

August 28, 2008

RT0813

Computer Science 25 pages

Research Report

A fluid limit for cache algorithms with general request processes

Takayuki Osogami

IBM Tokyo Research Laboratory
1623-14 Shimotsuruma
Yamato-shi, Kanagawa 242-8502, Japan

Limited Distribution Notice

This report has been submitted for publication outside of IBM and will be probably copyrighted if accepted. It has been issued as a Research Report for early dissemination of its contents. In view of the expected transfer of copyright to an outside publisher, its distribution outside of IBM prior to publication should be limited to peer communications and specific requests. After outside publication, requests should be filled only by reprints or copies of the article legally obtained (for example, by payment of royalties).



A fluid limit for cache algorithms with general request processes

Takayuki Osogami
IBM Tokyo Research Laboratory
1623-14 Shimotsuruma
Yamato-shi, Kanagawa 242-8502, Japan
Email: osogami@jp.ibm.com

August 28, 2008

Abstract

We introduce a formal limit, which we refer to as a fluid limit, of scaled stochastic models for a cache managed with the Least-Recently-Used algorithm when requests are issued according to general stochastic point processes, which may be non-stationary. We define our fluid limit as a superposition of dependent replications of the original system with smaller item sizes as the number of replications approaches infinity. We derive the average probability that a requested item is not in a cache (average miss probability) in the fluid limit. The usefulness of the fluid limit is demonstrated in two ways. First, our numerical experiments show that, when items are requested according to inhomogeneous Poisson processes, the average miss probability in the fluid limit closely approximates that in the original system as long as there are sufficient number of items. Second, we show that the asymptotic characteristics of the average miss probability as the cache size approaches infinity are often preserved in the fluid limit. This preservation is attractive since the asymptotic analysis in the fluid limit appears to be simpler than that in the original system. In addition, we show that the average miss probability in the fluid limit is asymptotically insensitive to particular dependencies in the requests when the request rates have a light tail, a property not known for the original system.

1 Introduction

Caching data is a widely used technique for scalability and efficiency in today's communication systems, including the World Wide Web [18], sensor networks [23], and peer-to-peer networks [22]. It is important to optimize the cache algorithms, since the response times perceived by users of these systems can be strongly affected by the cache algorithms. There have been two dominant approaches for analytically evaluating the performance of cache algorithms: stochastic analysis and competitive analysis [1]. When stochastic analysis is applied properly, we can understand the performance more precisely than with competitive analysis and also gain insights into the fundamental characteristics of the cache algorithms. Today, however, stochastic analysis is still limited in its applicability to cache algorithms. Our goal is to make stochastic analysis more applicable to cache algorithms.

1.1 Prior work

Least-Recently-Used (LRU) is a simple and popular cache algorithm and has been studied extensively with stochastic analysis. The stochastic analysis of LRU originates from the stochastic analysis of the Move-To-Front (MTF) list, where a requested item is moved to the head of the list. The miss probability (the

probability that a requested item is not in the cache) for LRU with a cache of size K coincides with the probability that the requested item is not at one of the first K positions of the MTF list. McCabe [16] derives the first two moments of the stationary position of a requested item in an MTF list with an “independent reference model,” which is essentially equivalent to the model where items are requested according to independent Poisson processes. The results of McCabe are extended to the probability distribution by Burville and Kingman [2] and to the generating function by Flajolet et al. [7] and Fill and Holst [6]. Unfortunately, these distribution and generating functions are computationally hard to evaluate numerically and provide little intuition due to the complexity of their expressions.

To gain greater insights from stochastic analysis and to evaluate performance more efficiently, researchers have studied the asymptotic characteristics of the MTF list and LRU. Fill [5] shows that the generating function of the stationary position of a requested item is simplified in the limiting case where the number of items approaches infinity. Jelenković [10] studies the miss probability for LRU in the limiting case where the cache size, K , approaches infinity. In particular, when the request rates, λ_i for $i = 1, 2, \dots$, have a heavy tail (i.e., $\lambda_i \sim c/i^\alpha$ for $i = 1, 2, \dots$ with $c > 0$ and $\alpha > 1$), it is shown that the miss probability for LRU decays with a power law as $K \rightarrow \infty$. Jelenković [10] also studies a fluid limit of the stationary position of a requested item. Roughly speaking, investigating the fluid limit results in breaking up each item into m items of size $1/m$ and formally taking the limit of $m \rightarrow \infty$. In particular, when the request rates have a light tail (i.e., $\lambda_i \sim c \exp(-\xi i^\beta)$ for $i = 1, 2, \dots$ with $c, \xi, \beta > 0$), it is shown that the miss probability for LRU decays exponentially in the fluid limit. Hirade and Osogami [9] show that the miss probabilities for LRU and the 2Q cache algorithm [14], respectively, can be closely approximated with those analyzed in a fluid limit.

An asymptotic analysis is also found to be useful in comparing the performance of cache algorithms. For example, Jelenković and Radovanović [12] discuss the asymptotic optimality of the Persistent-Access-Caching algorithm as $K \rightarrow \infty$ when the request rates have a heavy tail.

The prior work mentioned above assumes the independent reference model, but stochastic analysis has also been applied for various dependent request processes. When the request process forms a Markov chain, Lam et al. [15] and Rodrigues [19], respectively, derive the mean and the variance of the stationary position of a requested item in an MTF list, and Chu and Knott [3] derive an expression for the stationary miss probability for LRU. Coffman and Jelenković [4] derives the first two moments of the stationary position of a requested item in an MTF list when the probability of requesting each item depends on the state of a modulating process.

Similar to the case with the independent reference model, the analysis of the asymptotic characteristics is found to provide insight into the fundamental nature of LRU. Jelenković and Radovanović [11] and Sugimoto and Miyoshi [21] show that, when the request rates have a heavy tail, the miss probability for LRU is asymptotically insensitive to the types of dependencies in the request process studied in Coffman and Jelenković [4] as $K \rightarrow \infty$. Jelenković et al. [13] characterize the critical cache sizes where the miss probability for LRU becomes insensitive to the dependencies.

1.2 Our contributions

In this paper, we define a fluid limit of a stochastic model for a cache managed with LRU when the requests follow general stochastic point processes. Our fluid limit is a non-trivial extension of the fluid limits for the independent reference model in [10, 9]. We will explain how the dependencies in the request process would disappear with a *trivial* extension of their fluid limits. Then we formally derive an analytical expression, $\bar{p}^{(\infty)}$, for the average miss probability for LRU in our fluid limit. The definition of the fluid limit and the analysis of $\bar{p}^{(\infty)}$ constitute the primary contributions of this paper. The analysis in a fluid limit is

useful in two ways, and our secondary contributions are to demonstrate the usefulness with simulation and asymptotic analysis.

First, $\bar{p}^{(\infty)}$ can be used as an approximation of the average miss probability for LRU in the original system, \bar{p} , whose numerical analysis is intractable. We will study $\bar{p}^{(\infty)}$ when the requests follow inhomogeneous Poisson processes, which are non-stationary. Note that all of the prior work on stochastic analysis of cache algorithms assumes stationary request processes for tractability. Our numerical experiments will show that the error in approximating \bar{p} with $\bar{p}^{(\infty)}$ is typically within 1% for $N > 128$ and smaller for a larger N .

Second, $\bar{p}^{(\infty)}$ can provide insights into the fundamental nature of cache algorithms. We find that asymptotic characteristics of LRU are often preserved in our fluid limit. Specifically, we will see that, as $K \rightarrow \infty$, $\bar{p}^{(\infty)}$ is asymptotically insensitive to particular dependencies in the request processes when the request rates have a heavy tail, which agrees with the findings for \bar{p} in [4, 21]. We also find that the asymptotic analysis of $\bar{p}^{(\infty)}$ appears to be simpler than a corresponding analysis of \bar{p} . This simplicity allows us to find that the asymptotic insensitivity of $\bar{p}^{(\infty)}$ to the particular dependencies also holds for the case of a light tail. Note that asymptotic characteristics of \bar{p} as $K \rightarrow \infty$ is not known even for the independent reference model. Recall that Jelenković [10] studies the asymptotic characteristics for the case of a light tail in his fluid limit.

The rest of the paper is organized follows. In Section 2, we derive an expression for \bar{p} . In Section 3, we define the fluid limit and formally derive a general expression for $\bar{p}^{(\infty)}$. In Section 4, we study $\bar{p}^{(\infty)}$ when requests follow inhomogeneous Poisson processes. In particular, we evaluate the accuracy of approximating \bar{p} with $\bar{p}^{(\infty)}$. In Section 5, we show that $\bar{p}^{(\infty)}$ is asymptotically insensitive to particular dependencies in the request process.

2 LRU with general stochastic point processes

In this section, we derive an expression for the average miss probability for LRU when items are requested according to general stochastic point processes, Ψ . In Section 2.1, we formally define the model of caching with LRU and state assumptions on Ψ . In Section 2.2, we analyze the average miss probability for LRU, which will be used in Section 3 to study the fluid limit.

2.1 Model and assumptions

We consider a system with N items of size 1 and a cache of size K , where $0 < K < N \leq \infty$. The items are requested according to stochastic point processes, $\Psi = (\Psi_1, \dots, \Psi_N)$, where $\Psi_i = \{t_\ell^{(i)}, \ell \in \mathbf{Z}\}$ denotes the request process for the i -th item, e_i . For each e_i , we let $t_0^{(i)} \leq 0 < t_1^{(i)}$ and $t_\ell^{(i)} < t_{\ell+1}^{(i)}$ for $\ell \in \mathbf{Z}$, so that $t_\ell^{(i)}$ denotes the epoch of the ℓ -th request for e_i after time 0 for $\ell > 0$, although $t_\ell^{(i)}$ is also defined for $\ell \leq 0$.

When a requested item is not in the cache, LRU removes the item that was requested least recently from the cache, and the requested item is placed in the cache. When a requested item is in the cache, the cache remains unchanged. We assume that exactly K items are always stored in the cache. Also, we assume that items are requested one at a time, since simultaneous requests would require a tie breaking rule. Formally, we assume that $t_\ell^{(i)} \neq t_{\ell'}^{(j)}$ for any ℓ, ℓ', i, j . In addition, we assume that $t_\ell^{(i)} \rightarrow \infty$ as $\ell \rightarrow \infty$ and $t_\ell^{(i)} \rightarrow -\infty$ as $\ell \rightarrow -\infty$, so that a finite number of requests are issued in a bounded interval. When these assumptions hold, we say that Ψ is simple.

The metric of interest is the miss probability, the probability that a requested item is not in the cache. In contrast to the prior work, Ψ may be non-stationary in this paper. Thus, instead of the stationary

miss probability, which may not exist, we will study the average miss probability. Specifically, let $p_{i,\ell}$ be the probability that the ℓ -th request for e_i is a miss (i.e., the e_i is not in the cache). The average miss probability for e_i is defined as $\bar{p}_i \equiv \lim_{L \rightarrow \infty} \frac{1}{L} \sum_{\ell=1}^L p_{i,\ell}$.

To formally study \bar{p}_i , we use notations from [20] and make additional assumptions about Ψ . Let θ_t be the shift operator that shifts time by t and relabels the indices so that the index of the first request epoch after time 0 is 1. Formally, $\theta_t \Psi_i = \left\{ (t_{M^{(i)}(t)+\ell}^{(i)} - t), \ell \in \mathbf{Z} \right\}$, where $M^{(i)}(t)$ is the maximum ℓ such that $t_\ell^{(i)} \leq t$. Let $\theta_t \Psi = (\theta_t \Psi_1, \dots, \theta_t \Psi_N)$. We assume that Ψ is time-asymptotically stationary, so that there exists a distribution defined by

$$\mathbf{P}^*(\Psi \in \mathcal{E}) \equiv \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t \mathbf{P}(\theta_u \Psi \in \mathcal{E}) du.$$

For simplicity, we assume that Ψ is ergodic with respect to \mathbf{P}^* .

