Adv. Appl. Prob. **42,** 816–833 (2010)
Printed in Northern Ireland

A FLUID LIMIT FOR A CACHE ALGORITHM WITH GENERAL REQUEST PROCESSES

TAKAYUKI OSOGAMI,* *IBM Research - Tokyo*

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. We define our fluid limit as a superposition of dependent replications of the original system with smaller item sizes when the number of replications approaches ∞. We derive the average probability that a requested item is not in a cache (average miss probability) in the fluid limit. We show that, when requests follow inhomogeneous Poisson processes, the average miss probability in the fluid limit closely approximates that in the original system. Also, we compare the asymptotic characteristics, as the cache size approaches ∞, of the average miss probability in the fluid limit to those in the original system.

Keywords: Fluid limit; least recently used; point process; nonstationary; hit probability; insensitivity; approximation

2010 Mathematics Subject Classification: Primary 68W40
Secondary 60G55; 60F05

1. Introduction

Caching data is a widely used technique for scalability and efficiency in today's computer and communication systems, including the World Wide Web, sensor networks, and peer-to-peer networks. 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. 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.

The least-recently-used (LRU) algorithm 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 [15] derived 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 were extended to the probability distribution by

Received 16 March 2009; revision received 24 March 2010.
* Postal address: IBM Research - Tokyo, 1623-14 Shimotsuruma, Yamato-shi, Kanagawa 242-8502, Japan.
Email address: osogami@jp.ibm.com

Burville and Kingman [1] and to the generating function by Flajolet *et al.* [6] and Fill and Holst [5]. 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 [4] showed that the generating function of the stationary position of a requested item is simplified in the limiting case where the number of items approaches ∞. Jelenković [9] studied the miss probability for LRU in the limiting case where the cache size, K, approaches ∞. In particular, when the request rates, $\bar{\lambda}_i$ for $i = 1, 2, \ldots$, have a heavy tail (i.e. $\bar{\lambda}_i \sim c/i^\alpha$ for $i = 1, 2, \ldots$ with $c > 0$ and $\alpha > 1$), it was shown that the miss probability for LRU decays following a power law as $K \to \infty$. Jelenković [9] also studied 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 $m \to \infty$. In particular, when the request rates have a light tail (i.e. $\bar{\lambda}_i \sim c \exp(-\xi i^\beta)$ for $i = 1, 2, \ldots$ with $c, \xi, \beta > 0$), it was shown that the miss probability for LRU decays exponentially in the fluid limit. Hirade and Osogami [8] showed that the miss probabilities for LRU and the 2Q cache algorithm [13] 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 [11] discussed the asymptotic optimality of the persistent-access-caching algorithm as $K \to \infty$ 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.* [14] and Rodrigues [18] respectively derived the mean and the variance of the stationary position of a requested item in an MTF list, and Chu and Knott [2] derived an expression for the stationary miss probability for LRU. Coffman and Jelenković [3] derived 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 [10] and Sugimoto and Miyoshi [20] showed that, when the request rates have a heavy tail, the miss probability for LRU is asymptotically insensitive to the type of dependencies in the request process studied in Coffman and Jelenković [3] as $K \to \infty$. Jelenković *et al.* [12] characterized the critical cache sizes where the miss probability for LRU becomes insensitive to the dependencies.

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 nontrivial extension of the fluid limits for the independent reference model in [8] and [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 (Theorem 1). 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 to approximate 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 (Theorem 2), which are nonstationary. All of the

https://doi.org/10.1239/aap/1282924064 Published online by Cambridge University Press

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 $N \geq 128$ 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 $K \to \infty$, $\bar{p}^{(\infty)}$ is asymptotically insensitive to particular dependencies in the request processes when the request rates have a heavy tail (Theorem 3), which agrees with the findings for $\bar{p}$ in [3] and [20]. 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 (Theorem 4). Note that asymptotic characteristics of $\bar{p}$ as $K \to \infty$ is not known even for the independent reference model. Recall that Jelenković [9] studied the asymptotic characteristics for the case of a light tail in his fluid limit.

