \frac{1}{N}\sum_{i=1}^{N} G\left(\frac{\theta_C}{\lambda_i}\right) = 1 - \frac{C}{N}$$

with the notation introduced in (30); invoking the distribution L_N of traffic intensities λ_i, we may also write the above equation as

$$\int_0^\infty G\left(\frac{\theta_C}{\lambda}\right) L_N(d\lambda) = 1 - \frac{C}{N}. \tag{47}$$

We are now in position to prove the following.

Proposition 2. *Consider a family of local memory systems indexed by N with request processes satisfying Assumptions 2 and 3. Choose the memory size of the Nth system as $C_N = cN$ with $0 < c \leqslant 1$. If there exists a unique solution to θ^* satisfying*

$$G_\infty(\theta^*) = 1 - c,$$

then the sequence of thresholds $\theta_N := \theta_{C_N}$ defined by (47) satisfies $\theta_N \to \theta^*$.

The essential step of the proof is to show that $\overline{G}_N \Rightarrow G_\infty$; this part is identical to the proof of Theorem 5. The result then follows from the convergence of quantiles.

Let us now analyze the asymptotic performance; considering an object i, its miss probability is given by

$$\mathrm{P}^0_{\Phi_i}(i \notin \mathcal{C}_{\theta_N}(0)) = \mathrm{P}^0_{\Phi_i}(X_i \leqslant \theta_N) = G_0^{(i)}(\theta_N).$$

Therefore, the total miss probability for system N is given by

$$M_{\theta_N} = \sum_{i=1}^{N} \frac{\lambda_i^{(N)}}{\lambda^{(N)}} \mathrm{P}^0_{\Phi_i}(X_i \leqslant \theta_N) = \sum_{i=1}^{N} \frac{\lambda_i^{(N)}}{\lambda^{(N)}} G_0(\theta_N/\lambda_i^{(N)}) = \frac{\int_0^\infty \lambda G_0(\theta_N/\lambda) L_N(d\lambda)}{\int_0^\infty \lambda L_N(d\lambda)}.$$

We can then state a result analogous to Theorem 6 and Corollary 1 for the asymptotic performance.

Proposition 3. *Under the conditions of Theorem 6, the asymptotic miss probability satisfies*

$$\lim_{N\to\infty} M_{\theta_N} = \frac{\int_0^\infty \lambda G_0\left(\frac{\theta^*}{\lambda}\right) L(d\lambda)}{\int_0^\infty \lambda L(d\lambda)}.$$

The proof is substantially easier than in the case of Theorem 6 because the threshold θ_N is deterministic instead of depending on the state of the remaining processes.

Indeed, by Lemma 1 and the fact that $\theta_N \to \theta^*$, we see that the function $\lambda G_0(\theta_N/\lambda)$ is uniformly bounded for all N. Taking proper account of continuity as in Theorem 6, the numerator converges as stated. For the denominator, we invoke uniform integrability.

The main conclusion of this analysis is that under the above assumptions, a deterministic threshold policy with a soft memory constraint is asymptotically equivalent to the optimal causal policy derived in Section 3 both in terms of memory usage and in terms of performance.

We shall show now that there is a strong connection between threshold policies and timer-based ones.

6.3. Connection with Timer-Based Policies: Monotone Hazard Rates

To end this section, we would like to highlight a strong connection between threshold policies and *timer-based* ones in the case where the requests follow renewal processes. Timer-based or time-to-live caching has been a long-standing idea in local memory systems; here, each time a request for item i occurs, a timer is started, and the item is kept in memory up to the timer expiration. When a new request arrives, the timer is reset. TTL policies were first analyzed in terms of point process requests in Fofack et al. (2014), and the optimal timers under stationary requests were obtained in Ferragut et al. (2016).

Consider the policy defined in (43) with deterministic threshold $\theta_N = \theta_{C_N}$ from Equation (46). Assume that the incoming processes are renewal and that the interarrival hazard rates are strictly *monotonically decreasing* over the distribution domain (as in the Pareto parametric example). Assume for simplicity that the threshold θ_N lies in the range of the function $\eta_i(\cdot)$, and define

$$T_i^{(N)} = \eta_i^{-1}(\theta_N).$$

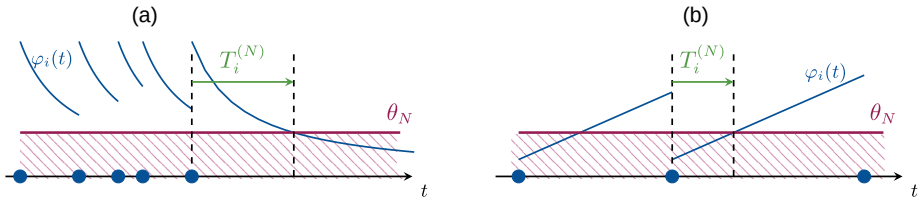
Then, because η_i is decreasing, we have

$$\varphi_i(t) > \theta_N \iff t < T_i^{(N)};$$

in panel (b) of Figure 4, we illustrate this condition. Consequently, in the case of monotonically decreasing hazard rates, the threshold policy is *exactly* equivalent to TTL caching.[4] This is consistent with the characterization of optimal timers in Ferragut et al. (2016). In addition, from the asymptotic optimality of threshold policies established above, we further conclude here that *TTL caching is asymptotically optimal*.

An analogous discussion is valid for monotonically *increasing* hazard rates; here, the likelihood of a subsequent request is *smallest* immediately after receiving one, and thus, caching is not useful. Instead, it was proposed in Ferragut et al. (2024) to remove the item and *prefetch* it at an appropriate later time. In fact, this strategy can also

Figure 4. Threshold Policies in Terms of Timers



Notes. (a) Timer-based caching for decreasing hazard rates. (b) Timer-based prefetching for increasing hazard rates.

be cast as a threshold policy by observing (see Figure 4) that

$$\varphi_i(t) > \theta_N \iff t > T_i^{(N)},$$

so the threshold policy removes the object from memory and recalls it after a time $T_i^{(N)}$; these timers coincide with the optimal timer prefetching policy from Ferragut et al. (2024). Once again, the asymptotical optimality of threshold policies allows us to further conclude that timer-based prefetching is asymptotically optimal for request processes with increasing hazard rates.

7. Extension to Heterogeneous Object Sizes

The preceding model assumes that objects have equal sizes; they all have the same impact on memory, and the natural performance criterion is the hit probability determined by whether the object can be (fully) recovered from local memory upon request. In practice, items can have different sizes and therefore, unequal memory cost. Also, one can consider in the performance the value of a *partial* recovery of the object. We now formalize this approach, which was first proposed in Panigrahy et al. (2022), and we show how the previous results can be extended to incorporate this case.

Consider a local memory system fed by a family of request processes $\{\Phi_i : i = 1, \dots, N\}$ with intensities λ_i, natural histories $\mathcal{F}_t^{(i)}$, and stochastic intensities $\varphi_i(t)$. Again, let Φ be their superposition, and let $\mathcal{F}_t$ be the aggregated history. Assume that each object i has a size $z_i > 0$ and that the system has total memory C.

A causal policy $\tilde{\mathcal{C}}(t)$ will now be an $\mathcal{F}_t$-predictable stochastic process $\tilde{\mathcal{C}} : \Omega \times \mathbb{R} \to [0,1]^N$, where $\tilde{\mathcal{C}}(t)$ represents the *fractions* of each file stored in the local memory. For feasibility, the policy should satisfy the memory constraint

$$\tilde{\mathcal{C}}(t) = (x_1, \dots, x_n) \Rightarrow \sum_i z_i x_i \leqslant C. \tag{48}$$

Thus, we are allowing *fractional caching* (i.e., only part of the item to be stored in memory).