Finally, we assume that the average request rate, λ_i , of e_i satisfies $0 < \lambda_i < \infty$ for $i = 1, \dots, N$. In addition, when $N = \infty$, we assume that $\sum_{i=1}^N \lambda_i < \infty$. Formally, λ_i is defined as follows:

$$\lambda_i = \mathbf{E}^*[M^{(i)}(1)], \quad (1)$$

where $M^{(i)}(1)$ denotes the number of requests for e_i in $(0, 1]$, and \mathbf{E}^* denotes the expectation with respect to \mathbf{P}^* .

2.2 Average miss probability

Under the above assumptions, Ψ_i is event-asymptotically stationary, so that the distribution defined by

$$\mathbf{P}^{0,i}(\Psi_i \in \mathcal{E}) \equiv \lim_{L \rightarrow \infty} \frac{1}{L} \sum_{\ell=1}^L \mathbf{P}(\theta_{t_\ell^{(i)}} \Psi_i \in \mathcal{E})$$

exists for $1 \leq i \leq N$ (see Theorem 2.9 from [20]). Then \bar{p}_i can be expressed conveniently using $\mathbf{P}^{0,i}$:

Lemma 1 *When Ψ is simple, time-asymptotically stationary, ergodic, and $\sum_{i=1}^N \lambda_i < \infty$, the average miss probability of e_i for LRU is*

$$\bar{p}_i = \mathbf{P}^{0,i} \left(\sum_{j \neq i} I\{t_1^{(j)} < t_1^{(i)}\} \geq K \right)$$

for $1 \leq i \leq N$, where I is the indicator random variable.

Before formally proving Lemma 1, we provide an intuitive explanation. We may see $\mathbf{P}^{0,i}(\mathcal{E})$ as the probability of an event, \mathcal{E} , when we “randomly observe way out at” [20] the epoch of a request for e_i , letting the time of the observation be zero. The next request for e_i after the observation is at time $t_1^{(i)}$ and is a miss iff at least K distinct items have been requested in the interval $(0, t_1^{(i)})$. Since items are requested one at a time, e_j is requested in the interval $(0, t_1^{(i)})$ iff $t_1^{(j)} < t_1^{(i)}$ for any $e_j \neq e_i$. Hence, the request for e_i at time $t_1^{(i)}$ is a miss iff $\sum_{j \neq i} I\{t_1^{(j)} < t_1^{(i)}\} \geq K$.

Proof of Lemma 1: The ℓ -th request for e_i is a miss iff at least K distinct items are requested in the interval, $(t_{\ell-1}^{(i)}, t_\ell^{(i)})$. Hence, we have

$$p_{i,\ell} = \mathbf{P} \left(\sum_{j \neq i} I\{\exists \kappa \text{ s.t. } t_{\ell-1}^{(i)} < t_\kappa^{(j)} < t_\ell^{(i)}\} \geq K \right).$$

Since there exists κ such that $t_{\ell-1}^{(i)} < t_\kappa^{(j)} < t_\ell^{(i)}$ iff the first request for e_j after time $t_{\ell-1}^{(i)}$ is before $t_\ell^{(i)}$, we obtain

$$p_{i,\ell} = \mathbf{P} \left(\theta_{t_{\ell-1}^{(i)}}^{(i)} \Psi \in \mathcal{E}_i \right),$$

where

$$\mathcal{E}_i = \left\{ \Psi \mid \sum_{j \neq i} I\{t_1^{(j)} < t_1^{(i)}\} \geq K \right\}.$$

Since $0 \leq p_{i,1} \leq 1$, we can calculate \bar{p}_i as the average miss probability from the second request for e_i :

$$\begin{aligned} \bar{p}_i &= \lim_{L \rightarrow \infty} \frac{1}{L-1} \sum_{\ell=2}^L \mathbf{P} \left(\theta_{t_{\ell-1}^{(i)}}^{(i)} \Psi \in \mathcal{E}_i \right) \\ &= \mathbf{P}^{0,i}(\mathcal{E}_i), \end{aligned}$$

which completes the proof of the lemma. \blacksquare

3 Fluid limit

In this section, we introduce a fluid limit of the stochastic model for caching with LRU and formally derive the average miss probability for LRU in the fluid limit.

3.1 Scaled systems and fluid limit

We consider a sequence of scaled systems, where the m -th scaled system has mN items, $e_{i,k}$ for $1 \leq k \leq m$ and $1 \leq i \leq N$, of size $1/m$. The first scaled system corresponds to the original system, and we call the scaled system with $m \rightarrow \infty$ the fluid limit of the original system. For $1 \leq k \leq m$, let $E_k = (e_{1,k}, \dots, e_{N,k})$ and let $\Phi_k = (\Phi_{1,k}, \dots, \Phi_{N,k})$ be the request processes for E_k . Let $t_\ell^{(i,k)}$ be the epoch of the ℓ -th request for $e_{i,k}$ after time 0, so that $\Phi_{i,k} = \{t_\ell^{(i,k)}, \ell \in \mathbf{Z}\}$.

Such scaled systems are also considered in [9, 10]. For example, the m -th scaled system, $\mathcal{S}^{(m)}$, of [9] can be seen as a superposition of independent replications of the original system. Specifically, in $\mathcal{S}^{(m)}$, Φ_k for $1 \leq k \leq m$ are independent and stochastically identical to Ψ . Unfortunately, the dependencies in Ψ would disappear in $\mathcal{S}^{(\infty)}$ in the sense that $\mathcal{S}^{(\infty)}$ with general Ψ is identical to that when Ψ is a vector of independent Poisson processes. We formally prove the above observation in Section A.

We will define our scaled system as a superposition of *dependent* replications of the original system. Also, in contrast to [9, 10], we will define a sequence of scaled systems for each e_i , so that the scaled systems

for different items have different dependencies in Φ_k . Let $\mathcal{T}_i^{(m)}$ be the m -th scaled system for e_i . For each e_i , we will study the miss probability for the e_i in $\mathcal{T}_i^{(\infty)}$. In $\mathcal{T}_i^{(m)}$, we assume that Φ_k is stochastically identical to Ψ (i.e., for $1 \leq k \leq m$, it holds that $\mathbf{P}(\Phi_k \in \mathcal{E}) = \mathbf{P}(\Psi \in \mathcal{E})$ for any measurable set, \mathcal{E}). However, we assume that Φ_k for $1 \leq k \leq m$ depend on each other. Specifically, in $\mathcal{T}_i^{(m)}$, we assume that $\Phi_{i,k}$ for $1 \leq k \leq m$ have the same sample path (i.e., $t_\ell^{(i,k)} = t_\ell^{(i,k')}$ for any $\ell \in \mathbf{Z}$ and $1 \leq k, k' \leq m$) and that Φ_k for $1 \leq k \leq m$ are conditionally independent given $\Phi_{i,1}$. Formally, for any measurable sets, \mathcal{E}_k for $1 \leq k \leq m$, it holds that

$$\mathbf{P}(\Psi_k \in \mathcal{E}_k, \forall k \in \{1, \dots, m\} \mid \Psi_{i,1}) = \prod_{k=1}^m \mathbf{P}(\Psi_k \in \mathcal{E}_k \mid \Psi_{i,1}). \quad (2)$$

To clarify the assumptions on Φ , consider a way to simulate $\Phi^{(m)} \equiv (\Phi_1, \dots, \Phi_m)$ in $\mathcal{T}_i^{(m)}$ for a bounded interval, $(0, T]$. We first simulate Ψ_i in the original system. This gives us a sequence of epochs, $\Psi_i(\omega) = \{t_1^{(i)}(\omega), \dots, t_{L_i(\omega)}^{(i)}(\omega)\}$. Then, for $1 \leq k \leq m$, we let $\Phi_{i,k}(\omega) = \Psi_i(\omega)$ (i.e., $t_\ell^{(i,k)}(\omega) = t_\ell^{(i)}(\omega)$ for $1 \leq \ell \leq L_i(\omega)$) be the simulated epochs of the requests for $e_{i,k}$ in $\mathcal{T}_i^{(m)}$. Next we simulate Ψ_j for all $j \neq i$ in the original system in such a way that $\Psi_i(\omega)$ and Ψ_j for $j \neq i$ have the desired dependency. This gives us a set of sequences of epochs, $\Psi_j(\omega_1) = \{t_1^{(j)}(\omega_1), \dots, t_{L_j(\omega_1)}^{(j)}(\omega_1)\}$ for $j \neq i$. Then, for $j \neq i$, we let $\Phi^{(j,1)}(\omega_1) = \Psi_j(\omega_1)$ be the simulated epochs of the requests for $e_{j,1}$ in $\mathcal{T}_i^{(m)}$. We repeat simulating Ψ_j for $j \neq i$ in the same way but independently of the previous repetitions. For $1 \leq k \leq m$, the results of the k -th repetition can be used to construct the simulated epochs, $\Phi_{j,k}(\omega_k)$ for $j \neq i$, of the requests for $e_{j,k}$ in $\mathcal{T}_i^{(m)}$.

To avoid introducing a tie-breaking rule, we assume that $\Phi^{(m)}$ is simple in the sense that, in $\mathcal{T}_i^{(m)}$, the items except $e_{i,k}$ for $1 \leq k \leq m$ are requested one at a time almost surely. This means, in the original system, that there is no mass probability: $\mathbf{P}(t_\ell^{(i)} = t) = 0$ for any ℓ, t , and e_i .

3.2 Miss probability in the fluid limit

We say that a request for e_i is a miss in $\mathcal{T}_i^{(m)}$ iff more than half of $e_{i,k}$ for $1 \leq k \leq m$ is not in the cache upon the request. Let $p_{i,\ell}^{(m)}$ be the probability that the ℓ -th request for e_i is a miss in $\mathcal{T}_i^{(m)}$. Let $C_{i,\ell}^{(m)}$ be the total size of distinct items that are requested after $t_{\ell-1}^{(i,1)}$ and before $t_\ell^{(i,1)}$. Then $p_{i,\ell}^{(m)} = \mathbf{P}\left(C_{i,\ell}^{(m)} > K - 1/2\right)$.

Although $e_{i,k}$ for $1 \leq k \leq m$ are requested at the same epochs in $\mathcal{T}_i^{(m)}$, $p_{i,\ell}^{(m)}$ does not depend on a particular tie-breaking rule. Hence, we do not specify the tie-breaking rule.

We first study $C_{i,\ell}^{(\infty)}$, which will be used to derive $p_{i,\ell}^{(\infty)}$. Notice that we can understand $C_{i,\ell}^{(m)}$ in the context of an MTF list, where the mN items in $\mathcal{T}_i^{(m)}$ are in the decreasing order of the time of the last request. Immediately before $t_\ell^{(i,k)}$, one of $e_{i,k}$ for $1 \leq k \leq m$ is at $C_{i,\ell}^{(m)}$ in the MTF list.