The rest of the paper is organized as 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 evaluate the accuracy of approximating $\bar{p}$ with $\bar{p}^{(\infty)}$ when requests follow inhomogeneous Poisson processes. 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, $\mathbf{\Psi}$. In Section 2.1 we define a model of caching with LRU and state assumptions on $\mathbf{\Psi}$. 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, $\mathbf{\Psi} = (\Psi_1, \ldots, \Psi_N)$, where $\Psi_i = \{t_\ell^{(i)},\ \ell \in \mathbb{Z}\}$ denotes the request process for the ith 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 $\ell \in \mathbb{Z}$, so that $t_\ell^{(i)}$ denotes the epoch of the ℓth 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 $(\ell, i) \neq (\ell', j)$. In addition, we assume that $t_\ell^{(i)} \to \infty$ and $t_{-\ell}^{(i)} \to -\infty$ as $\ell \to \infty$, so that a finite number of requests are issued in a bounded interval. When these assumptions hold, we say that $\mathbf{\Psi}$ 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, $\mathbf{\Psi}$ may be nonstationary 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 ℓth 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 $\bar{p}_i \equiv \lim_{L\to\infty} 1/L \sum_{\ell=1}^{L} p_{i,\ell}$.

To formally study $\bar{p}_i$, we use notation from [19] and make additional assumptions about $\mathbf{\Psi}$. 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, $\theta_t \Psi_i = \{(t^{(i)}_{M^{(i)}(t)+\ell} - t),\ \ell \in \mathbb{Z}\}$, where

https://doi.org/10.1239/aap/1282924064 Published online by Cambridge University Press

$M^{(i)}(t)$ is the maximum ℓ such that $t_\ell^{(i)} \le t$. Let $\theta_t \mathbf{\Psi} = (\theta_t \Psi_1, \dots, \theta_t \Psi_N)$. We assume that $\mathbf{\Psi}$ is time-asymptotically stationary, so that there exists a distribution defined by $\mathrm{P}^\star(\mathbf{\Psi} \in \mathcal{E}) \equiv \lim_{t\to\infty} 1/t \int_0^t \mathrm{P}(\theta_u \mathbf{\Psi} \in \mathcal{E})\,\mathrm{d}u$. Note that a nonstationary $\mathbf{\Psi}$ can be time-asymptotically stationary. For simplicity, we assume that $\mathbf{\Psi}$ is ergodic with respect to $\mathrm{P}^\star$.

Finally, we assume that the average request rate, $\bar\lambda_i$, of e_i satisfies $0 < \bar\lambda_i < \infty$ for $i = 1, \dots, N$. Formally, $\bar\lambda_i \equiv \mathrm{E}^\star[M^{(i)}(1)]$, where $M^{(i)}(1)$ denotes the number of requests for e_i in $(0, 1]$, and $\mathrm{E}^\star$ denotes the expectation with respect to $\mathrm{P}^\star$. When $N = \infty$, we also assume that $\sum_{i=1}^N \bar\lambda_i < \infty$.

2.2. Average miss probability

Under the above assumptions, Ψ_i is event-asymptotically stationary, so that the distribution defined by $\mathrm{P}^{0,i}(\Psi_i \in \mathcal{E}) \equiv \lim_{L\to\infty} 1/L \sum_{\ell=1}^L \mathrm{P}(\theta_{t_\ell^{(i)}} \Psi_i \in \mathcal{E})$ exists for $1 \le i \le N$ (see Theorem 2.9 of [19]). Then $\bar p_i$ can be expressed conveniently using $\mathrm{P}^{0,i}$.

Lemma 1. *When $\mathbf{\Psi}$ is simple, time-asymptotically stationary, ergodic, and $\sum_{i=1}^N \bar\lambda_i < \infty$, the average miss probability of e_i for LRU is $\bar p_i = \mathrm{P}^{0,i}(\sum_{j\ne i} I\{t_1^{(j)} < t_1^{(i)}\} \ge K)$ for $1 \le i \le N$, where $I\{\cdot\}$ is the indicator random variable.*

We provide a formal proof in Appendix A, but Lemma 1 can be explained intuitively as follows. We may see $\mathrm{P}^{0,i}(\mathcal{E})$ as the probability of an event, $\mathcal{E}$, when we 'randomly observe way out at' [19] the epoch of a request for e_i, letting the time of the observation be 0. The next request for e_i after the observation is at time $t_1^{(i)}$ and is a miss if and only if 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 $(0, t_1^{(i)})$ if and only if $t_1^{(j)} < t_1^{(i)}$ for any $e_j \ne e_i$. Hence, the request for e_i at time $t_1^{(i)}$ is a miss if and only if $\sum_{j\ne i} I\{t_1^{(k)} < t_1^{(i)}\} \ge K$.