Assume that whenever a request for item i arrives at time t, it can recover from the local memory $z_i x_i(t)$ and derives a *utility* from this amount of content $U_i(z_i x_i)$, where U_i is a concave function. In this scenario, the analogue of the hit process is the following *utility process*:

$$\Upsilon_{\tilde{\mathcal{C}}}(B) = \sum_{i=1}^{N} \int_B U_i(z_i x_i(t)) \Phi_i(dt). \tag{49}$$

The process $\Upsilon_{\tilde{\mathcal{C}}}$ is a measure-valued process that stores the utility derived from the local memory retrievals. For the special case where $U_i(z) = z$, we recover the *byte-counting* process in Panigrahy et al. (2022) because it keeps track of the amount of data recovered from the local memory in any time interval.

Now, let this local memory system be fed by a superposition of request processes $\{\Phi_i : i = 1, \dots, N\}$ satisfying Assumption 1. We have the analogue of Lemma 2.

Lemma 4. *Let $\tilde{\mathcal{C}}(t)$ be a causal policy. The stochastic intensity of the utility process (49) with respect to the aggregated history $\mathcal{F}_t$ is*

$$u_{\tilde{\mathcal{C}}}(t) = \sum_{i=1}^{N} \tilde{\varphi}_i(t) U_i(z_i x_i(t)). \tag{50}$$

The proof is similar to the proof of Lemma 2, changing the predictable process $\mathbf{1}_{\{i \in \mathcal{C}(t)\}}$ for the fractional caching version $U_i(z_i x_i(t))$, where $x_i(t)$ is predictable by assumption.

Consider now the policy $\tilde{\mathcal{C}}^*(t)$ defined by solving the following optimization problem:

$$\max_{x_i \in [0,1]} \sum_i \tilde{\varphi}_i(t) U_i(z_i x_i) \tag{51a}$$

$$\text{subject to } \sum_i z_i x_i \leqslant C. \tag{51b}$$

We then have the analogue of Theorem 3.

Theorem 8. *Given a local memory system with total memory C fed by a family of stationary point processes, $\{\Phi_i : i = 1, \dots, N\}$ with aggregated natural history $\mathcal{F}_t$, satisfying Assumption 1. Assume that the item sizes are $\{z_i : i = 1, \dots, N\}$ and that the system may store fractional items. Then, for any causal policy $\tilde{\mathcal{C}}(t)$, the stationary utility rate $u_{\tilde{\mathcal{C}}}$ satisfies*

$$u_{\tilde{\mathcal{C}}} = \mathrm{E}[u_{\tilde{\mathcal{C}}}(0)] \leqslant \mathrm{E}[u_{\tilde{\mathcal{C}}^*}(0)] = u_{\tilde{\mathcal{C}}^*}$$

with $u_{\mathcal{C}}(t)$ as in (50) and $\tilde{\mathcal{C}}^*$ as defined in (51).

The proof follows along the same lines of Theorem 3 by observing that the policy $\tilde{\mathcal{C}}^*$ satisfies Equations (51) at all times.

Example 8 (Maximizing the Byte Count). If we take $U_i(z) = z$ for all i, then we recover the byte-counting process of Panigrahy et al. (2022). In that case, the optimization problem defining $\tilde{\mathcal{C}}^*$ becomes a linear program, and its solution can be given explicitly in terms of the stochastic intensities $\tilde{\varphi}_i(t)$ as follows.

- Rank the contents in decreasing order of $\tilde{\varphi}_i(t)$.
- Choose the first K objects satisfying

$$\sum_{k=1}^{K-1} z_{i_k} < C \text{ and } \sum_{k=1}^{K} z_{i_k} > C,$$

where i_k denotes the ordering permutation of the first step.

- Take $x_{i_k}(t) = 1$ for $k = 1, \dots, K-1$ and $x_{i_K}(t) = \frac{C - \sum_{i=1}^{K-1} z_{i_k}}{z_{i_K}}$.

In other words, at all times, store the most likely upcoming requests as dictated by the stochastic intensity up to filling up your memory.

Note that the policy in Example 8 ranks the objects again in terms of the stochastic intensity independently of the object size. In particular, the optimal policy can be cast as a dynamic threshold policy, where the threshold is the stochastic intensity of the Kth object (i.e., the one that completes the memory allocation). Because of this fact, we think that the asymptotic results of Sections 5 and 6 can also be extended to this case under the appropriate scaling assumptions. This topic will be pursued in future work.

8. Parametric Examples and Simulations

In this section, we provide some interesting parametric examples to highlight the power of the results, in particular enabling us to compute sharp estimates for the performance of a local memory system. We must specify the interarrival distribution of our scale family of request processes and the distribution of popularities. We begin with the Pareto–Zipf combination, which was already introduced in examples above. We then analyze a further example with increasing hazard rates.

8.1. Pareto Interarrival Times and Zipf Popularities

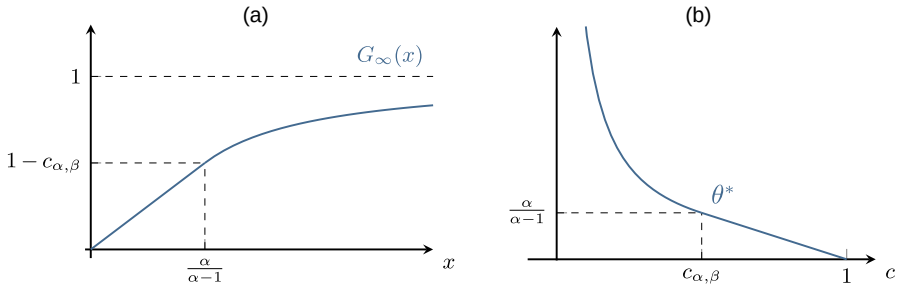
Consider first the Pareto interrequest times introduced in Section 2.3 with tail parameter α. As we already mentioned, this family represents bursty traffic with decreasing hazard rates. For the base (unit intensity) process, we have the following distribution for (nonsynchronized) observed stochastic intensity:

$$G(x) = \left(\frac{\alpha - 1}{\alpha} x\right)^{\alpha - 1} \text{ for } 0 \leqslant x \leqslant \frac{\alpha}{\alpha - 1}; \qquad G(x) = 1 \text{ for } x > \frac{\alpha}{\alpha - 1}.$$

Combine it with the Zipf popularities with tail parameter β introduced in Example 6 with limit distribution

$$L(\lambda) = 1 - \lambda^{-1/\beta} \text{ for } \lambda \geqslant 1.$$

Figure 5. Pareto Example with Zipf Popularities, $\alpha = 2$, and $\beta = 1$



Notes. (a) Limit distribution G_∞. (b) Threshold fixed point.

By appropriately integrating Equation (29), we can obtain the asymptotic distribution:

$$G_\infty(x) = \begin{cases} (1 - c_{\alpha,\beta})\left(\frac{\alpha-1}{\alpha}x\right)^{\alpha-1} & x \leqslant \dfrac{\alpha}{\alpha-1}; \\ 1 - c_{\alpha,\beta}\left(\frac{\alpha-1}{\alpha}x\right)^{-1/\beta} & x > \dfrac{\alpha}{\alpha-1}; \end{cases} \quad \text{where} \quad c_{\alpha,\beta} := \frac{(\alpha-1)\beta}{(\alpha-1)\beta+1}, \tag{52}$$

which is a relevant parameter in the following discussion. Note that $G_\infty(x)$ is continuous and takes the value $1 - c_{\alpha,\beta}$ at $x = \frac{\alpha}{\alpha-1}$. A depiction of $G_\infty(x)$ is given in Figure 5.