Lemma 2 *Let $I_\ell(j)$ be the indicator random variable such that $I_\ell(j) = 1$ iff e_j is requested in the interval $(t_{\ell-1}^{(i)}, t_\ell^{(i)})$ in the original system for $1 \leq j \leq N$. Note that $I_\ell(i) = 0$. Then $C_{i,\ell}^{(m)} \Rightarrow \sum_{j=1}^N \mathbf{E}[I_\ell(j) \mid \Psi_i]$ as $m \rightarrow \infty$, where \Rightarrow denotes convergence in distribution.*

Proof: We prove the convergence in distribution by showing the convergence of the Laplace transform, $\varphi_{i,\ell}^{(m)}(s) \equiv \mathbf{E}\left[\exp(-s C_{i,\ell}^{(m)})\right]$ for $0 \leq s < \infty$, of $C_{i,\ell}^{(m)}$. Let $I_\ell(j, k)$ be the indicator random variable such

that $I_\ell(j, k) = 1$ iff $e_{j,k}$ is requested after $t_{\ell-1}^{(i,1)}$ and before $t_\ell^{(i,1)}$. By definition, $I_\ell(i, k) = 0$ for $1 \leq k \leq m$. Then

$$\begin{aligned}\varphi_{i,\ell}^{(m)}(s) &= \mathbf{E} \left[\exp \left(-\frac{s}{m} \sum_{j=1}^N \sum_{k=1}^m I_\ell(j, k) \right) \right] \\ &= \mathbf{E} \left[\prod_{k=1}^m \exp \left(-\frac{s}{m} \sum_{j=1}^N I_\ell(j, k) \right) \right].\end{aligned}$$

The conditional independence assumed in Equation (2) implies

$$\varphi_{i,\ell}^{(m)}(s) = \mathbf{E} \left[\prod_{k=1}^m \mathbf{E} \left[\exp \left(-\frac{s}{m} \sum_{j=1}^N I_\ell(j, k) \right) \mid \Phi_{i,1} \right] \right].$$

Also, since Φ_k for $1 \leq k \leq m$ are stochastically identical, we obtain

$$\varphi_{i,\ell}^{(m)}(s) = \mathbf{E} \left[\mathbf{E} \left[\exp \left(-\frac{s}{m} \sum_{j=1}^N I_\ell(j, 1) \right) \mid \Phi_{i,1} \right]^m \right].$$

Let Q_r be the conditional probability that r distinct items in E_1 are requested in the interval $(t_{\ell-1}^{(i,1)}, t_\ell^{(i,1)})$ given $\Phi_{i,1}$. Formally, let $Q_r = \mathbf{P} \left(\sum_{j=1}^N I_\ell(j, 1) = r \mid \Phi_{i,1} \right)$ for $0 \leq r \leq N-1$. Then

$$\varphi_{i,\ell}^{(m)}(s) = \mathbf{E} \left[\left(\sum_{r=0}^{N-1} Q_r \exp(-s r/m) \right)^m \right].$$

Since $\sum_{r=0}^{N-1} Q_r = 1$, the dominated convergence theorem can be used to show that

$$\lim_{m \rightarrow \infty} \varphi_{i,\ell}^{(m)}(s) = \mathbf{E} \left[\lim_{m \rightarrow \infty} \left(\sum_{r=0}^{N-1} Q_r \exp(-s r/m) \right)^m \right].$$

By Lemma 9 in Appendix C, we obtain

$$\lim_{m \rightarrow \infty} \varphi_{i,\ell}^{(m)}(s) = \mathbf{E} \left[\exp \left(-s \sum_{r=0}^{N-1} r Q_r \right) \right].$$

Since $\sum_{r=0}^{N-1} r Q_r = \mathbf{E} \left[\sum_{j=1}^N I_\ell(j, 1) \mid \Phi_{i,1} \right]$, we obtain

$$\lim_{m \rightarrow \infty} \varphi_{i,\ell}^{(m)}(s) = \mathbf{E} \left[\exp \left(-s \mathbf{E} \left[\sum_{j=1}^N I_\ell(j, 1) \mid \Phi_{i,1} \right] \right) \right].$$

Therefore, the continuity theorem (e.g., see p. 262 from [8]) implies that $C_{i,\ell}^{(m)} \Rightarrow \mathbf{E} \left[\sum_{j=1}^N I_\ell(j, 1) \mid \Phi_{i,1} \right]$ as $m \rightarrow \infty$. Since the pair $(I_\ell(j, 1), \Phi_{i,1})$ and the pair $(I_\ell(j), \Psi_i)$ are stochastically identical, this completes

the proof of the lemma via the linearity of expectation. ■

Next, we study the average miss probability of e_i ,

$$\bar{p}_i^{(m)} = \lim_{L \rightarrow \infty} \frac{1}{L} \sum_{\ell=1}^L p_{i,\ell}^{(m)}, \quad (3)$$

as $m \rightarrow \infty$. The next theorem characterizes $\bar{p}_i^{(\infty)}$, which should be compared against Lemma 1, which characterizes $\bar{p}_i = \bar{p}_i^{(1)}$. In particular, a random variable, $I\{t_1^{(j)} < t_1^{(i)}\}$, in \bar{p}_i is replaced with a conditional expectation, $\mathbf{E}\left[I\{t_1^{(j)} < t_1^{(i)}\} \mid \Psi_i\right]$, in $\bar{p}_i^{(\infty)}$. This suggests that some randomness disappears in $\mathcal{T}_i^{(\infty)}$. Roughly speaking, in $\mathcal{T}_i(\infty)$, whether or not a request for e_i is a miss is determined only by Ψ_i and by the expected impact that Ψ_i has on Ψ_j for $j \neq i$ via the dependencies between Ψ_i and Ψ_j for $j \neq i$.

Theorem 1 *Under the conditions of Lemma 1, we have*

$$\lim_{m \rightarrow \infty} \bar{p}_i^{(m)} = \mathbf{P}^{0,i} \left(\sum_{j=1}^N \mathbf{E}\left[I\{t_1^{(j)} < t_1^{(i)}\} \mid \Psi_i\right] > K - \frac{1}{2} \right),$$

where $\bar{p}_i^{(m)}$ is defined with Equation (3) for $\mathcal{T}_i^{(m)}$.

Proof: Recall that $p_{i,\ell}^{(m)} = \mathbf{P}\left(C_{i,\ell}^{(m)} > K - 1/2\right)$. Lemma 2 suggests that

$$p_{i,\ell}^{(m)} \rightarrow \mathbf{P} \left(\sum_{j=1}^N \mathbf{E}\left[I_\ell(j) \mid \Psi_i\right] > K - \frac{1}{2} \right) \quad (4)$$

as $m \rightarrow \infty$. By the definition of $I_\ell(j)$ given in Lemma 2, the last expression is equivalent to

$$p_{i,\ell}^{(m)} \rightarrow \mathbf{P} \left(\theta_{\ell-1}^{(i)} \Psi \in \mathcal{D}_i \right) \quad (5)$$

where

$$\mathcal{D}_i = \left\{ \Psi \mid \sum_{j=1}^N \mathbf{E}\left[I\{t_1^{(j)} < t_1^{(i)}\} \mid \Psi_i\right] > K - \frac{1}{2} \right\}.$$

Since $0 \leq p_{i,1}^{(m)} \leq 1$, we can calculate $\bar{p}_i^{(m)}$ from the second request for e_i , so that

$$\bar{p}_i^{(m)} = \lim_{L \rightarrow \infty} \frac{1}{L-1} \sum_{\ell=2}^L p_{i,\ell}^{(m)}. \quad (6)$$

Since Ψ is time-asymptotically stationary, $\Phi^{(m)}$ is time-asymptotically stationary for any m . Hence, $\bar{p}_i^{(m)}$ exists for any m . Thus, we can exchange the limits to obtain

$$\lim_{m \rightarrow \infty} \bar{p}_i^{(m)} = \lim_{L \rightarrow \infty} \frac{1}{L-1} \sum_{\ell=2}^L \lim_{m \rightarrow \infty} p_{i,\ell}^{(m)}, \quad (7)$$

which together with Relation (5) proves the theorem. ■

4 Inhomogeneous Poisson requests

In this section, we study the $\bar{p}_i^{(\infty)}$ derived in Section 3 in more detail for the particular case when the requests are issued according to inhomogeneous Poisson processes. In Section 4.1, we derive an explicit expression for $\bar{p}_i^{(\infty)}$ in this particular case. Our derivation uses $H = \lambda G$, an extension of Little's law, to convert the event-average expression in Theorem 1 to a time-average expression. In Section 4.2, we numerically evaluate the expression derived in Section 4.1 and study the accuracy of approximating \bar{p}_i with $\bar{p}_i^{(\infty)}$.

4.1 Miss probability in the fluid limit

Here, we study $\bar{p}_i^{(\infty)}$ when Ψ is a vector of independent inhomogeneous Poisson processes. Let $\lambda_i(t)$ be the request rate for e_i at time t . Then the average request rate defined by Equation (1) is $\lambda_i = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \lambda_i(t) dt$.

Theorem 2 *In addition to the conditions of Lemma 1, if Ψ_i is an inhomogeneous Poisson process with rate $\lambda_i(t)$ at time t for $1 \leq i \leq N$, then the average miss probability of e_i in $\mathcal{T}_i^{(\infty)}$ is*

$$\bar{p}_i^{(\infty)} = \frac{1}{\lambda_i} \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \exp(-\Lambda_i(t, \tau_i(t))) \lambda_i(t) dt,$$

where, for $1 \leq i \leq N$, we define

$$\Lambda_i(t, u) \equiv \int_t^u \lambda_i(v) dv \quad (8)$$

and $\tau_i(t)$ is the maximum u such that

$$\sum_{j \neq i} (1 - \exp(-\Lambda_j(t, u))) \leq K - \frac{1}{2}. \quad (9)$$

Proof: We first consider $p_{i,\ell}^{(\infty)}$. By Relation (4) and the independence of Ψ_i and Ψ_j for $i \neq j$, we obtain

$$p_{i,\ell}^{(m)} \rightarrow \mathbf{P} \left(\sum_{j=1}^N \mathbf{E} \left[I_\ell(j) \mid (t_{\ell-1}^{(i)}, t_\ell^{(i)}) \right] > K - \frac{1}{2} \right)$$

as $m \rightarrow \infty$ for $\ell > 1$. Since $\mathbf{E} \left[I_\ell(j) \mid (t_{\ell-1}^{(i)}, t_\ell^{(i)}) \right]$ is the conditional probability that e_j is requested in the interval, $(t_{\ell-1}^{(i)}, t_\ell^{(i)})$, we obtain, by a property of the inhomogeneous Poisson process (e.g., see p. 246 of [17]), that

$$\mathbf{E} \left[I_\ell(j) \mid (t_{\ell-1}^{(i)}, t_\ell^{(i)}) \right] = 1 - \exp \left(-\Lambda_j(t_{\ell-1}^{(i)}, t_\ell^{(i)}) \right),$$

where $\Lambda_j(u, t)$ is defined with Equation (8).

Let $\lambda_{i,\ell-1}(u)$ be the probability density function for the epoch of the $(\ell-1)$ -th request of e_i , and let $U_{i,\ell}(t_{\ell-1}^{(i)})$ be the epoch of the ℓ -th request for e_i given $t_{\ell-1}^{(i)}$. By the Markovian property, given $t_{\ell-1}^{(i)}$, $U_{i,\ell}(t)$

is conditionally independent of ℓ , so that we will write $U_i(t) \equiv U_{i,\ell}(t)$, which can be understood as the epoch of the first request for e_i after time t . By conditioning on $t_{\ell-1}^{(i)}$, we obtain

$$p_{i,\ell}^{(\infty)} = \int_0^\infty P_i(t) \lambda_{i,\ell-1}(t) dt, \quad (10)$$

where we define

$$P_i(t) \equiv \mathbf{P} \left(\sum_{j \neq i} (1 - \exp(-\Lambda_j(t, U_i(t)))) > K - \frac{1}{2} \right).$$

Since $1 - \exp(-\Lambda_j(t, u))$ is non-decreasing with u for any t , we can express $P_i(t)$, using the definitions of $\tau_i(t)$, as

$$\begin{aligned} P_i(t) &= \mathbf{P}(U_i(t) > \tau_i(t)) \\ &= \exp(-\Lambda_i(t, \tau_i(t))), \end{aligned} \quad (11)$$

where the last equality follows from a property of the inhomogeneous Poisson process (e.g., see p. 246 of [17]).

Finally, we derive $\bar{p}_i^{(\infty)}$. By (7), (10), and (11), we obtain

$$\bar{p}_i^{(\infty)} = \lim_{L \rightarrow \infty} \frac{1}{L-1} \sum_{\ell=2}^L \int_0^\infty P_i(t) \lambda_{i,\ell-1}(t) dt. \quad (12)$$

We will show that the event-average expression with Equation (12) is equivalent to the time-average expression,

$$\bar{p}_i^{(\infty)} = \frac{1}{\lambda_i} \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T P_i(t) \lambda_i(t) dt, \quad (13)$$

using $H = \lambda G$, an extension of Little's law. Let $G_\ell \equiv \int_0^\infty P_i(t) \lambda_{i,\ell}(t) dt$. Observe that G_ℓ denotes the miss probability of the $(\ell + 1)$ -th request for e_i . Let $H(t) \equiv \sum_{\ell=1}^\infty P_i(t) \lambda_{i,\ell}(t)$. Since $t_0^{(i)} \leq 0 < t_1^{(i)}$, there is a relationship between $\lambda_{i,\ell}(u)$ and $\lambda_i(u)$ for $u \geq 0$ such that $\lambda_i(u) = \sum_{\ell=1}^\infty \lambda_{i,\ell}(u)$. Hence, it follows that $H(t) = P_i(t) \lambda_i(t)$, which denotes the miss probability of a request for e_i given that the request is issued at time t , multiplied by $\lambda_i(t)$.