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 the fluid limit

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

Such scaled systems are also considered in [8] and [9]. For example, the mth scaled system, $\mathcal{S}^{(m)}$, of [8] can be seen as a superposition of independent replications of the original system. Specifically, in $\mathcal{S}^{(m)}$, the $\mathbf{\Phi}_k$ for $1 \le k \le m$ are independent and stochastically identical to $\mathbf{\Psi}$. Unfortunately, the dependencies in $\mathbf{\Psi}$ would disappear in $\mathcal{S}^{(\infty)}$ in the sense that $\mathcal{S}^{(\infty)}$ with general $\mathbf{\Psi}$ is identical to that when $\Psi_1, \dots, \Psi_N$ are independent. We formally prove the above observation in Appendix B.

We will define our scaled system as a superposition of *dependent* replications of the original system. Also, in contrast to [8] and [9], we will define a sequence of scaled systems *for each* e_i, so that the scaled systems for different items have different dependencies in $\mathbf{\Phi}_k$. Let $\mathcal{T}_i^{(m)}$ be the mth 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 $\mathbf{\Phi}_k$ is stochastically identical to $\mathbf{\Psi}$ (i.e. for $1 \le k \le m$, it holds

https://doi.org/10.1239/aap/1282924064 Published online by Cambridge University Press

that $\mathrm{P}(\mathbf{\Phi}_k \in \mathcal{E}) = \mathrm{P}(\mathbf{\Psi} \in \mathcal{E})$ for any measurable set, $\mathcal{E}$). However, we assume that the $\mathbf{\Phi}_k$ for $1 \le k \le m$ depend on each other. Specifically, in $\mathcal{T}_i^{(m)}$, we assume that $\Phi_{i,k}$ for $1 \le k \le m$ have the same sample path (i.e. $t_\ell^{(i,k)} = t_\ell^{(i,k')}$ for any $\ell \in \mathbb{Z}$ and $1 \le k, k' \le m$) and that the $\mathbf{\Phi}_k$ for $1 \le k \le m$ are conditionally independent given $\Phi_{i,1}$. Formally, for any measurable sets, $\mathcal{E}_k$ for $1 \le k \le m$, it holds that

$$\mathrm{P}(\mathbf{\Psi}_k \in \mathcal{E}_k \text{ for all } k \in \{1, \dots, m\} \mid \Phi_{i,1}) = \prod_{k=1}^{m} \mathrm{P}(\mathbf{\Psi}_k \in \mathcal{E}_k \mid \Phi_{i,1}). \tag{1}$$

To clarify the assumptions on $\mathbf{\Phi}$, consider a way to simulate $\mathbf{\Phi}^{(m)} \equiv (\mathbf{\Phi}_1, \dots, \mathbf{\Phi}_m)$ 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, $\Psi_i(\omega) = \{t_1^{(i)}(\omega), \dots, t_{L_i(\omega)}^{(i)}(\omega)\}$, where ω denotes a sample path and $L_i(\omega)$ denotes the number of the requests in $(0, T]$. Then, for $1 \le k \le m$, we let $\Phi_{i,k}(\omega) = \Psi_i(\omega)$ be the simulated epochs of the requests for $e_{i,k}$ in $\mathcal{T}_i^{(m)}$ (i.e. $t_\ell^{(i,k)}(\omega) = t_\ell^{(i)}(\omega)$ for $1 \le \ell \le L_i(\omega)$). Next we simulate $\mathbf{\Psi}_{-i} \equiv \{\Psi_j \mid j \ne i\}$ in the original system in such a way that $\Psi_i(\omega)$ and $\mathbf{\Psi}_{-i}$ have the desired dependency. The sample path, ω_1, from the simulation of $\mathbf{\Psi}_{-i}$ is used to construct the simulated epochs of the requests for $\{e_{j,1} \mid j \ne i\}$ in $\mathcal{T}_i^{(m)}$ such that $\Phi_{j,1}(\omega_1) = \Psi_j(\omega_1) = \{t_1^{(j)}(\omega_1), \dots, t_{L_j(\omega_1)}^{(j)}(\omega_1)\}$ for each $j \ne i$. We repeat simulating $\mathbf{\Psi}_{-i}$ in the same way, but independently of the previous repetitions. For $2 \le k \le m$, the sample path, ω_k, from the kth repetition is used in the same way as ω_1 to construct the simulated epochs of the requests for $\{e_{j,k} \mid j \ne i\}$ in $\mathcal{T}_i^{(m)}$.