Because G_∞ is strictly increasing, we can now solve for the unique asymptotic threshold θ^* as a function of the memory size c. Imposing $G_\infty(\theta^*) = 1 - c$, we get the following result:

$$\theta^* = \begin{cases} \left(\dfrac{\alpha}{\alpha-1}\right)\left(\frac{c_{\alpha,\beta}}{c}\right)^{\beta} & c \leqslant c_{\alpha,\beta}, \\ \left(\dfrac{\alpha}{\alpha-1}\right)\left(\frac{1-c}{1-c_{\alpha,\beta}}\right)^{\frac{1}{\alpha-1}} & c \geqslant c_{\alpha,\beta}. \end{cases}$$

In the case $\beta < 1$, it is easy to see that the measures $\{L_N\}$ are uniformly integrable, and thus, we can compute the miss rate estimate from (37):

$$\int_0^\infty \lambda G_0(x/\lambda)L(d\lambda) = \begin{cases} (1 - c_{\alpha,\beta})\left(\frac{\alpha-1}{\alpha}x\right)^{\alpha} & x \leqslant \frac{\alpha}{\alpha-1}; \\ \frac{1}{1-\beta}\left[1 - \alpha\beta(1 - c_{\alpha,\beta})\left(\frac{\alpha-1}{\alpha}x\right)^{1-1/\beta}\right] & x \geqslant \frac{\alpha}{\alpha-1}. \end{cases}$$

Substituting the appropriate threshold and noting that $\int_0^\infty \lambda L(d\lambda) = \frac{1}{1-\beta}$, we reach the following result for the asymptotic miss probability:

$$M = \begin{cases} 1 - \alpha\beta(1 - c_{\alpha,\beta})\left(\frac{c}{c_{\alpha,\beta}}\right)^{1-\beta} & c \leqslant c_{\alpha,\beta}; \\ (1-\beta)(1 - c_{\alpha,\beta})^{-\frac{1}{\alpha-1}}(1-c)^{\frac{\alpha}{\alpha-1}} & c \geqslant c_{\alpha,\beta}. \end{cases} \tag{53}$$

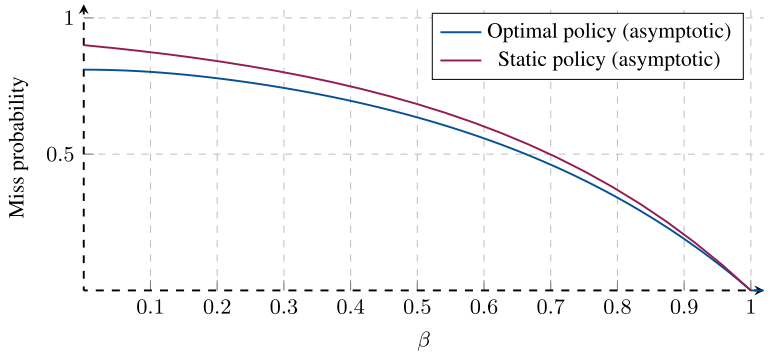
In Figure 6, we plot the asymptotic miss probability of Equation (53) for fixed α and c as a function of the popularity tail parameter β. We also compare it with the asymptotic miss probability of the static policy, which is derived in Ferragut et al. (2016) and is given by $1 - c^{1-\beta}$ for $0 \leqslant \beta \leqslant 1$. The optimal policy has the most advantage when popularities are more homogeneous; in particular, as $\beta \to 0$, we have $M \to (1-c)^{\frac{\alpha}{\alpha-1}} < 1 - c$ for any $\alpha > 1$. The gain is larger as α decreases; the optimal policy capitalizes on the burstiness of the incoming traffic captured by the hazard rates.

A second observation is that convergence of the empirical distribution of the OSIs is very fast. In Figure 7, we plot a simulated sample in the steady state of $\hat{G}_N(x)$ for $N = 100$ items for $\alpha = 2$ and $\beta = 0.5$ as an example. We see the convergence to $G_\infty(x)$ as established in Theorem 5.

Moreover, because convergence to G_∞ is fast, convergence of the random threshold $\hat{\theta}_N := \hat{Q}_N(1 - c) \to \theta^*$ is also fast. We plot panels (a) and (b) of Figure 8 two examples with the same parameters as above for $N = 1{,}000$ and 10,000, respectively. The threshold process $\theta_N(t)$ is approximately constant for large N around the value θ^*.

This last observation is crucial for developing finite N approximations for the performance. Because of the fact that when $\beta \to 1$, the family of distributions $\{L_N\}$ approaches nonuniform integrability, convergence of the miss

Figure 6. Asymptotic Miss Probability for Zipf Popularities with Varying Parameter β



Notes. Pareto interrequest times with $\alpha = 2$ and $c = 0.1$. The static policy was added for comparison.

rate estimate $m_{\mathcal{C}^*}^{(N)}$, the total rate $\lambda^{(N)}$, and the miss probability $M_{\mathcal{C}^*}^{(N)}$ is slow around this value as depicted in the simulations shown in Figure 9.

In order to compute finite N approximations, we use the intuition developed in Equation (35); that is, we compute the asymptotic threshold θ^* by solving $G_\infty(\theta^*) = 1 - c$ and plug in this estimate in place of the random threshold $\hat{\theta}_N$. Then, we estimate $M_{\mathcal{C}^*}^{(N)}$ as

$$M_{\mathcal{C}^*}^{(N)} \approx \frac{\int_0^\infty \lambda G_0(\theta^*/\lambda) L_N(d\lambda)}{\int_0^\infty \lambda L_N(d\lambda)} = \frac{1}{\lambda^{(N)}} \sum_{i=1}^{N} \lambda_i^{(N)} G_0(\theta^*/\lambda_i^{(N)}). \tag{54}$$

This turns out to be equivalent to approximating the optimal performance for that of the fixed threshold policy discussed in Section 6, and it is numerically easy to compute, even for distributions that do not have closed-form expressions as in this case. Therefore, this procedure provides *sharp estimates* of the maximum achievable performance for any policy and finite N. An example of this approximation is depicted in the dashed lines in Figure 9.

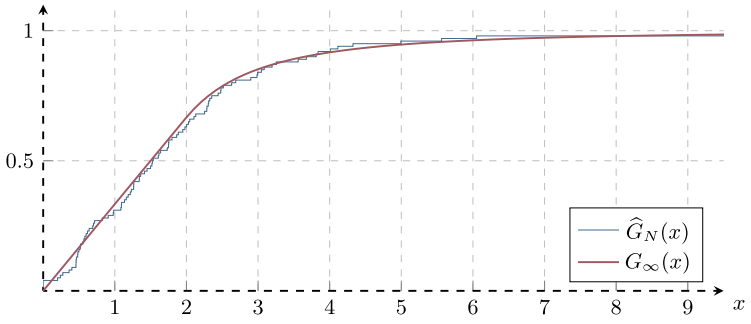
8.2. Erlang Interrequest Times and Zipf Popularities

We now turn our attention to a different example, where the hazard rate of the interrequest distribution is *increasing*. This leads to a totally different behavior; for instance, caching is not a good idea in this setting because upon receiving a request, a subsequent request becomes *less* likely. It is actually preferable to remove the content from memory and prefetch it again closer to request time (Ferragut et al. 2024).

Of course, the optimal memory management policy designed in Section 5 does this automatically by keeping track of the hazard rates, and all of the asymptotic results derived for it are still valid in this case.