Since Ψ is time-asymptotically stationary and ergodic, $H = \lambda_i G$ holds (see Theorem 6.4 from [20]) for $G \equiv \lim_{L \rightarrow \infty} \frac{1}{L} \sum_{\ell=1}^L G_\ell$ and $H \equiv \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T H(t) dt$. Thus, we can conclude that Equation (13) is valid, which completes the proof of the theorem. ■

4.2 Accuracy of approximation with fluid limit

Now, we study the accuracy of approximating \bar{p}_i with $\bar{p}_i^{(\infty)}$. Let $r_i \equiv \lambda_i / \sum_{j=1}^N \lambda_j$ denote the fraction of the requests for e_i . We will estimate the overall average miss probability, $\bar{p} \equiv \sum_{i=1}^N r_i \bar{p}_i$, with a simulation, and we will compare it against $\bar{p}^{(\infty)} \equiv \sum_{i=1}^N r_i \bar{p}_i^{(\infty)}$ as evaluated numerically. Recall that $\bar{p}_i^{(\infty)}$ is defined

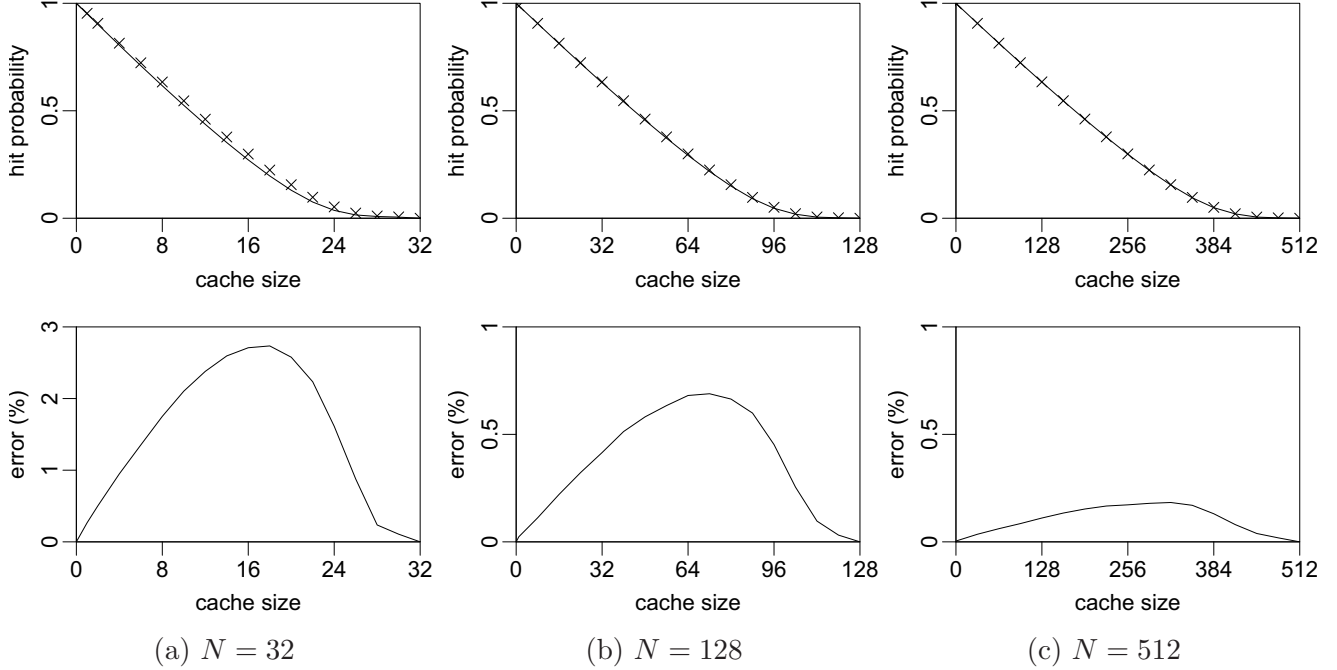


Figure 1: The accuracy of approximating \bar{p} with $\bar{p}^{(\infty)}$ when requests follow inhomogeneous Poisson processes, where N is set as shown in each column. In the top row, solid lines show $\bar{p}^{(\infty)}$, and \times marks show \bar{p} . The bottom row shows the error (%) of $\bar{p}^{(\infty)}$.

for each $\mathcal{T}_i^{(\infty)}$. We will refer to the formal average, $\bar{p}^{(\infty)}$, as the overall average miss probability in the fluid limit. The error (%) of $\bar{p}^{(\infty)}$ is defined as $100 |\bar{p}^{(\infty)} - \bar{p}|$.

For each data point, the simulation is run at least 20 times, where $10^4 N$ requests are generated in each run. Hence, on average, each item receives 10^4 requests in each run. When the 20 runs do not suffice to provide the confidence level that the estimated value is within 1% with probability 0.95, the simulation is repeated until this confidence level is achieved. Before the first run, we warm up the system by generating requests until every item is requested at least once. Each new run is started from the last state of the previous run.

In Figure 1, we consider the settings where the values of $\lambda_i(\cdot)$ for $1 \leq i \leq N$ fluctuate as trigonometric functions. Specifically, we set $\lambda_i(t) = 2 \sin^2\left(\frac{\pi}{4N}t + \frac{i}{8}\pi\right)$ for each e_i . Observe that, for any e_i , the period of $\lambda_i(\cdot)$ is $4N$ and its average rate is $\lambda_i = 1$, so that e_i is expected to be requested four times in a period. The phase of $\lambda_i(0)$ is chosen depending on $(i \bmod 8)$. Therefore, items are classified into eight types, and items with different types become popular (requested frequently) in different epochs.

The top row of Figure 1 shows $\bar{p}^{(\infty)}$ with solid lines and \bar{p} with \times marks. The number of items, N , is set as shown in each row. The horizontal axis represents the cache size, K . Although we have defined \bar{p} and $\bar{p}^{(\infty)}$ only for $1 \leq K \leq N - 1$, Figure 1 shows the range of $0 \leq K \leq N$. Here, we define $\bar{p} = \bar{p}^{(\infty)} = 0$ for $K = 0$ and $\bar{p} = \bar{p}^{(\infty)} = 1$ for $K = N$. Observe that the solid lines and the \times marks are on top of each other when $N \geq 128$. We can see that $\bar{p}^{(\infty)}$ slightly underestimates \bar{p} for $N = 32$.

To take a close look, the bottom row of Figure 1 shows the error (%) of $\bar{p}^{(\infty)}$. Observe that the error of $\bar{p}^{(\infty)}$ is within 3% for $N = 32$ and within 1% for $N \geq 128$. We find that, in general, the error of $\bar{p}^{(\infty)}$ is smaller for a larger N . This makes intuitive sense, since the original system approaches its fluid limit

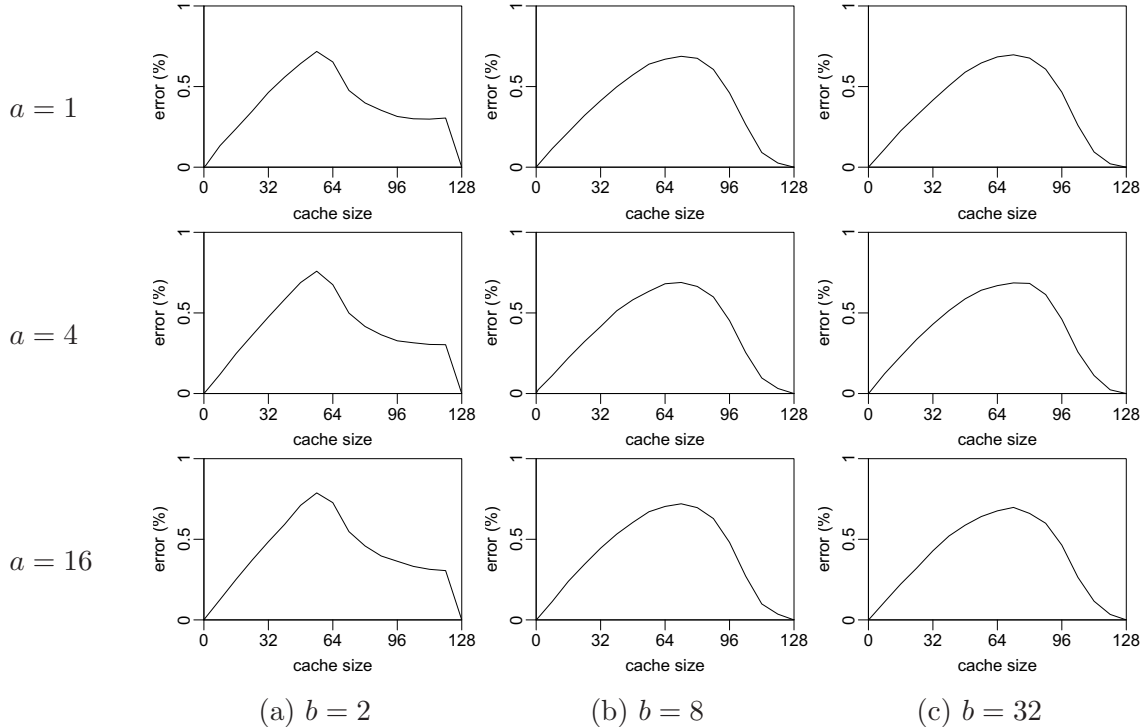


Figure 2: The error (%) of $\bar{p}^{(\infty)}$ for $N = 128$ when requests follow inhomogeneous Poisson processes.

as $N \rightarrow \infty$. We find that $\bar{p}^{(\infty)} \leq \bar{p}$ for all of the data points in Figure 1. Also, observe that the error of $\bar{p}^{(\infty)}$ tends to become smaller as K approaches 0 or N . It appears that, for a fixed N , the largest error is achieved when $K \approx N/2$.

In Figure 2, we consider $\lambda_i(\cdot)$ of the form, $\lambda_i(t) = 2 \sin^2\left(\frac{\pi}{aN}t + \frac{i}{b}\pi\right)$. We fix $N = 128$ and vary (a, b) as shown in the figure. Here, a denotes the expected number of requests for each item within a period, and b denotes the number of different types of requests. The settings of Figure 1(b) corresponds to the case with $a = 4$ and $b = 8$. Observe that the error (%) of $\bar{p}^{(\infty)}$ is not very sensitive to the particular (a, b) . This insensitivity is also found for other settings of N (figures omitted).

5 Asymptotic analysis with fluid limit

In this section, we study the request processes that are similar to those studied in [4, 11, 13, 21]. Specifically, let $J(\cdot)$ be a modulating stochastic-process on a general state space that determines the request rate for e_i at time t with $\lambda_i(J(t))$ for $1 \leq i \leq N$. Thus, given $J(\cdot)$, Ψ_i is an inhomogeneous Poisson process with rate $\lambda_i(J(t))$ at time t . Observe that Ψ_i for $i = 1, \dots, N$ are conditionally independent given $J(\cdot)$. We assume that Ψ is stationary, which is also assumed in [4, 11, 13, 21], so that the stationary miss probability exists (see Lemma 2.1 from [21]) and agrees with the average miss probability. In Section 5.1, we derive an explicit expression for $\bar{p}_i^{(\infty)}$ in this particular case. In Section 5.2, we study the limit as $K \rightarrow \infty$ when λ_i for $i = 1, 2, \dots$ have a heavy tail and a light tail, respectively.

5.1 Miss probability in the fluid limit

The purpose of this section is to derive $\bar{p}_i^{(\infty)}$ for the particular modulated request processes under consideration, which will be used in Section 5.2 to study the asymptotic characteristics.