To avoid introducing a tie-breaking rule, we assume that, in $\mathcal{T}_i^{(m)}$, the items in $\{e_{j,k} \mid j \ne i,\ 1 \le k \le m\} \cup \{e_{i,1}\}$ are requested one at a time almost surely (recall that the items in $\{e_{i,k} \mid 1 \le k \le m\}$ are requested simultaneously). This means, in the original system, that there is no mass probability: $\mathrm{P}(t_\ell^{(i)} = t) = 0$ 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 $\mathcal{T}_i^{(m)}$ if and only if more than half of $e_{i,k}$ for $1 \le k \le m$ are not in the cache upon the request. Let $p_{i,\ell}^{(m)}$ be the probability that the ℓth request for e_i is a miss in $\mathcal{T}_i^{(m)}$. We study the average miss probability, $\bar{p}_i^{(m)} \equiv \lim_{L\to\infty} 1/L \sum_{\ell=1}^{L} p_{i,\ell}^{(m)}$, of e_i as $m \to \infty$. Note that $\bar{p}_i^{(\infty)}$ and $\bar{p}_j^{(\infty)}$ are defined with different fluid limits, $\mathcal{T}_i^{(\infty)}$ and $\mathcal{T}_j^{(\infty)}$, respectively, for $i \ne j$. Recall how the dependencies in $\{\Phi_{i,k} \mid 1 \le i \le N,\ k = 1, 2, \dots\}$ are constructed differently between $\mathcal{T}_i^{(\infty)}$ and $\mathcal{T}_j^{(\infty)}$ for $i \ne j$.

Theorem 1. *In addition to the conditions of Lemma 1, suppose that* $\mathrm{P}(t_\ell^{(i)} = t) = 0$ *for any* ℓ, t, *and* i. *Then* $\lim_{m\to\infty} \bar{p}_i^{(m)} = \mathrm{P}^{0,i}(\sum_{j=1}^{N} \mathrm{E}[I\{t_1^{(j)} < t_1^{(i)}\} \mid \Psi_i] > K - \frac{1}{2})$.

The theorem 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, $\mathrm{E}[I\{t_1^{(j)} < t_1^{(i)}\} \mid \Psi_i]$, 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 $\mathbf{\Psi}_{-i}$ via the dependencies between Ψ_i and $\mathbf{\Psi}_{-i}$.

Proof of Theorem 1. Let $C_{i,\ell}^{(m)}$ be the total size of distinct items that are requested after $t_{\ell-1}^{(i,1)}$ and before $t_\ell^{(i,1)}$ in $\mathcal{T}^{(m)}$. Note that $p_{i,\ell}^{(m)} = \mathrm{P}(C_{i,\ell}^{(m)} > K - \frac{1}{2})$. We will first show that, as $m \to \infty$,

$$C_{i,\ell}^{(m)} \xrightarrow{\mathrm{D}} \sum_{j=1}^{N} \mathrm{E}[I_\ell(j) \mid \Psi_i], \tag{2}$$

where '$\xrightarrow{\mathrm{D}}$' denotes convergence in distribution and $I_\ell(j)$ is the indicator random variable such

https://doi.org/10.1239/aap/1282924064 Published online by Cambridge University Press

that $I_\ell(j) = 1$ if and only if e_j is requested in the interval $(t_{\ell-1}^{(i)}, t_\ell^{(i)})$ in the original system for $1 \le j \le N$. Note that $I_\ell(i) = 0$.

We prove the convergence in distribution by showing the convergence of the Laplace transform, $\varphi_{i,\ell}^{(m)}(s) \equiv \mathrm{E}[\exp(-sC_{i,\ell}^{(m)})]$ for $0 \le s < \infty$, of $C_{i,\ell}^{(m)}$. Let $I_\ell(j,k)$ be the indicator random variable such that $I_\ell(j,k) = 1$ if and only if $e_{j,k}$ is requested after $t_{\ell-1}^{(i,1)}$ and before $t_\ell^{(i,1)}$. By definition, $I_\ell(i,k) = 0$ for $1 \le k \le m$. Then

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

The conditional independence assumed in (1) implies that

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

Also, since the $\mathbf{\Phi}_k$ for $1 \le k \le m$ are stochastically identical, we obtain

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

For $0 \le n \le N-1$, let $Q_n = \mathrm{P}(\sum_{j=1}^{N} I_\ell(j,1) = n \mid \Phi_{i,1})$ be the conditional probability that n distinct items in E_1 are requested in the interval $(t_{\ell-1}^{(i,1)}, t_\ell^{(i,1)})$ given $\Phi_{i,1}$. Then

$$\varphi_{i,\ell}^{(m)}(s) = \mathrm{E}\left[\left(\sum_{n=0}^{N-1} Q_n \exp\left(-\frac{sn}{m}\right)\right)^m\right].$$