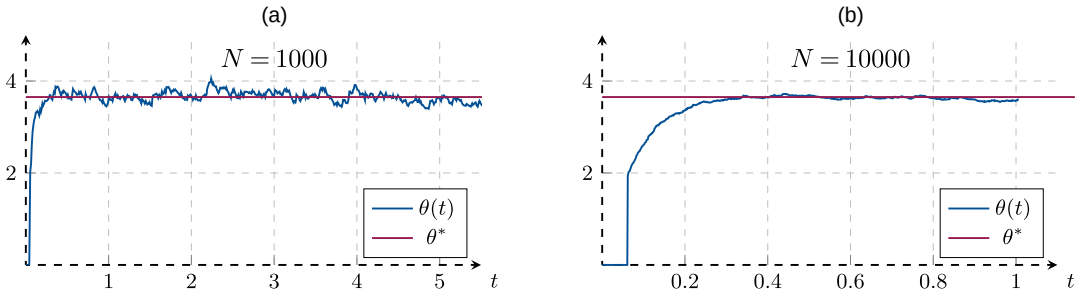
To illustrate this behavior, we choose the interrequest times to be distributed as an Erlang distribution with k stages and appropriate means. For $k = 1$, this corresponds to the Poisson process because interrequest times become exponential. As k grows, the process approaches deterministic interrequest times, and thus, the traffic pattern becomes more regular.

Figure 7. The Empirical Distribution of the Observed Hazard Rates for $N = 100$ and Its Limit $G_\infty(x)$



Note. Pareto requests with $\alpha = 2$ and Zipf popularities with $\beta = 0.5$.

Figure 8. Threshold Evolution and Large-Scale Limit for Pareto Requests $\alpha = 2$, $\beta = 0.5$, and $c = 0.1$



Notes. (a) Catalog size $N = 1{,}000$. (b) Catalog size $N = 10{,}000$.

To be precise, let the base interrequest distribution be Erlang with k stages and mean of one (so $\lambda = k$). This corresponds to the following:

$$f_0(t) = \frac{1}{(k-1)!} k^k t^{k-1} e^{-kt}, \quad F_0(t) = 1 - \sum_{j=0}^{k-1} \frac{1}{j!} (kt)^j e^{-kt} \quad t \geqslant 0.$$

This yields the following nice formula for the hazard rate function:

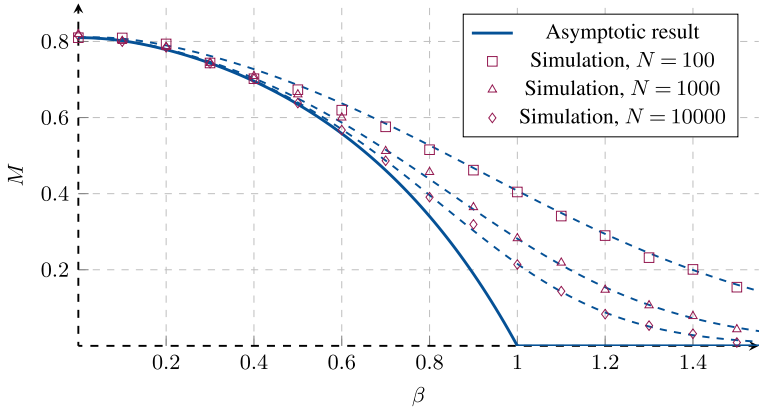
$$\eta(t) = k \frac{\frac{(kt)^{k-1}}{(k-1)!}}{\sum_{j=0}^{k-1} \frac{1}{j!}(kt)^j} = kB(kt, k-1),$$

where $B(A, C)$ is the classical Erlang-B formula for the blocking probability of telephone systems. For $k > 1$, this becomes strictly increasing in t, $\eta(0) = 0$, and $\eta(t) \uparrow k$ when $t \to \infty$.

We do not have analytical expressions for the age distribution or for the observed hazard rates, so G_∞ must be estimated numerically. We do so for the Zipf popularity distribution limit $L(\lambda)$ introduced above. We plot the resulting distribution in Figure 10 together with a simulation of the empirical version $\hat{G}_N(x)$ for $N = 1{,}000$, showing good convergence. Also, in Figure 10, we show the threshold $\hat{\theta}_N$ evolution over time for the optimal policy, showing convergence to the numerically computed limit θ^*.

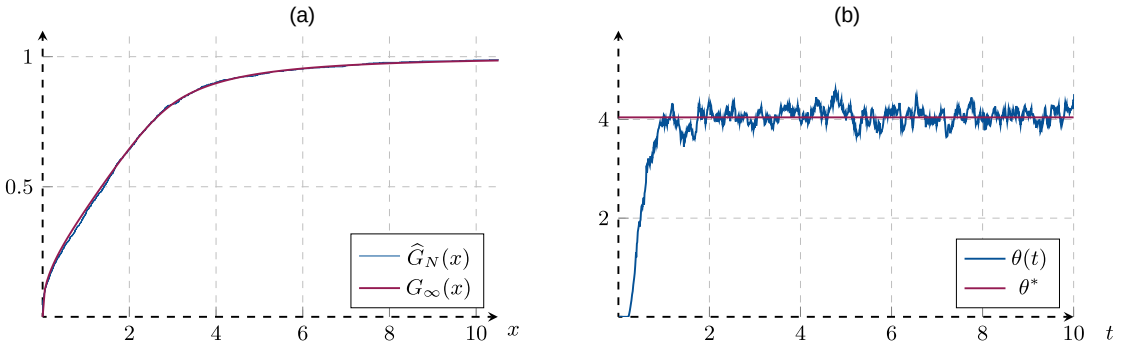
Finally, in Figure 11, we plot the asymptotic miss probability of the optimal policy numerically computed from Equation (38) as well as a simulation of the optimal policy for $N = 1{,}000$ as a function of parameter β. As discussed above, convergence to the performance limit is slow when β approaches one, so we compute the performance estimate from (54), which also corresponds to the optimal prefetching policy from Ferragut et al. (2024), showing good fit. As a comparison, the classical LRU policy is also simulated for the traffic pattern. As we discussed above, classical caching performs badly because of the regularity of the request process.

Figure 9. Miss Probability of the Optimal Policy for Pareto Requests, $\alpha = 2$, $c = 0.1$, and Varying β



Notes. The solid curve represents the asymptotic result (53). Dots represent simulation results for different values of N. The dotted lines represent the finite N approximation (54).

Figure 10. Observed Hazard Rates and Threshold Evolution for an Example with Erlang Requests with $k = 4$ Stages and Zipf Popularities with $\beta = 0.5$ and $c = 0.1$



Notes. (a) Empirical distribution of the observed hazard rates for $N = 1{,}000$ and its limit $G_\infty(x)$. (b) Threshold process evolution and predicted limit θ^* numerically computed.

8.3. Practical Implementation

Although our paper is focused on theoretical results for the optimal management of local memory systems, it is important to discuss possible practical implementations of the optimal policy. From Theorem 3, we know that the key point to implement the optimal policy is correctly estimating the stochastic intensities of the underlying request processes, a well-known hard problem in statistics (Daley and Vere-Jones 2003, chapter 7.5) with few results.

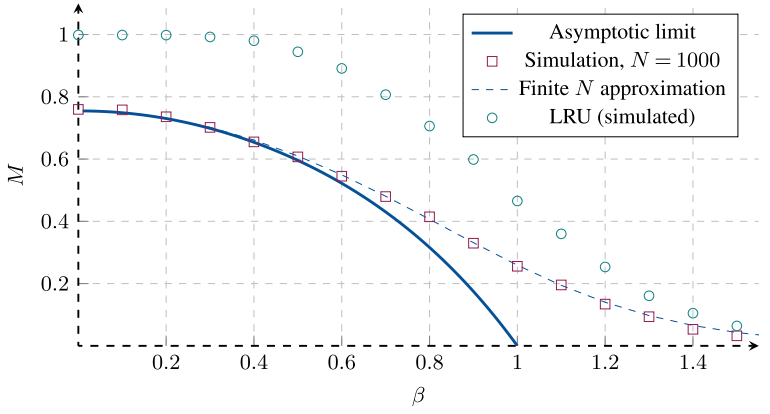
In the case where the incoming processes are independent and renewal processes, the problem boils down to estimating the hazard rate function of each stream based on previously recorded requests. Hazard rate estimation from life tables and for discrete random variables is a well-studied issue, with estimators such as the Kaplan–Meier estimator for the hazard rate and the Nelson–Aalen estimator for the cumulative hazard rate, respectively (Colosimo et al. 2002). In the continuous setting, some sort of *kernel smoothing* must be done (Tanner and Wong 1983) because by definition, the hazard rate depends on the underlying density of the distribution, so the problem is completely related to kernel density estimation.