Theorem 3 *Let Ψ be a modulated stochastic-process such that Ψ_i is an inhomogeneous Poisson process with rate $\lambda_i(J(t))$ at time t for $1 \leq i \leq N$, where $J(\cdot)$ is a modulating stochastic-process on a general state space. We assume that Ψ is stationary and ergodic and that $\lambda_i(J(t)) < \infty$ for any t and e_i . Then*

$$\lim_{m \rightarrow \infty} \bar{p}_i^{(m)} = \frac{1}{\lambda_i} \mathbf{E} [\exp(-\Lambda_i(\tau_i(K; J); J)) \lambda_i(J(0))],$$

where we define $\Lambda_i(u; J) \equiv \int_0^u \lambda_i(J(v)) dv$, and $\tau_i(K; J)$ is the maximum u such that

$$\sum_{j \neq i} (1 - \exp(-\Lambda_j(u; J))) \leq K - \frac{1}{2}. \quad (14)$$

Proof: Let $\bar{p}_i^{(m)}(J)$ be the conditional average miss probability for e_i in $\mathcal{T}_i^{(m)}$ given J . Then $\bar{p}_i^{(m)} = \mathbf{E} [\bar{p}_i^{(m)}(J)]$. Since $0 \leq \bar{p}_i^{(m)}(J) \leq 1$, the dominated convergence theorem can be used to show that

$$\lim_{m \rightarrow \infty} \bar{p}_i^{(m)} = \mathbf{E} \left[\lim_{m \rightarrow \infty} \bar{p}_i^{(m)}(J) \right].$$

By Theorem 2, we obtain

$$\bar{p}_i^{(m)} \rightarrow \frac{1}{\lambda_i} \mathbf{E} \left[\lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T e^{-\Lambda_i(t, \tau_i(t, K; J); J)} \lambda_i(J(t)) dt \right]$$

as $m \rightarrow \infty$, where we define $\Lambda_i(t, u; J) \equiv \int_t^{t+u} \lambda_i(J(v)) dv$, and $\tau_i(t, K; J)$ is the maximum u such that

$$\sum_{j \neq i} (1 - \exp(-\Lambda_j(t, u; J))) \leq K - \frac{1}{2}.$$

The pair $(\lambda_i(J(t)), \Lambda(t, \tau_i(t, K; J); J))$ has the same joint distribution as the pair $(\lambda_i(J(0)), \Lambda(0, \tau_i(0, K; J); J))$, since $J(\cdot)$ is stationary. Therefore, we obtain

$$\begin{aligned} \lim_{m \rightarrow \infty} \bar{p}_i^{(m)} &= \frac{1}{\lambda_i} \mathbf{E} \left[\lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T e^{-\Lambda_i(0, \tau_i(0, K; J); J)} \lambda_i(J(0)) dt \right] \\ &= \frac{1}{\lambda_i} \mathbf{E} \left[e^{-\Lambda_i(0, \tau_i(0, K; J); J)} \lambda_i(J(0)) \right], \end{aligned}$$

which proves the theorem, since $\Lambda_i(0, u; J) = \Lambda_i(u; J)$ and $\tau_i(0, K; J) = \tau_i(K; J)$. \blacksquare

When $J(\cdot)$ is constant, each Ψ_i is an independent Poisson process. The following corollary can be compared against the stationary miss probabilities in the fluid limits obtained in [10, 9].

Corollary 1 *If Ψ_i is an independent Poisson process with rate λ_i for each e_i , then $\bar{p}_i^{(m)} \rightarrow \exp(-\lambda_i \tau_i(K))$ as $m \rightarrow \infty$, where $\tau_i(K) = C_i^{-1}(K - 1/2)$, and $C_i^{-1}(\cdot)$ is the inverse function of*

$$C_i(t) \equiv \sum_{j \neq i} (1 - \exp(-\lambda_j t)). \quad (15)$$

The corollary can be understood as follows. Suppose that $e_{i,1}$ is requested and moved to the head of the MTF list at time 0. Then, until $e_{i,1}$ is requested again, the position of $e_{i,1}$ in the MTF list of $\mathcal{T}_i^{(\infty)}$ is $C_i(t)$ at time t . Observe that the term, $1 - \exp(-\lambda_j t)$, in the right-hand side of Equation (15) is the probability that, in the original system, e_j is requested in the interval $(0, t)$. Also, this term agrees with the fraction of $e_{j,k}$ for $1 \leq k \leq m$ that are requested in $(0, t)$ as $m \rightarrow \infty$. In $\mathcal{T}_i^{(\infty)}$, the position of $e_{i,1}$ reaches $K - 1/2$ at time $\tau_i(K)$. The probability that the next request for $e_{i,1}$ is issued after time $\tau_i(K)$ is $\exp(-\lambda_i \tau_i(K))$. Note that, in the MTF list of $\mathcal{T}_i^{(\infty)}$, $e_{i,1}$ moves up following a deterministic function until $e_{i,1}$ is requested at a random time.

Since our fluid limit differs from the fluid limits defined in [10, 9], our $\bar{p}_i^{(\infty)}$ differs from those derived in [10, 9]. However, the only difference between our $\bar{p}_i^{(\infty)}$ and that in [9] is that, in [9], $\tau_i(K)$ is replaced with $\tau(K) = C^{-1}(K)$, where $C^{-1}(\cdot)$ is the inverse function of $C(t) = \sum_{j=1}^N (1 - \exp(-\lambda_j t))$. The differences between the fluid limits in [10] and [9] are discussed in [9]. We find that these differences become negligible when we study the asymptotic characteristics as in Section 5.2. In Appendix B, we also compare our $\bar{p}_i^{(\infty)}$ against that in [9] for a range of K .

5.2 Asymptotic characteristics of miss probability

In this section, we consider the overall average miss probability in the fluid limit, $\bar{p}^{(\infty)}(K) \equiv \sum_{i=1}^{\infty} r_i \bar{p}_i^{(\infty)}$, for a cache size K as $K \rightarrow \infty$. We assume that $N = \infty$ and that $\sum_{j=1}^N \lambda_j = 1$ (without loss of generality), so that $r_i = \lambda_i$.

We consider the request process studied in Section 5.1 for the case when $J(\cdot)$ is an ergodic semi-Markov chain on a finite state space. Our results rely on the following properties of the $J(\cdot)$

Lemma 3 *Let $J(\cdot)$ be an ergodic semi-Markov chain on a finite state space. Then almost surely*

$$\left| \frac{1}{t} \int_0^t \lambda_i(J(u)) du - \lambda_i \right| / \lambda_i \rightarrow 0$$

as $t \rightarrow \infty$ uniformly for $i = 1, 2, \dots$

Proof: Let $\{1, \dots, \Upsilon\}$ be the state space of J . For $1 \leq v \leq \Upsilon$, let $V_v(t; J)$ be the time that J spends at state v by time t . Then, for any i ,

$$\int_0^t \lambda_i(J(u)) du = \sum_{v=1}^{\Upsilon} \lambda_i(v) V_v(t, J). \quad (16)$$

Since J is an ergodic semi-Markov chain, almost surely $V_v(t; J)/t \rightarrow \pi_v$ as $t \rightarrow \infty$, where π_v is the stationary probability. Notice that

$$\lambda_i = \sum_{v=1}^{\Upsilon} \lambda_i(v) \pi_v, \quad (17)$$

for any i . By Equation (16) and Equation (17), we obtain

$$\left| \frac{1}{t} \int_0^t \lambda_i(J(u)) du - \lambda_i \right| = \left| \frac{1}{t} \sum_{v=1}^{\Upsilon} \lambda_i(v) V_v(t, J) - \sum_{v=1}^{\Upsilon} \lambda_i(v) \pi_v \right| \quad (18)$$

$$\leq \sum_{v=1}^{\Upsilon} \lambda_i(v) \left| \frac{V_v(t, J)}{t} - \pi_v \right| \quad (19)$$

Since Υ is finite, we have that, for any ε , there exists T such that, for all $t > T$, it holds that almost surely $|V_v(t; J)/t - \pi_v| < \varepsilon \pi_v$ for $1 \leq v \leq \Upsilon$. Therefore, Inequality (19) suggests

$$\left| \frac{1}{t} \int_0^t \lambda_i(J(u)) du - \lambda_i \right| \leq \varepsilon \lambda_i$$

almost surely for all $t > T$. We complete the proof of the lemma by taking $\varepsilon \rightarrow 0$. \blacksquare

5.2.1 Heavy tail

We first consider the case when λ_i has a heavy tail so that $\lambda_i \sim c/i^\alpha$ for $i = 1, 2, \dots$, where $\alpha > 1$, $c > 0$, and $a_i \sim b_i$ denotes $\lim_{i \rightarrow \infty} a_i/b_i = 1$. Also, let $a_i \lesssim b_i$ denote $\lim_{i \rightarrow \infty} a_i/b_i \leq 1$. We will use the following two lemmas, where $\Gamma(z) \equiv \int_0^\infty e^{-y} y^{z-1} dy$ denotes the gamma function.

Lemma 4 *Let $g_i(t) = \sum_{j \neq i} (1 - \exp(-\lambda_j t))$. If $\lambda_i \sim c/i^\alpha$ with $\alpha > 1$ and $c > 0$, then $g_i(t) \sim c^{1/\alpha} \Gamma(1 - 1/\alpha) t^{1/\alpha}$.*

Lemma 5 *Let $f(t) = \sum_{i=1}^\infty \lambda_i \exp(-\lambda_i t)$. If $\lambda_i \sim c/i^\alpha$ with $\alpha > 1$ and $c > 0$, then $f(t) \sim c^{1/\alpha} \alpha^{-1} \Gamma(1 - 1/\alpha) t^{-1+1/\alpha}$.*

Lemma 4 is a direct consequence of Corollary 1 from [10]. Lemma 5 can be proved similarly to Lemma 3.1 from [21], and we provide a proof of Lemma 5 in Appendix C. The following theorem characterizes $\bar{p}^{(\infty)}(K)$ for a large K .

Theorem 4 *In addition to the conditions in Theorem 3, if $J(\cdot)$ is an ergodic semi-Markov chain on a finite state space and $\lambda_i \sim c/i^\alpha$ for $i = 1, 2, \dots$, where $\alpha > 1$ and $c > 0$, then $\bar{p}^{(\infty)}(K)$ is asymptotically insensitive to $J(\cdot)$ as $K \rightarrow \infty$, and it holds that*

$$\bar{p}^{(\infty)}(K) \sim \frac{c \Gamma(1 - 1/\alpha)^\alpha}{\alpha} K^{1-\alpha},$$

where $\Gamma(z) \equiv \int_0^\infty e^{-y} y^{z-1} dy$ denotes the gamma function.