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

$$\lim_{m\to\infty}\varphi_{i,\ell}^{(m)}(s) = \mathrm{E}\left[\lim_{m\to\infty}\left(\sum_{n=0}^{N-1} Q_n \exp\left(-\frac{sn}{m}\right)\right)^m\right].$$

By Lemma 8 in Appendix A we obtain

$$\lim_{m\to\infty}\varphi_{i,\ell}^{(m)}(s) = \mathrm{E}\left[\exp\left(-s\sum_{n=0}^{N-1} nQ_n\right)\right].$$

Since $\sum_{n=0}^{N-1} nQ_n = \mathrm{E}[\sum_{j=1}^{N} I_\ell(j,1) \mid \Phi_{i,1}]$, we obtain

$$\lim_{m\to\infty}\varphi_{i,\ell}^{(m)}(s) = \mathrm{E}\left[\exp\left(-s\,\mathrm{E}\left[\sum_{j=1}^{N} I_\ell(j,1)\,\middle|\,\Phi_{i,1}\right]\right)\right]. \tag{3}$$

Therefore, the continuity theorem (see, e.g. p. 262 of [7]) implies that

$$C_{i,\ell}^{(m)} \xrightarrow{\mathrm{D}} \mathrm{E}\left[\sum_{j=1}^{N} I_\ell(j,1)\,\middle|\,\Phi_{i,1}\right] \quad \text{as } m \to \infty.$$

https://doi.org/10.1239/aap/1282924064 Published online by Cambridge University Press

Since the pair $(I_\ell(j,1), \Phi_{i,1})$ and the pair $(I_\ell(j), \Psi_i)$ are stochastically identical, this establishes (2) via the linearity of expectation.

Now, (2) suggests that, as $m \to \infty$,

$$p_{i,\ell}^{(m)} \to \mathrm{P}\left(\sum_{j=1}^{N} \mathrm{E}[I_\ell(j) \mid \Psi_i] > K - \frac{1}{2}\right) = \mathrm{P}(\theta_{\ell-1}^{(i)} \boldsymbol{\Psi} \in \mathcal{D}_i), \tag{4}$$

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

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

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

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

$$\lim_{m\to\infty} \bar{p}_i^{(m)} = \lim_{L\to\infty} \frac{1}{L-1} \sum_{\ell=2}^{L} \lim_{m\to\infty} p_{i,\ell}^{(m)}, \tag{5}$$

which together with (4) 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 follow 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 study the accuracy of approximating $\bar{p}_i$ with $\bar{p}_i^{(\infty)}$.

4.1. Miss probability in the fluid limit

The expression of $\bar{p}_i^{(\infty)}$ in Theorem 1 can be made more explicit when a specific $\boldsymbol{\Psi}$ is assumed, which we will demonstrate for the case where $\boldsymbol{\Psi}$ is a vector of independent inhomogeneous Poisson processes. We will also consider a special case where $\boldsymbol{\Psi}$ is a vector of independent Poisson processes and compare our results against those in [8] and [9].

Theorem 2. *In addition to the conditions of Lemma 1, suppose that Ψ_i is an inhomogeneous Poisson process with rate $\lambda_i(t)$ at time t for $1 \le i \le N$. Let $\Lambda_i(t,u) \equiv \int_t^u \lambda_i(v)\,\mathrm{d}v$, and let $\tau_i(t)$ be the maximum u such that $\sum_{j\ne i}(1 - \exp(-\Lambda_j(t,u))) \le K - \frac{1}{2}$ for $1 \le i \le N$. Then*

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

where $\bar{\lambda}_i = \lim_{T\to\infty} 1/T \int_0^T \lambda_i(t)\,\mathrm{d}t$.

Proof. We first consider $p_{i,\ell}^{(\infty)}$. By (4) and the independence of Ψ_i and $\boldsymbol{\Psi}_{-i}$, we obtain

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

as $m \to \infty$ for $\ell > 1$. Since $\mathrm{E}[I_\ell(j) \mid (t_{\ell-1}^{(i)}, t_\ell^{(i)})]$ is the conditional probability that e_j is

https://doi.org/10.1239/aap/1282924064 Published online by Cambridge University Press

requested in the interval, $(t_{\ell-1}^{(i)}, t_\ell^{(i)})$, we find, by a property of the inhomogeneous Poisson process (see, e.g. p. 246 of [16]), that $\mathrm{E}[I_\ell(j) \mid (t_{\ell-1}^{(i)}, t_\ell^{(i)})] = 1 - \exp(-\Lambda_j(t_{\ell-1}^{(i)}, t_\ell^{(i)}))$.