Nevertheless, in our case, we can propose a practical algorithm along the lines of the Nelson–Aalen estimator. We first discuss this for a given item i, so we drop the dependence on i to ease the notation. Begin by observing that the hazard rate is given by

$$\eta(t) = -\frac{\partial}{\partial t}\log(1 - F_0(t)),$$

where F_0 is the interarrival distribution. In particular, $-\log(1 - F_0(t))$ is the cumulative hazard rate function. The algorithm will proceed as follows.

Figure 11. Miss Probability of the Optimal Policy for Erlang Requests with $k = 4$ Stages, $c = 0.1$, and Varying β



Notes. The solid curve represents the asymptotic result. Squares represent simulation results ($N = 1{,}000$), and the numerical approximation from (54) is shown by the dashed line. For comparison, LRU simulations are also shown.

- Keep track of the observed interrequest times $\{\tilde{\tau}_k : k = 1, \dots, M\}$ up to time t.
- Construct the ordered sample $\{\tilde{\tau}^*_k : k = 1, \dots, M\}$.
- Measure the current elapsed time because the last arrival $t_0 := t - \tau^-(t)$.
- Find k such that

$$\tilde{\tau}^*_{k-1} \leqslant t_0 < \tilde{\tau}^*_k$$

(i.e., the interval in the sample that contains the current elapsed time).

- Estimate the hazard rate at elapsed time t_0 as the slope of the empirical cumulative hazard rate between times $\tilde{\tau}^*_{k-1}$ and $\tilde{\tau}^*_k$, which can be directly computed as

$$\hat{\eta}(t_0) = \frac{\log\left(1 + \frac{1}{N-k}\right)}{\tilde{\tau}^*_k - \tilde{\tau}^*_{k-1}}.$$

Note that the computational complexity of this algorithm is high because we must keep track of a large-enough sample of interarrival times (something that is difficult to achieve for less popular items), and we must do so for each item in the catalog. We must also keep these interarrival times sorted and find the correct interval to evaluate the slope dynamically as time elapses.

This is why more computationally efficient algorithms are needed. In particular, the classical and simple LRU algorithm seems well suited but only in the case of *decreasing* hazard rates or bursty traffic. For more regular traffic, like in the increasing hazard rate case, the problem is still open.

9. Conclusions

This paper analyzes the optimal management of local memory systems with tools of stationary point processes. As a first contribution, we provided a rigorous setting for this problem, and under suitable assumptions, we characterized the optimal policy (in the sense of maximizing hit rates) to store in memory at every moment the items corresponding to the highest stochastic intensities.

From this starting point, for the case of independent request streams, we analyzed the limiting behavior of the optimal policy as the catalog size $N \to \infty$ when a fixed fraction c of items can be stored. Assuming that item request processes come from a common scale family with intensities that have a limit in distribution, we proved that the optimal policy amounts to comparing the stochastic intensity (observed hazard rate) of the process with a fixed *threshold* defined by the $1 - c$ quantile of a certain limiting distribution function. We further characterized the asymptotic performance (miss probability) of this optimal policy.

We also analyzed optimal threshold policies for the finite N case that satisfy the memory constraint in the mean. We showed that they have the same limiting behavior as the optimal policy. Moreover, we found that for renewal traffic and monotonic hazard rates in the interrequest distribution, these are equivalent to *timer-based* policies, where caching or prefetching times are determined by a hazard rate threshold. We also showed that some of the optimality results presented extend quite nicely to the case of heterogeneous object sizes and fractional caching.

Finally, we presented two detailed examples of our results for the standard Zipf model of item popularities and two cases for the interrequest process. For the bursty, decreasing hazard rate case of Pareto interarrival times, we obtain closed-form expressions for the optimal asymptotic threshold and the corresponding performance. We also provide sharp estimates of the optimal performance for finite N validated by detailed stochastic simulations. For the regular, increasing hazard rate case of Erlang interarrival times, closed-form expressions are not available, but we still carry out a numerical validation of the asymptotic threshold and performance. We also use this example to exhibit the significant superiority of the optimal policy in comparison with the popular LRU policy in contrast with the bursty case where LRU has good behavior.

For future research, we may explore the application of our machinery to characterize optimality for other kinds of request traffic: Markov-modulated Poisson processes to account for time variations, more general mixtures of traffic from different sources, and possibly, heterogeneity in item sizes as well as different popularity distributions.

Also, note that as presented, the optimal policy serves as a fundamental limit on the achievable performance but not necessarily a practical algorithm because it requires knowledge of the relevant item intensities and hazard rate distributions. This poses the question of possibly *learning* this information from the actual request data and alternatively, designing an eviction policy that does not make explicit use of these quantities but

approximates optimal performance in practice; to some extent, LRU achieves this in the bursty case, but a counterpart for the increasing hazard rate is currently open.

Appendix A. Proof of Asymptotic Performance

In this section, we will provide a proof for our main result on asymptotic performance of the optimal policy for the uniformly integrable case. Referring back to Section 4, we begin by restating Equation (34):

$$m_{C^*}^{(N)} = \sum_{i=1}^{N} \lambda_i^{(N)} \mathrm{P}_{\Phi_i}^0(\hat{G}_N^{(-i)}(X_i^{(N)}) \leqslant p_N) = \sum_{i=1}^{N} \lambda_i^{(N)} \mathrm{P}_{\Phi_i}^0(X_i^{(N)} \leqslant \hat{Q}_N^{(-i)}(p_N)), \tag{A.1}$$

where $p_N = 1 - \frac{C}{N-1}$, $\hat{G}_N^{(-i)}$ was defined in (33) to be the empirical distribution of the OSIs of items different from i, and $\hat{Q}_N^{(-i)}$ is its corresponding quantile function.

To compute each term on the right, we must invoke the distribution of $X_i^{(N)}$ under the Palm probability; it would be more convenient if the right-hand side of the inequality does not depend on i. For this purpose, we will bound the quantiles obtained from $\hat{G}_N^{(-i)}$ with those of $\hat{G}_N$ defined in (24), which include all items; as usual, $\hat{Q}_N$ is the corresponding quantile function. The key observation is that the contribution of the ith item to the empirical distribution $\hat{G}_N$ of all OSIs is at most $\frac{1}{N}$ and therefore, negligible in the limit as N tends to infinity. More precisely, we have Lemma A.1.

Lemma A.1. $\max_{1 \leqslant i \leqslant N} \|\hat{G}_N - \hat{G}_N^{(-i)}\|_\infty \leqslant \frac{1}{N}$.