Proof: Let $C_i(t; J) \equiv \sum_{j \neq i} (1 - \exp(-\lambda_j t))$ be the left-hand side of Inequality (14). Then Lemma 3 implies that, for any ε , there exists t_0 such that, for all $t > t_0$, we have that almost surely

$$\sum_{j \neq i} (1 - \exp(-(1 - \varepsilon)\lambda_j t)) \leq C_i(t; J) \leq \sum_{j \neq i} (1 - \exp(-(1 + \varepsilon)\lambda_j t)) \quad (20)$$

for any e_i . Hence, Lemma 4 suggests that almost surely

$$\begin{aligned} C_i(t; J) &\lesssim (1 + \varepsilon)^{1/\alpha} c^{1/\alpha} \Gamma(1 - 1/\alpha) t^{1/\alpha} \\ C_i(t; J) &\gtrsim (1 - \varepsilon)^{1/\alpha} c^{1/\alpha} \Gamma(1 - 1/\alpha) t^{1/\alpha}. \end{aligned}$$

uniformly for $i = 1, 2, \dots$. Taking $\varepsilon \rightarrow 0$, we obtain that almost surely

$$C_i(t; J) \sim c^{1/\alpha} \Gamma(1 - 1/\alpha) t^{1/\alpha} \quad (21)$$

uniformly for $i = 1, 2, \dots$

Recall that $\tau_i(K; J)$ is the maximum u such that $C_i(t; J) \leq K - 1/2$. Since $\tau_i(K; J) \rightarrow \infty$ almost surely as $K \rightarrow \infty$, Relation (21) implies that almost surely

$$K \sim C_i(\tau_i(K); J) \sim c^{1/\alpha} \Gamma(1 - 1/\alpha) (\tau_i(K; J))^{1/\alpha}.$$

Hence, we obtain that almost surely

$$\tau_i(K; J) \sim \tau(K) \equiv \frac{K^\alpha}{c \Gamma(1 - 1/\alpha)^\alpha} \quad (22)$$

uniformly for $i = 1, 2, \dots$

Finally, we consider $\bar{p}^{(\infty)}$. By Theorem 3, we have

$$\bar{p}^{(\infty)}(K) = \sum_{i=1}^{\infty} \mathbf{E} [\exp(-\Lambda_i(\tau_i(K; J); J)) \lambda_i(J(0))]. \quad (23)$$

The uniform convergence of Relation (22) and Lemma 3 imply that, for any ε , there exists K_0 such that, for all $K > K_0$, we have that almost surely

$$(1 - \varepsilon) \lambda_i \tau_i(K; J) \leq \Lambda_j(\tau_i(K; J); J) \leq (1 + \varepsilon) \lambda_i \tau_i(K; J) \quad (24)$$

for $i = 1, 2, \dots$. Now, (22), (23), and (24) imply that, for all $K > K_0$, we have

$$\begin{aligned} \bar{p}^{(\infty)}(K) &\leq \sum_{i=1}^{\infty} \mathbf{E} [\exp(-(1 - \varepsilon) \lambda_i \tau(K)) \lambda_i(J(0))] \\ &= \sum_{i=1}^{\infty} \exp(-(1 - \varepsilon) \lambda_i \tau(K)) \mathbf{E} [\lambda_i(J(0))] \\ &= \frac{1}{1 - \varepsilon} \sum_{i=1}^{\infty} (1 - \varepsilon) \lambda_i \exp(-(1 - \varepsilon) \lambda_i \tau(K)), \end{aligned} \quad (25)$$

where the last equality follows from the stationality of $J(\cdot)$. By Lemma 5, we obtain

$$\bar{p}^{(\infty)}(K) \lesssim (1 - \varepsilon)^{1/\alpha - 1} c^{1/\alpha} \alpha^{-1} \Gamma(1 - 1/\alpha) t^{-1+1/\alpha}.$$

Similarly, we obtain the asymptotic lower bound:

$$\bar{p}^{(\infty)}(K) \gtrsim (1 + \varepsilon)^{1/\alpha - 1} c^{1/\alpha} \alpha^{-1} \Gamma(1 - 1/\alpha) t^{-1+1/\alpha}.$$

Taking $\varepsilon \rightarrow 0$, we complete the proof of the theorem. \blacksquare

Theorem 4, which is obtained for the fluid limit, is in agreement with the asymptotic results for the original system derived in [11, 21].

5.2.2 Light tail

In this section, we study $\bar{p}^{(\infty)}(K)$ as $K \rightarrow \infty$ for the case when λ_i has a light tail so that $\lambda_i \sim c \exp(-\xi i^\beta)$, where $c, \xi, \beta > 0$. This case has not been fully investigated in the prior work. Jelenković [10] studies asymptotic properties of the overall stationary miss probability in his fluid limit when λ_i has a light tail, assuming that requests follow the independent reference model (equivalently, independent Poisson processes), but no asymptotic results are known for other request processes. We will use the following two lemmas:

Lemma 6 *Let $g(t) \equiv \int_0^\infty (1 - \exp(-\lambda_x t)) dx + \delta$, where $|\delta| < \infty$. If $\lambda_x \sim c \exp(-\xi x^\beta)$, where $c, \xi, \beta > 0$, then $g^{-1}(v + \delta') \sim e^{-\gamma} c^{-1} \exp(\xi v^\beta)$ for any $|\delta'| < \infty$.*

Lemma 7 *Let $f(t) \equiv \int_0^\infty \lambda_x \exp(-\lambda_x t) dx$. If $\lambda_x \sim c \exp(-\xi x^\beta)$, where $c, \xi, \beta > 0$, then $f(t) \sim (\ln(ct))^{1/\beta-1} \xi^{-1/\beta} \beta^{-1} t^{-1}$.*

Lemma 6 is a trivial extension of Lemma 6 from [10], and Lemma 7 equivalent to Lemma 3 from [10] by Equation (7.49) from [10].

Theorem 5 *In addition to the conditions in Theorem 3, if $J(\cdot)$ is an ergodic semi-Markov chain on a finite state space and $\lambda_i \sim c \exp(-\xi i^\beta)$ for $i = 1, 2, \dots$, where $c, \xi, \beta > 0$, then $\bar{p}^{(\infty)}(K)$ is asymptotically insensitive to $J(\cdot)$ as $K \rightarrow \infty$, and it holds that*

$$\bar{p}^{(\infty)}(K) \sim \frac{c e^\gamma}{\beta \xi} K^{1-\beta} \exp(-\xi K^\beta),$$

where $\gamma \equiv \int_0^\infty \exp(-y) \ln y dy \approx 0.577$ is Euler's constant.

Our proof of Theorem 5 follows a slightly different procedure than that of Theorem 4. In Theorem 4, an asymptotic expression for $\tau_i(K; J)$ in Relation (22) is obtained from an asymptotic expression for $C_i(t; J)$ in Relation (21). If we were to follow the same procedure as the proof of Theorem 4, the asymptotic upper bound would not match the asymptotic lower bound in Theorem 5. This difference stems from the fact that $C_i(t; J)$ is asymptotically polynomial in Theorem 4 and asymptotically poly-logarithmic in Theorem 5. Therefore, we will study the asymptotic property of the inverse function of $C_i(t; J)$, which is the same approach as the proof of Theorem 6 from [10].

Proof of Theorem 5: We first study $\tau_i(K; J) = C_i^{-1}(K - 1/2; J)$, where $C_i^{-1}(\cdot; J)$ is the inverse function of $C_i(\cdot; J)$. Observe that Inequality (20) remains valid when λ_i has a light tail. Let

$$\begin{aligned} C(t, \varepsilon) &\equiv \sum_{j=1}^{\infty} (1 - \exp(-(1 + \varepsilon) \lambda_j t)) \\ &= \int_0^\infty (1 - \exp(-(1 + \varepsilon) \lambda_x t)) dx, \end{aligned}$$

where we extend the domain of λ_i to non-negative real numbers, so that $\lambda_x = \lambda_{\lceil x \rceil}$. Note that $\lambda_x \sim c \exp(-\xi x^\beta)$. Let $D(t, \varepsilon) \equiv C(t, -\varepsilon) - 1$. Then Inequality (20) implies that almost surely $D(t, \varepsilon) \leq C_i(t; J) \leq C(t, \varepsilon)$ for any i and t . Let $C^{-1}(\cdot, \varepsilon)$ and $D^{-1}(\cdot, \varepsilon)$, respectively, denote the inverse functions of $C(\cdot, \varepsilon)$ and $D(\cdot, \varepsilon)$. Since $C_i(\tau_i(K; J); J) = K - 1/2$, we have that almost surely $C^{-1}(K - 1/2, \varepsilon) \leq \tau_i(K; J) \leq D^{-1}(K - 1/2, \varepsilon)$.

Let $\tau(K) \equiv e^{-\gamma} c^{-1} \exp(\xi K^\beta)$. Applying Lemma 6 with $\delta = 0$ and $\delta' = -1$, we obtain $C^{-1}(K - 1/2, \varepsilon) \sim \tau(K)/(1 + \varepsilon)$. Applying Lemma 6 with $\delta = -1$ and $\delta' = -1$, we obtain $D^{-1}(K - 1/2, \varepsilon) \sim \tau(K)/(1 - \varepsilon)$. Taking $\varepsilon \rightarrow 0$, we obtain that $\tau_i(K; J) \sim \tau(K)$ uniformly for $i = 1, 2, \dots$

The uniform convergence of $\tau_i(K; J)$ and Lemma 3 imply that Inequality (25) remains valid when λ_i has a light tail. Hence, Lemma 7 implies that

$$\bar{p}^{(\infty)}(K) \lesssim \frac{1}{1 - \varepsilon} \frac{(\ln((1 - \varepsilon) c \tau(K)))^{1/\beta - 1}}{\xi^{1/\beta} \beta \tau(K)}.$$

Substituting $\tau(K)$, we obtain

$$\bar{p}^{(\infty)}(K) \lesssim \frac{1}{1 - \varepsilon} \frac{c e^\gamma}{\xi \beta} K^{1 - \beta} \exp(-\xi K^\beta) \left(1 - \frac{\gamma - \ln(1 - \varepsilon)}{\xi K^\beta}\right)^{1/\beta - 1}.$$

Similarly, we obtain an asymptotic lower bound:

$$\bar{p}^{(\infty)}(K) \gtrsim \frac{1}{1 + \varepsilon} \frac{c e^\gamma}{\xi \beta} K^{1 - \beta} \exp(-\xi K^\beta) \left(1 - \frac{\gamma - \ln(1 + \varepsilon)}{\xi K^\beta}\right)^{1/\beta - 1}.$$

Now the theorem follows by taking $\varepsilon \rightarrow 0$. \blacksquare

6 Conclusion

We have introduced and demonstrated the usefulness of the fluid limit of a stochastic model for LRU with possibly non-stationary and dependent request processes. In particular, our numerical experiments show that the average miss probability derived in the fluid limit closely approximates that in the original system for a moderate cache size. For a large cache, we find that the average miss probability in the fluid limit often has the same asymptotic characteristics as those in the original system and that the asymptotic analysis is often simpler in the fluid limit than in the original system.

Our expectation is that the fluid limit and the average miss probability derived in the fluid limit will find applications beyond those investigated in this paper. An interesting future direction is to seek an optimal cache algorithm with dependent and non-stationary request processes in the fluid limit. To this end, Hirade and Osogami [9] show that, in a fluid limit, the 2Q cache algorithm [14] can be made to have a lower miss probability than LRU by choosing the right value of the parameter of 2Q, assuming that the requests follow independent Poisson processes. However, it is also shown that the 2Q that has the minimum stationary miss probability can have a high transient miss probability, which suggests the importance of studying the optimality with non-stationary request processes.

References

- [1] A. Borodin and R. El-Yaniv. *Online computation and competitive analysis*. Cambridge University Press, Cambridge, UK, 1998.
- [2] P. J. Burville and J. F. C. Kingman. On a model for storage and search. *Journal of Applied Probability*, 10(3):697–701, 1973.