Proof. Recalling the definitions (24) and (33), we can write

$$N\hat{G}_N(x) - (N-1)\hat{G}_N^{(-i)}(x) = \sum_{j=1}^{N} \mathbf{1}_{\{X_j^{(N)} \leqslant x\}} - \sum_{j \neq i} \mathbf{1}_{\{X_j^{(N)} \leqslant x\}} = \mathbf{1}_{\{X_i^{(N)} \leqslant x\}};$$
$$\Rightarrow N\left(\hat{G}_N(x) - \hat{G}_N^{(-i)}(x)\right) = \mathbf{1}_{\{X_i^{(N)} \leqslant x\}} - \hat{G}_N^{(-i)}(x).$$

In the preceding identity, the right-hand side is the difference of two numbers, both in [0, 1]; therefore, $|\hat{G}_N(x) - \hat{G}_N^{(-i)}(x)| \leqslant \frac{1}{N}$. This holds for every x and for each i. □

We will now make use of this lemma to bound the distribution of $\hat{G}_N^{(-i)}(X_i^{(N)})$ as required for (A.1) with that of $\hat{G}_N(X_i^{(N)})$. However, this change brings a new difficulty because $\hat{G}_N(\cdot)$ and $X_i^{(N)}$ would not be independent. For this reason, to calculate the distribution, we will use a coupling argument and consider independent versions of the corresponding random variables.

Proposition A.1 (Miss Rate Bounds for Fixed N). *Let $p_N^{\pm} = p_N \pm \frac{1}{N}$, where $p_N = 1 - \frac{C}{N-1}$. Let $\Theta_N^{\pm}$ be the probability distribution function of the random variable $\hat{\theta}_N^{\pm} := \hat{Q}_N(p_N^{\pm})$ under the probability* P. *Then, the miss rate is bounded from above and below by*

$$\int_0^\infty \int_0^\infty \lambda G_0\left(\frac{\theta}{\lambda}\right) \Theta_N^-(d\theta) L_N(d\lambda) \leqslant \frac{m_{C^*}^{(N)}}{N} \leqslant \int_0^\infty \int_0^\infty \lambda G_0\left(\frac{\theta}{\lambda}\right) \Theta_N^+(d\theta) L_N(d\lambda). \tag{A.2}$$

Proof. The proof is based on a coupling argument where we consider independent copies of the OSIs. The distributions of these copies mimic the distributions of the OSIs at an arbitrary time $t = 0$ (that is, under the probability measure P) and upon an arrival at time $t = 0$ (that is, under the Palm distribution P_Φ^0). More precisely, consider two independent, row-independent triangular arrays

$$\{Y_i^N : N \geqslant 1, 1 \leqslant i \leqslant N\} \text{ and } \{Z_i^N : N \geqslant 1, 1 \leqslant i \leqslant N\}$$

defined on a common probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathrm{P}})$ with distributions

$$Y_i^N \sim G(\cdot/\lambda_i^N) \text{ and } Z_i^N \sim G_0(\cdot/\lambda_i^N) \text{ for all } 1 \leqslant i \leqslant N \text{ and } N \geqslant 1.$$

Let

$$\hat{H}_N(x) = \frac{1}{N}\sum_{j=1}^{N} 1_{\{Y_j^N \leqslant x\}} \text{ and } \hat{H}_N^{(-i)}(x) = \frac{1}{N-1}\sum_{j \neq i} 1_{\{Y_j^N \leqslant x\}}.$$

Analogously to Lemma A.1, we have

$$\max_{1 \leqslant i \leqslant N} \left\|\hat{H}_N^{(-i)} - \hat{H}_N\right\|_\infty \leqslant \frac{1}{N}. \tag{A.3}$$

For each $N \geqslant 1$ and $1 \leqslant i \leqslant N$, the random variables

$$Y_1^N, \ldots, Y_{i-1}^N, Z_i^N, Y_{i+1}^N, \ldots, Y_N^N$$

are independent, so if we let $U_i^N = \hat{H}_N^{(-i)}(Z_i^N)$, its distribution under $\tilde{\mathrm{P}}$ will coincide with that of $\hat{G}_N^{(-i)}(X_i^{(N)})$ under the Palm probability $\mathrm{P}_{\Phi_i}^0$: that is,

$$\mathrm{P}_{\Phi_i}^0(\hat{G}_N^{(-i)}(X_i^{(N)}) \leqslant p_N) = \tilde{\mathrm{P}}(U_i^N \leqslant p_N). \tag{A.4}$$

To bound the probability on the right-hand side, observe that from (A.3), we have

$$\{\hat{H}_N(Z_i^N) \leqslant p_N^-\} \subset \{U_i^N \leqslant p_N\} \subset \{\hat{H}_N(Z_i^N) \leqslant p_N^+\}; \tag{A.5}$$

the proof proceeds by computing the probabilities of the extreme sets in (A.5), noting that under the constructed $\tilde{\mathrm{P}}$, Z_i^N is independent of $\hat{H}_N$ (a function of $Y_1^N, \dots Y_N^N$).

Focusing momentarily on the upper bound, we can write

$$\tilde{\mathrm{P}}(\hat{H}_N(Z_i^N) \leqslant p_N^+) = \tilde{\mathrm{P}}(Z_i^N \leqslant \hat{Q}_N^H(p_N^+)), \tag{A.6}$$

where $\hat{Q}_N^H$ denotes the quantile function of $\hat{H}_N$; by construction, the distribution of $\hat{Q}_N^H(p_N^+)$ under $\tilde{\mathrm{P}}$ coincides with that of $\hat{\theta}_N^+ = \hat{Q}_N(p_N^+)$ under P, which by hypothesis, is denoted by $\Theta_N^+(\theta)$. Because Z_i^N is independent and has distribution $G_0(\cdot/\lambda_i^N)$, we can compute the right-hand side of (A.6) to be

$$\int_0^\infty \tilde{\mathrm{P}}(Z_i^N \leqslant \theta)\Theta_N^+(d\theta) = \int_0^\infty G_0\left(\frac{\theta}{\lambda_i^N}\right)\Theta_N^+(d\theta).$$

Multiplying by λ_i^N and averaging over I, we obtain from (A.5)

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^n \lambda_i^N \tilde{P}(U_i^N \leqslant p_N) &\leqslant \frac{1}{N}\sum_{i=1}^n \tilde{\mathrm{P}}\left(\hat{H}_N(Z_i^N) \leqslant p_N^+\right) \\ &= \frac{1}{N}\sum_{i=1}^n \int_0^\infty \lambda_i^N G_0\left(\frac{\theta}{\lambda_i^N}\right)\Theta_N^+(d\theta) = \int_0^\infty \int_0^\infty \lambda G_0\left(\frac{\theta}{\lambda}\right)\Theta_N^+(d\theta)L_N(d\lambda),\end{aligned}$$

where the last step invokes the definition of L_N. An analogous calculation provides the lower bound

$$\int_0^\infty \int_0^\infty \lambda G_0\left(\frac{\theta}{\lambda}\right)\Theta_N^-(d\theta)L_N(d\lambda) \leqslant \frac{1}{N}\sum_{i=1}^n \lambda_i^N \tilde{P}(U_i^N \leqslant p_N).$$

Invoking (A.4) and Expression (A.1) for the miss rate, we conclude the proof. □

A.1. Convergence of Quantiles

Our next step is to compute the limit as $N \to \infty$ under suitable conditions of the lower and upper bounds in (A.2). We will use the convergence $L_N \Rightarrow L$ of Assumption 3, but we also need to address the convergence of $\Theta_N^\pm$ and the distribution of quantiles for the random function $\hat{G}_N$.

Recall that by Theorem 5, we have $\hat{G}_N \Rightarrow G_\infty$ with P probability 1, where G_∞ is defined by (29). A standard result that we have already invoked (e.g., van der Vaart 1998, lemma 21.2) is that convergence in distribution implies the convergence of quantiles at a fixed point p provided that the limit quantile function is continuous at p. We will need a slight generalization of this property for the case where quantiles of the sequence are evaluated at a variable point p_N convergent to p. This is stated and proven next.

Proposition A.2 (Convergence of Quantiles). *Let $\{F_N\}_{N\geqslant 1}$ be a sequence of cumulative distribution functions such that $F_N \Rightarrow F$ for some distribution F. Let Q_N and Q be the quantile functions of F_N and F, respectively. Let $p \in [0,1]$ be a continuity point of Q and $\{p_N\}_{N\geqslant 1}$ be a sequence in $[0,1]$ such that $p_N \to p$. Then, $Q_N(p_N) \to Q(p)$.*

Proof. In the proof, we use the characterization of weak convergence in terms of the Lévy distance

$$d_L(F,G) := \inf\{\epsilon > 0 | F(x-\epsilon) - \epsilon \leqslant G(x) \leqslant F(x+\epsilon) + \epsilon,\ \forall x \in \mathbb{R}\}$$

between cdfs $F, G : \mathbb{R} \to [0,1]$. Convergence in the Lévy distance $d_L(F_N, F) \to 0$ is equivalent to weak convergence $F_N \Rightarrow F$.

Let $\epsilon > 0$. By continuity of Q at p, there exists $\delta > 0$ such that $|q-p| < \delta$ implies $|Q(q) - Q(p)| < \epsilon/2$. Let $\delta_N := d_L(F_N, F)$. Choose N_0 such that $|p_N - p| < \delta/2$ and $\delta_N \leqslant \min\{\epsilon, \delta\}/2$ for all $N \geqslant N_0$.

We claim that

$$Q(p_N - \delta_N) - \delta_N \leqslant Q_N(p_N) \leqslant Q(p_N + \delta_N) + \delta_N. \tag{A.7}$$

Indeed, for the left inequality of (A.7), let $x \in \mathbb{R}$ be such that $F_N(x) \geqslant p_N$. Then,

$$F(x + \delta_N) \geqslant F_N(x) - \delta_N \geqslant p_N - \delta_N,$$

which implies $x + \delta_N \geqslant Q(p_N - \delta_N)$. Taking infimum over such x, we get $Q_N(p_N) \geqslant Q(p_N - \delta_N) - \delta_N$.

For the right inequality of (A.7), let $x \in \mathbb{R}$ be such that $F(x-\delta_N) \geqslant p_N + \delta_N$. Then,

$$F_N(x) \geqslant F(x-\delta_N) - \delta_N \geqslant p_N,$$

which implies $x \geqslant Q_N(p_N)$. Taking infimum over such x, we get $Q(p_N+\delta_N)+\delta_N \geqslant Q_N(p_N)$. This proves the claim.

Let $N \geqslant N_0$. Then,

$$Q(p) - \epsilon \leqslant Q(p) - \frac{\epsilon}{2} - \delta_N < Q(p_N - \delta_N) - \delta_N \leqslant Q_N(p_N),$$

and

$$Q_N(p_N) \leqslant Q(p_N+\delta_N) + \delta_N < Q(p) + \frac{\epsilon}{2} + \delta_N \leqslant Q(p) + \epsilon.$$

Thus, $|Q_N(p_N) - Q(p)| < \epsilon$ for all $N \geqslant N_0$. □

A.2. Proof of Theorem 6

We are now in a position to complete the proof of our main result on performance.

We first establish that $\hat{\theta}_N^{\pm} := \hat{Q}_N(p_N^{\pm}) \to_{N\to\infty} \theta^* := Q_\infty(1-c)$ almost surely in P. This follows by invoking Proposition A.2 at each ω in the set where $\hat{G}_N \Rightarrow G_\infty$, which has unit probability by Theorem 5; note that $p_N^{\pm} \to 1-c$, which by hypothesis, is a point of continuity of Q_∞.

Almost sure convergence implies convergence in probability and in distribution. Therefore, the distributions $\Theta_N^{\pm}$ of the random variables $\hat{\theta}_N^{\pm}$ both satisfy $\Theta_N^{\pm} \Rightarrow \delta_{\theta^*}$, a unit mass at the limit threshold. As a consequence, under Assumption 3, we have convergence of the product measures $\Theta_N^{\pm} \otimes L_N \Rightarrow \delta_{\theta^*} \otimes L$. We will use this fact to show that both bounds in (A.2) converge to the same limit, and it coincides with the right-hand side of (37).

Define for this purpose the function $h : \mathbb{R}_+^2 \to \mathbb{R}_+$ by $h(\lambda,\theta) = \lambda G_0\left(\frac{\theta}{\lambda}\right)$. Our next claim is that

$$(\Theta_N^{\pm} \otimes L_N)h^{-1} \Rightarrow (\delta_{\theta^*} \otimes L)h^{-1}. \tag{A.8}$$

This claim is established in the mapping theorem (Billingsley 1999, theorem 2.7) provided that h is measurable and that its set of discontinuities D_h has zero measure in the limit (i.e., $(\delta_{\theta^*} \otimes L)(D_h) = 0$).

Because the limit measure in θ is a point mass, we have

$$(\delta_{\theta^*} \otimes L)(D_h) = L\left(\left\{\lambda : \frac{\theta^*}{\lambda} \in D_{G_0}\right\}\right). \tag{A.9}$$

Also, note that G_0 being a distribution function, its set of discontinuities is countable. Therefore, the measure on the right of (A.9) can only be positive if there exists $\lambda_0 \in D_L$ (an atom of the measure L) such that $\frac{\theta^*}{\lambda_0} \in D_{G_0}$. This would imply that $\theta^* \in D_L \cdot D_{G_0}$ and is ruled out by hypothesis. Therefore, the claim holds.

To finish the theorem, we must go from the convergence in distribution given in (A.8) to a first-moment condition. Here is where uniform integrability will come into play. It is easiest to express this via independent random variables $\hat{\theta}_N^{\pm}$ and Λ_N with respective distributions $\Theta_N^{\pm}$ and L_N with $\hat{\theta}_N^{\pm} \xrightarrow{d} \theta^*$ and $\Lambda_N \xrightarrow{d} \Lambda$, the latter with distribution L. Condition (A.8) is equivalent to the statement

$$h\left(\Lambda_N, \hat{\theta}_N^{\pm}\right) = \Lambda_N G_0\left(\frac{\hat{\theta}_N^{\pm}}{\Lambda_N}\right) \xrightarrow{d} \Lambda G_0\left(\frac{\theta^*}{\Lambda}\right) = h(\Lambda, \theta^*).$$

Because $G_0 \leqslant 1$, we have $h(\Lambda_N, \hat{\theta}_N^{\pm}) \leqslant \Lambda_N$; now, $\{\Lambda_N\}_{N\geqslant 1}$ is uniformly integrable by hypothesis, so the same happens with $\{h(\Lambda_N, \hat{\theta}_N^{\pm})\}_{N\geqslant 1}$. Now, invoke Billingsley (1999, theorem 3.5) to obtain

$$\int_0^\infty \int_0^\infty \lambda G_0\left(\frac{\theta}{\lambda}\right) L_N(d\lambda)\Theta_N^{\pm}(d\theta) = \mathrm{E}\left[\Lambda_N G_0\left(\frac{\hat{\theta}_N^{\pm}}{\Lambda_N}\right)\right] \underset{N\to\infty}{\to} \mathrm{E}\left[\Lambda G_0\left(\frac{\theta^*}{\Lambda}\right)\right] = \int_0^\infty \lambda G_0\left(\frac{\theta^*}{\lambda}\right) L(d\lambda).$$

Thus, both upper and lower bounds in (A.2) converge to the given limit, which concludes the proof.

Endnotes

[1] Note that τ_0 is a proper random variable because $\{\tau_0 \leqslant t\} = \{\Phi((t,0]) = 0\}$ and thus, is measurable for all $t \leqslant 0$. Similarly, any point τ_k is a random variable.