- [3] J.-H. Chu and G. D. Knott. A new method for computing page-fault rates. *SIAM Journal on Computing*, 22(6):1319–1330, 1993.
- [4] E. G. Coffman and P. Jelenković. Performance of the move-to-front algorithm with Markov-modulated request sequences. *Operations Research Letters*, 25(3):109–118, 1999.
- [5] J. A. Fill. Limits and rates of convergence for the distribution of search cost under the move-to-front rule. *Theoretical Computer Science*, 164(1-2):185–206, 1996.
- [6] J. A. Fill and L. Holst. On the distribution of search cost for the move-to-front rule. *Random Structures & Algorithms*, 8(3):179–186, 1996.
- [7] P. Flajolet, D. Gardy, and L. Thimonier. Birthday paradox, coupon collectors, caching algorithms and self-organizing search. *Discrete Applied Mathematics*, 39(3):207–229, 1992.
- [8] B. Fristedt and L. Gray. *A Modern Approach to Probability Theory*. Birkhäuser, Boston, MA, 1997.
- [9] R. Hirade and T. Osogami. Analysis of page replacement policies in the fluid limit. Technical Report RT0768 (submitted for publication), IBM Research, 2007.
- [10] P. R. Jelenković. Asymptotic approximation of the move-to-front search cost distribution and least-recently-used caching fault probabilities. *Annals of Applied Probability*, 9(2):430–464, 1999.
- [11] P. R. Jelenković and A. Radovanović. Least-recently-used caching with dependent requests. *Theoretical Computer Science*, 326(1-3):293–327, 2004.
- [12] P. R. Jelenković and A. Radovanović. The persistent-access-caching algorithm. *Random Structures & Algorithms*, in press, 2008.
- [13] P. R. Jelenković, A. Radovanović, and M. S. Squillante. Critical sizing of LRU caches with dependent requests. *Journal of Applied Probability*, 43(4):1013–1027, 2006.
- [14] T. Johnson and D. Shasha. 2Q: A low overhead high performance buffer management replacement algorithm. In *Proceedings of the 20th International Conference on Very Large Data Bases*, pages 439–450, September 1994.
- [15] K. Lam, M. Leung, and M. Siu. Self-organizing files with dependent accesses. *Journal of Applied Probability*, 21(2):343–359, 1984.
- [16] J. McCabe. On serial files with relocatable records. *Operations Research*, 13(4):609–618, 1965.
- [17] R. Nelson. *Probability, Stochastic Processes, and Queueing Theory*. Springer-Verlag, New York, NY, 1995.
- [18] S. Podlipnig and L. Böszörményi. A survey of Web cache replacement strategies. *ACM Computing Surveys*, 35(4):374–398, 2003.
- [19] E. R. Rodrigues. The performance of the move-to-front scheme under some particular forms of Markov requests. *Journal of Applied Probability*, 32(4):1089–1102, 1995.
- [20] K. Sigman. *Stationary Marked Point Processes: An Intuitive Approach*. Chapman & Hall, New York, NY, 1995.

- [21] T. Sugimoto and N. Miyoshi. On the asymptotics of fault probability in least-recently-used caching with Zipf-type request distribution. *Random Structures & Algorithms*, 29(3):296–323, 2006.
- [22] S. Tewari and L. Kleinrock. Proportional replication in peer-to-peer networks. In *Proceedings of the IEEE INFOCOM*, pages 1–12, April 2006.
- [23] L. Ying, Z. Liu, D. Towsley, and C. H. Xia. Distributed operator placement and data caching in large-scale sensor networks. In *Proceedings of the IEEE INFOCOM*, pages 977–985, April 2008.

A Fluid limit defined with independent replications

In this section, we prove that the dependencies in Ψ would disappear in the fluid limit defined with $\mathcal{S}^{(m)}$. Recall that, in $\mathcal{S}^{(m)}$, Φ_k for $1 \leq k \leq m$ are independent and have the distribution identical to Ψ . In $\mathcal{S}^{(\infty)}$, we have the following lemma corresponding to Lemma 2.

Lemma 8 *Let $C_{i,k,\ell}^{(m)}$ be the total size of distinct items that are requested after $t_{\ell-1}^{(i,k)}$ and before $t_{\ell}^{(i,k)}$ in $\mathcal{S}^{(m)}$. Then, as $m \rightarrow \infty$, $C_{i,k,\ell}^{(m)} \Rightarrow \sum_{j=1}^N \mathbf{E}[H_{\ell}(j)]$, where we define the indicator random variable, $H_{\ell}(j)$, as follows. Consider a system, \mathcal{I}_i , having $N+1$ items, e_j for $0 \leq j \leq N$. For each e_j , let $\Psi'_j = \{u_{\ell}^{(j)}, \ell \in \mathbf{Z}\}$ be the request process for e_j in \mathcal{I}_i . In \mathcal{I}_i , we assume that Ψ'_j for $0 \leq j \leq N$ are independent, Ψ'_j has the same distribution as Ψ_j for $1 \leq j \leq N$, and Ψ'_0 has the same distribution as Ψ_i . For $1 \leq j \leq N$, let $H_{\ell}(j)$ be the indicator random variable such that $H_{\ell}(j) = 1$ iff e_j is requested after $u_{\ell-1}^{(0)}$ and before $u_{\ell}^{(0)}$ in \mathcal{I}_i .*

Observe that $C_{i,k,\ell}^{(\infty)}$ does not depend on the dependencies in the Ψ .

Proof of Lemma 8: Without loss of generality, we study only the convergence of $C_{i,1,\ell}^{(m)}$. We prove the convergence in distribution by showing the convergence of the Laplace transform, $\psi_{i,\ell}^{(m)}(s)$ for $0 \leq s < \infty$, of $C_{i,1,\ell}^{(m)}$. Let $H_{\ell}(j,k)$ be the indicator random variable such that $H_{\ell}(j,k) = 1$ iff $e_{j,k}$ is requested after $t_{\ell-1}^{(i,1)}$ and before $t_{\ell}^{(i,1)}$. By definition, $H_{\ell}(i,1) = 0$. Then

$$\psi_{i,\ell}^{(m)}(s) = \mathbf{E} \left[\exp \left(-\frac{s}{m} \sum_{j=1}^N \sum_{k=1}^m H_{\ell}(j,k) \right) \right]$$

Since Φ_k for $1 \leq k \leq m$ are independent, we obtain

$$\psi_{i,\ell}^{(m)}(s) = \prod_{k=1}^m \mathbf{E} \left[\exp \left(-\frac{s}{m} \sum_{j=1}^N H_{\ell}(j,k) \right) \right].$$

Since Φ_k for $1 \leq k \leq m$ are identically distributed but $H_{\ell}(j,k)$ is defined with respect to the interval $(t_{\ell-1}^{(i,1)}, t_{\ell}^{(i,1)})$, we obtain

$$\psi_{i,\ell}^{(m)}(s) = \mathbf{E} \left[\exp \left(-\frac{s}{m} \sum_{j=1}^N H_{\ell}(j,2) \right) \right]^{m-1} \mathbf{E} \left[\exp \left(-\frac{s}{m} \sum_{j=1}^N H_{\ell}(j,1) \right) \right].$$

For $0 \leq r \leq N - 1$, let $q_r = \mathbf{P} \left(\sum_{j=1}^N H_\ell(j, 2) = r \right)$ be the probability that r distinct items in E_2 are requested in the interval $(t_{\ell-1}^{(i,1)}, t_\ell^{(i,1)})$. Then

$$\psi_{i,\ell}^{(m)}(s) = \left(\sum_{r=0}^{N-1} q_r \exp(-s r/m) \right)^{m-1} \mathbf{E} \left[\exp \left(-\frac{s}{m} \sum_{j=1}^N H_\ell(j, 1) \right) \right].$$

By Lemma 9 in Appendix C, we obtain

$$\lim_{m \rightarrow \infty} \psi_{i,\ell}^{(m)}(s) = \exp \left(-s \sum_{r=0}^{N-1} r q_r \right) \lim_{m \rightarrow \infty} \mathbf{E} \left[\exp \left(-\frac{s}{m} \sum_{j=1}^N H_\ell(j, 1) \right) \right].$$

By the dominated convergence theorem, we can exchange the limit and the expectation, so that

$$\begin{aligned} \lim_{m \rightarrow \infty} \psi_{i,\ell}^{(m)}(s) &= \exp \left(-s \sum_{r=0}^{N-1} r q_r \right) \mathbf{E} \left[\lim_{m \rightarrow \infty} \exp \left(-\frac{s}{m} \sum_{j=1}^N H_\ell(j, 1) \right) \right] \\ &= \exp \left(-s \sum_{r=0}^{N-1} r q_r \right) \end{aligned}$$

Observe that the summation in the last expression denotes an expectation. Hence, we obtain

$$\lim_{m \rightarrow \infty} \psi_{i,\ell}^{(m)}(s) = \exp \left(-s \mathbf{E} \left[\sum_{j=1}^N H_\ell(j, 2) \right] \right).$$

Therefore, the continuity theorem (e.g., see p. 262 from [8]) implies that $C_{i,\ell}^{(m)} \Rightarrow \mathbf{E} \left[\sum_{j=1}^N H_\ell(j, 2) \right]$ as $m \rightarrow \infty$. Since $H_\ell(j, 2)$ and $H_\ell(j)$ have the same distribution, this completes the proof of the lemma via the linearity of expectation. ■

The ℓ -th request for $e_{i,k}$ is a miss in $\mathcal{S}^{(m)}$ iff $C_{i,\ell}^{(m)} \geq K$. Hence, the miss probability of the ℓ -th request for $e_{i,k}$ in $\mathcal{S}^{(m)}$ is $p_{i,k,\ell}^{(m)} = \mathbf{P} \left(C_{i,k,\ell}^{(m)} \geq K \right)$. The corresponding average miss probability is defined as follows:

$$\bar{p}_{i,k}^{(m)} = \lim_{L \rightarrow \infty} \frac{1}{L} \sum_{\ell=1}^L p_{i,k,\ell}^{(m)}.$$

The next corollary can be proved in the same way as Theorem 1 using Lemma 8

Corollary 2

$$\lim_{m \rightarrow \infty} \bar{p}_{i,k}^{(m)} = \mathbf{P}^{0,i} \left(\sum_{j=1}^N \mathbf{E} \left[I_{\{t_1^{(j)} < t_1^{(0)}\}} \right] \geq K \right).$$

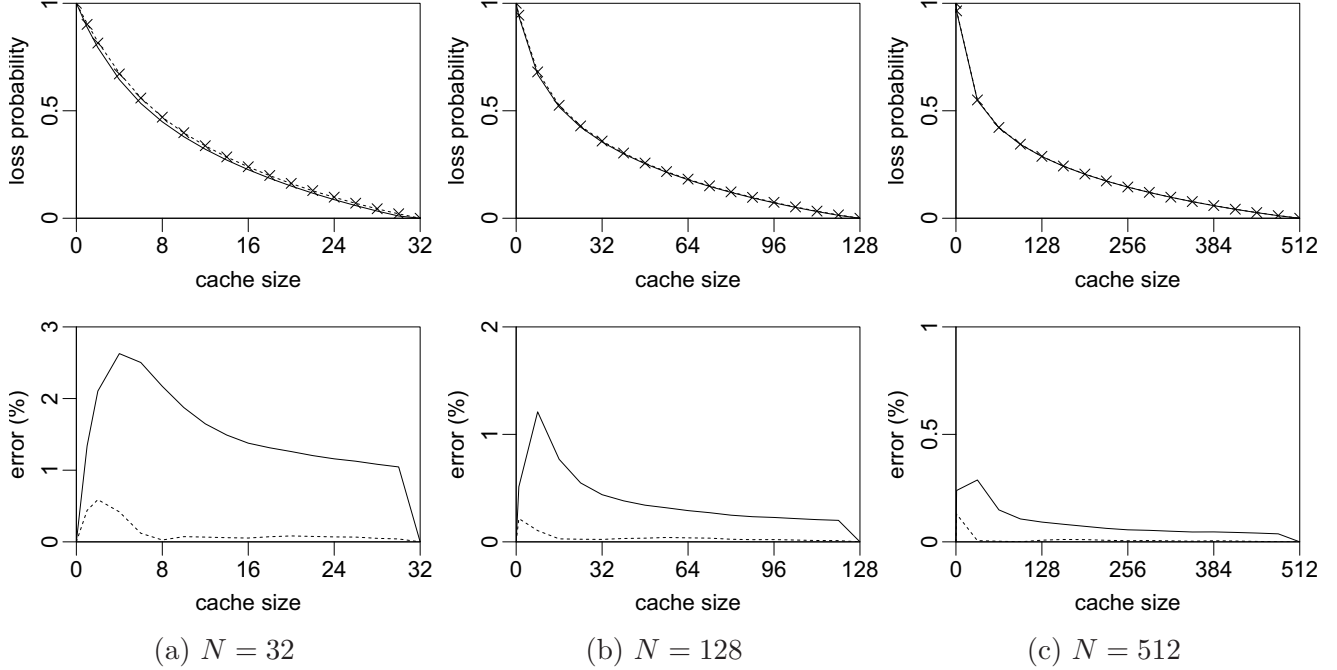


Figure 3: The accuracy of approximating \bar{p} with $\bar{p}^{(\infty)}$ and with $\bar{q}^{(\infty)}$ when requests follow Poisson processes and λ_i has Zipf's law. N is set as shown in each column. In top row, solid lines show $\bar{p}^{(\infty)}$, dashed lines show $\bar{q}^{(\infty)}$, and \times marks show \bar{p} . The bottom row shows the error (%) of $\bar{p}^{(\infty)}$ with solid lines and that of $\bar{q}^{(\infty)}$ with dashed lines.