[2] We specifically avoid using the term *cache* for our system because in caching, normally only the recently requested objects are stored, and our model aims to generalize these policies.

[3] We now make explicit the dependence on the system size N because we will study the limits as $N \to \infty$.

[4] If θ_N is not in the range of valid hazard rates, the equivalence also holds by allowing zero (or infinite) timers (i.e., items that are never (or always) stored).

References

Baccelli F, Brémaud P (2013) *Elements of Queueing Theory* (Springer-Verlag, Berlin).

Bahat O, Makowski AM (2005) Measuring consistency in TTL-based caches. *Performance Evaluation* 62(1):439–455.

Barrera J, Fontbona J (2010) The limiting move-to-front search-cost in law of large numbers asymptotic regimes. *Ann. Appl. Probab.* 20(2):722–752.

Berger DS, Gland P, Singla S, Ciucu F (2014) Exact analysis of TTL cache networks. *Performance Evaluation* 79:2–23.

Berger DS, Henningsen S, Ciucu F, Schmitt JB (2015) Maximizing cache hit ratios by variance reduction. *ACM SIGMETRICS Performance Evaluation Rev.* 43(2):57–59.

Bianchi G, Detti A, Caponi A, Blefari Melazzi N (2013) Check before storing: What is the performance price of content integrity verification in LRU caching? *ACM SIGCOMM Comput. Comm. Rev.* 43(3):59–67.

Bilal M, Kang SG (2017) A cache management scheme for efficient content eviction and replication in cache networks. *IEEE Access* 5:1692–1701.

Billingsley P (1999) *Convergence of Probability Measures*, Wiley Series in Probability and Statistics, 2nd ed. (Wiley, Chichester, UK).

Borst S, Gupta V, Walid A (2010) Distributed caching algorithms for content distribution networks. *2010 Proc. IEEE INFOCOM* (IEEE, Piscataway, NJ), 1–9.

Brémaud P (2020) *Point Process Calculus in Time and Space* (Springer, New York).

Che H, Tung Y, Wang Z (2002) Hierarchical web caching systems: Modeling, design and experimental results. *IEEE J. Selected Areas Comm.* 20(7):1305–1314.

Colosimo E, Ferreira F, Oliveira M, Sousa C (2002) Empirical comparisons between Kaplan–Meier and Nelson–Aalen survival function estimators. *J. Statist. Comput. Simulation* 72(4):299–308.

Daley DJ, Vere-Jones D (2003) *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods* (Springer, New York).

Dan A, Towsley D (1990) An approximate analysis of the LRU and FIFO buffer replacement schemes. *Proc. ACM/SIGMETRICS 1990* (ACM, New York), 143–152.

Dehghan M, Massoulie L, Towsley D, Menasche DS, Tay YC (2019) A utility optimization approach to network cache design. *IEEE/ACM Trans. Networking* 27(3):1013–1027.

Ferragut A, Carrasco M, Paganini F (2024) Caching or pre-fetching? The role of hazard rates. *2024 60th Annual Allerton Conf. Comm. Control Comput.* (IEEE, Piscataway, NJ), 1–8.

Ferragut A, Rodriguez I, Paganini F (2016) Optimizing TTL caches under heavy tailed demands. *Proc. ACM/SIGMETRICS 2016* (ACM, New York), 101–112.

Ferragut A, Rodríguez I, Paganini F (2018) Optimal timer-based caching policies for general arrival processes. *Queueing Systems* 88(3–4):207–241.

Fill JA (1996) Limits and rates of convergence for the distribution of search cost under the move-to-front rule. *Theoret. Comput. Sci.* 164(1):185–206.

Fofack NC, Nain P, Neglia G, Towsley D (2012) Analysis of TTL-based cache networks. *Proc. Internat. Conf. Performance Evaluation Methodologies Tools (VALUETOOLS)* (IEEE, Piscataway, NJ), 1–10.

Fofack NC, Nain P, Neglia G, Towsley D (2014) Performance evaluation of hierarchical TTL-based cache networks. *Comput. Networks* 65:212–231.

Fricker C, Robert P, Roberts J (2012) A versatile and accurate approximation for LRU cache performance. *Proc. 24th Internat. Teletraffic Congress* (IEEE, Piscataway, NJ), 57–64.

Garetto M, Leonardi E, Martina V (2016) A unified approach to the performance analysis of caching systems. *ACM Trans. Model. Performance Evaluation Comput. Systems* 1(3):1–28.

Gast N, Houdt BV (2015) Transient and steady-state regime of a family of list-based cache replacement algorithms. *Proc. ACM/SIGMETRICS 2015* (ACM, New York), 123–136.

Gelenbe E (1973) A unified approach to the evaluation of a class of replacement algorithms. *IEEE Trans. Comput.* C-22(6):611–618.

Ioannidis S, Yeh E (2016) Adaptive caching networks with optimality guarantees. *ACM SIGMETRICS Performance Evaluation Rev.* 44(1):113–124.

Ioannidis S, Massoulie L, Chaintreau A (2010) Distributed caching over heterogeneous mobile networks. *Proc. ACM SIGMETRICS Internat. Conf. Measurement Model. Comput. Systems* (ACM, New York), 311–322.

Jelenković PR (1999) Asymptotic approximation of the move-to-front search cost distribution and least-recently used caching fault probabilities. *Ann. Appl. Probab.* 9(2):430–464.

Jelenković P, Radovanović A (2003) Asymptotic insensitivity of least-recently-used caching to statistical dependency. *Proc. IEEE/INFOCOM 2003* (IEEE, Piscataway, NJ), 438–447.

Jelenković PR, Radovanović A (2004) Least-recently-used caching with dependent requests. *Theoret. Comput. Sci.* 326(1):293–327.

Jelenković PR, Radovanović A (2008) The persistent-access-caching algorithm. *Random Structures Algorithms* 33(2):219–251.

Jelenković PR, Radovanović A, Squillante MS (2006) Critical sizing of LRU caches with dependent requests. *J. Appl. Probab.* 43(4):1013–1027.

Jung J, Berger AW, Balakrishnan H (2003) Modeling TTL-based internet caches. *Proc. IEEE/INFOCOM 2003* (IEEE, Piscataway, NJ), 417–426.

King W (1971) Analysis of paging algorithms. *Proc. IFIP Congress 1971* (North-Holland Publishing Company, Amsterdam), 485–490.

Martina V, Garetto M, Leonardi E (2014) A unified approach to the performance analysis of caching systems. *Proc. IEEE/INFOCOM 2014* (IEEE, Piscataway, NJ), 2040–2048.

Panigrahy NK, Li J, Towsley D (2017) Hit rate vs. hit probability based cache utility maximization. *ACM SIGMETRICS Performance Evaluation Rev.* 45(2):21–23.

Panigrahy NK, Li J, Towsley D, Hollot CV (2020) Network cache design under stationary requests: Exact analysis and Poisson approximation. *Comput. Networks* 180:107379.

Panigrahy NK, Nain P, Neglia G, Towsley D (2022) A new upper bound on cache hit probability for non-anticipative caching policies. *ACM Trans. Model. Performance Evaluation Comput. Systems* 7(2–4):5.

Rosensweig EJ, Kurose J, Towsley D (2010) Approximate models for general cache networks. *Proc. IEEE/INFOCOM 2010* (IEEE, Piscataway, NJ), 1–9.

Shorack GR (1979) The weighted empirical process of row independent random variables with arbitrary distribution functions. *Statistica Neerlandica* 33(4):169–189.

Tanner MA, Wong WH (1983) The estimation of the hazard function from randomly censored data by the kernel method. *Ann. Statist.* 11(3):989–993.

van der Vaart AW (1998) *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, vol. 3 (Cambridge University Press, Cambridge, UK).