B Accuracy of approximation with fluid limit under Poisson requests

In this section, we study the accuracy of approximating \bar{p} with $\bar{p}^{(infty)}$ when requests follow Poisson processes. See Corollary 1 for the expression of $\bar{p}^{(infty)}$. In [9], Hirade and Osogami derive, $\bar{q}^{(infty)}$, the overall loss probability in their fluid limit. In particular, we will compare $\bar{p}^{(\infty)}$ against $\bar{q}^{(\infty)}$. We do not consider the corresponding expression derived by Jelenković [10], but $\bar{q}^{(\infty)}$ is a special case of the general expression in [10]. We define the error (%) of $\bar{q}^{(\infty)}$ as $100|\bar{q}^{(\infty)} - \bar{p}|$, similar to the definition of the error of $\bar{p}^{(\infty)}$ in Section 4.2. The conditions of the simulation to evaluate \bar{p} are the same as in Section 4.2.

In Figure 3, we consider the case where λ_i has Zipf's law so that $\lambda_i = 1/i$ for $1 \leq i \leq N$. The top row shows $\bar{p}^{(\infty)}$ with solid lines, $\bar{q}^{(\infty)}$ with dashed lines, and \bar{p} with \times marks. The number of items is varied as indicated in each column. Observe that the two lines and the \times marks are on top of each other when $N \geq 128$. We find that, in general, the error of $\bar{p}^{(\infty)}$ and the error of $\bar{q}^{(\infty)}$ are smaller for a larger N .

To take a close look, the bottom row shows the error (%) of $\bar{p}^{(\infty)}$ with solid lines and that of $\bar{q}^{(\infty)}$ with dashed lines. We find that, in general, the error of $\bar{p}^{(\infty)}$ is larger than the error of $\bar{q}^{(\infty)}$. However, the error of $\bar{p}^{(\infty)}$ is within 3% even with $N = 32$ and become negligible for a large N . The error of $\bar{p}^{(\infty)}$ is also sensitive to the cache size, K . It appears that the error of $\bar{p}^{(\infty)}$ is small when K is close to 0 or close to N . The largest error is observed when K is small ($K < N/4$) but not too close to 0.

In Figure 4, we consider the case where $\lambda_i = 1/N^{\frac{i-1}{N-1}}$ for $1 \leq i \leq N$. Thus, λ_i for $1 \leq i \leq N$ is a geometric series. We choose $\lambda_1 = 1$ and $\lambda_N = 1/N$, so that the largest λ_i and the smallest λ_i agree with those in the settings of Figure 3. Similar to Figure 3, the top row shows $\bar{p}^{(\infty)}$, $\bar{q}^{(\infty)}$, and \bar{p} , and the bottom row shows the error of $\bar{p}^{(\infty)}$ and the error of $\bar{q}^{(\infty)}$.

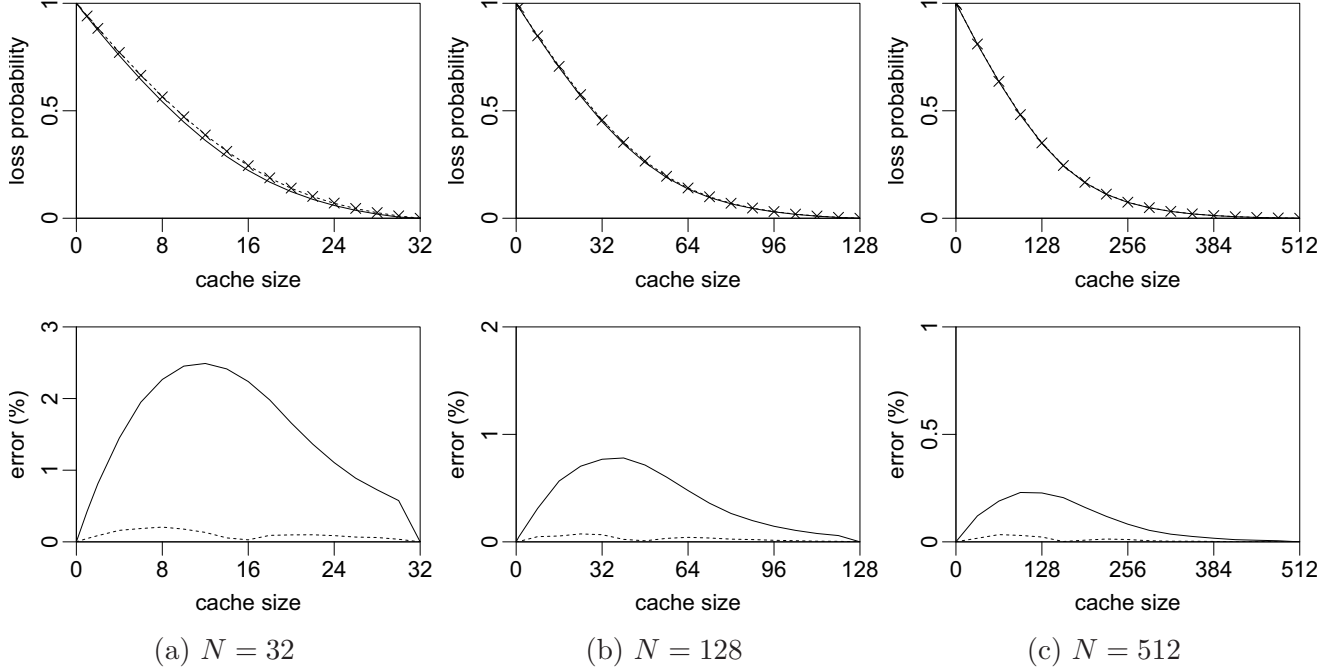


Figure 4: The accuracy of approximating \bar{p} with $\bar{p}^{(\infty)}$ and with $\bar{q}^{(\infty)}$ when requests follow Poisson processes and λ_i for $1 \leq i \leq N$ is a geometric series. N is set as shown in each column. In top row, solid lines show $\bar{p}^{(\infty)}$, dashed lines show $\bar{q}^{(\infty)}$, and \times marks show \bar{p} . The bottom row shows the error (%) of $\bar{p}^{(\infty)}$ with solid lines and that of $\bar{q}^{(\infty)}$ with dashed lines.

Observe that \bar{p} in Figure 4 is significantly different from \bar{p} in Figure 3 for some settings of K and N . For example, when $K = 32$ and $N = 512$, we have $\bar{p} = 0.81$ in Figure 3 and $\bar{p} = 0.55$ in Figure 4. However, the characteristics of the error of $\bar{p}^{(\infty)}$ and the error of $\bar{q}^{(\infty)}$ in Figure 4 are similar to those in Figure 3. In Figure 4, the error of $\bar{p}^{(\infty)}$ and the error of $\bar{q}^{(\infty)}$ are again within 3% for $N = 32$ and smaller for a larger N . Also, the error of $\bar{p}^{(\infty)}$ and the error of $\bar{q}^{(\infty)}$ are small when K is close to 0 or close to N . Figure 4 suggests that, for a fixed N , the largest error is achieved at a slightly larger K than the K where the error becomes largest in Figure 3.

C Technical Lemma and proofs

Lemma 9 Let $h(m) \equiv \left(\sum_{j=0}^M c_j \exp(-j \chi/m) \right)^m$, where $\sum_{j=0}^M c_j = 1$, $|\chi| < \infty$, and $0 \leq M \leq \infty$. Then $h(m) \rightarrow \exp\left(-\chi \sum_{j=0}^M j c_j\right)$ as $m \rightarrow \infty$.

Proof: It suffices to show that $\ln h(m) \rightarrow -\chi \sum_{j=0}^M j c_j$ as $m \rightarrow \infty$. Changing the variable with $x = \chi/m$, we obtain

$$\lim_{m \rightarrow \infty} \ln h(m) = \chi \lim_{x \downarrow 0} \frac{\ln \left(\sum_{j=0}^M c_j \exp(-j x) \right)}{x}.$$

Since $\ln(\sum_{j=0}^M c_j) = 0$, we use L'Hopital's rule to obtain

$$\begin{aligned}\lim_{m \rightarrow \infty} \ln h(m) &= -\chi \lim_{x \downarrow 0} \frac{\sum_{j=0}^M j c_j \exp(-j x)}{\sum_{j=0}^M c_j \exp(-j x)} \\ &= -\chi \sum_{j=0}^M j c_j,\end{aligned}$$

where the last expression follows from $\sum_{j=0}^M c_j = 1$. \blacksquare

Proof of Lemma 5: We first consider the asymptotic upper bound for the special case where $\lambda_i = c/i^\alpha$. Let $\nu(x) = c x^{-\alpha} \exp(-c t x^{-\alpha})$. Then $\nu(\cdot)$ is increasing in $[0, (c t)^{1/\alpha}]$ and decreasing in $[(c t)^{1/\alpha}, \infty)$, so that $\nu(x) \leq g((c t)^{1/\alpha}) = 1/(e t)$. Let $i_0 = \lceil (c t)^{1/\alpha} \rceil$. Then

$$\begin{aligned}f(t) &\leq \int_0^{i_0} c x^{-\alpha} \exp(-c t x^{-\alpha}) dx + \frac{1}{e t} + \int_{i_0}^{\infty} c x^{-\alpha} \exp(-c t x^{-\alpha}) dx \\ &= \int_0^{\infty} c x^{-\alpha} \exp(-c t x^{-\alpha}) dx + \frac{1}{e t}\end{aligned}$$

Changing the variable with $y = c t/x^\alpha$, we obtain that

$$f(t) \leq \frac{c^{1/\alpha} t^{-1+1/\alpha}}{\alpha} \int_0^{\infty} e^{-y} y^{-1/\alpha} dy + \frac{1}{e t}.$$

Since the integral in the above expression is a gamma function, $\Gamma(1 - 1/\alpha)$, we obtain that

$$f(t) \lesssim \frac{c^{1/\alpha} \Gamma(1 - 1/\alpha) t^{-1+1/\alpha}}{\alpha}. \quad (26)$$

Next, we consider the general case, where $\lambda_i \sim c/i^\alpha$. Then, for any ε , there exists j_0 such that $\forall i > j_0$, we have that $(1 - \varepsilon) c/i^\alpha < \lambda_i < (1 + \varepsilon) c/i^\alpha$. Hence,

$$f(t) \leq \sum_{i=1}^{j_0} \lambda_i \exp(-\lambda_i t) + \sum_{i=j_0+1}^{\infty} (1 + \varepsilon) c/i^\alpha \exp(-(1 - \varepsilon) c t/i^\alpha)$$

Let $\lambda^* \equiv \min(\lambda_1, \dots, \lambda_{j_0})$. Since $\lambda_i \leq 1$ for any i , we obtain

$$f(t) \leq j_0 \exp(-\lambda^* t) + \frac{1 + \varepsilon}{1 - \varepsilon} \sum_{i=1}^{\infty} (1 - \varepsilon) c/i^\alpha \exp(-(1 - \varepsilon) c t/i^\alpha)$$

By Relation (26), we obtain

$$f(t) \lesssim \frac{1 + \varepsilon ((1 - \varepsilon) c)^{1/\alpha} \Gamma(1 - 1/\alpha) t^{-1+1/\alpha}}{1 - \varepsilon \alpha}.$$

By taking $\varepsilon \rightarrow 0$, we obtain

$$f(t) \lesssim \frac{c^{1/\alpha} \Gamma(1 - 1/\alpha) t^{-1+1/\alpha}}{\alpha}.$$

The asymptotic lower bound can be proved using Lemma 3.1 from [21]. Since $e^{-x} \geq 1 - x$ for any x , we have that

$$f(t) \geq \sum_{i=1}^{\infty} \lambda_i (1 - \lambda_i)^t.$$

By Lemma 3.1 from [21], we obtain the asymptotic lower bound. ■