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 $T_\ell^{(i)}(t_{\ell-1}^{(i)})$ be the epoch of the ℓth request for e_i given $t_{\ell-1}^{(i)}$. By the Markovian property, given $t_{\ell-1}^{(i)}$, $T_\ell^{(i)}(t)$ is conditionally independent of ℓ, so that we write $T^{(i)}(t) \equiv T_\ell^{(i)}(t)$, 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

$$p_{i,\ell}^{(\infty)} = \int_0^\infty P_i(t)\lambda_{i,\ell-1}(t)\,\mathrm{d}t, \tag{6}$$

where $P_i(t) \equiv \mathrm{P}(\sum_{j\neq i}(1 - \exp(-\Lambda_j(t, T^{(i)}(t)))) > K - \frac{1}{2})$. Since $1 - \exp(-\Lambda_j(t,u))$ is nondecreasing with u for any t, we have $P_i(t) = \mathrm{P}(T^{(i)}(t) > \tau_i(t)) = \exp(-\Lambda_i(t, \tau_i(t)))$, where the last equality follows from the property of the inhomogeneous Poisson process.

Finally, we derive $\bar{p}_i^{(\infty)}$. By (5) and (6), we obtain

$$\bar{p}_i^{(\infty)} = \lim_{L\to\infty} \frac{1}{L-1} \sum_{\ell=2}^{L} \int_0^\infty P_i(t)\lambda_{i,\ell-1}(t)\,\mathrm{d}t. \tag{7}$$

We will use $H = \lambda G$, an extension of Little's law, to show that the event-average expression with (7) is equivalent to the time-average expression

$$\bar{p}_i^{(\infty)} = \frac{1}{\bar{\lambda}_i} \lim_{T\to\infty} \frac{1}{T} \int_0^T P_i(t)\lambda_i(t)\,\mathrm{d}t. \tag{8}$$

Let $G_\ell \equiv \int_0^\infty P_i(t)\lambda_{i,\ell}(t)\,\mathrm{d}t$. 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 $t_0^{(i)} \le 0 < t_1^{(i)}$, there is a relationship between $\lambda_{i,\ell}(u)$ and $\lambda_i(u)$ for $u \ge 0$ 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 $\boldsymbol{\Psi}$ is time-asymptotically stationary and ergodic, $H = \bar{\lambda}_i G$ holds (see Theorem 6.4 of [19]) for $G \equiv \lim_{L\to\infty} 1/L \sum_{\ell=1}^L G_\ell$ and $H \equiv \lim_{T\to\infty} 1/T \int_0^T H(t)\,\mathrm{d}t$. Thus, we can conclude that (8) is valid, which completes the proof of the theorem.

To gain further insights into our fluid limit, we consider the case where $\lambda_i(\cdot)$ is a constant (i.e. Ψ_i is a Poisson process) for each i. The following corollary can be compared against the stationary miss probabilities in the fluid limits obtained in [8] and [9].

Corollary 1. *If Ψ_i is an independent Poisson process with rate $\bar{\lambda}_i$ for each e_i, then $\bar{p}_i^{(m)} \to \exp(-\bar{\lambda}_i\tau_i(K))$ as $m \to \infty$, where $\tau_i(K) = C_i^{-1}(K - \frac{1}{2})$ and $C_i^{-1}(\cdot)$ is the inverse function of $C_i(t) \equiv \sum_{j\neq i}(1 - \exp(-\bar{\lambda}_j t))$.*

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. Note that the term $1 - \exp(-\bar{\lambda}_j t)$ 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 $1 \le k \le m$ that are requested in $(0,t)$ as $m \to \infty$. In $\mathcal{T}_i^{(\infty)}$, the position of $e_{i,1}$ reaches $K - \frac{1}{2}$ at $t = \tau_i(K)$. The probability that the next request for $e_{i,1}$ is issued after $t = \tau_i(K)$ is $\exp(-\bar{\lambda}_i\tau_i(K))$. 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.

https://doi.org/10.1239/aap/1282924064 Published online by Cambridge University Press

Since our fluid limit differs from the fluid limits defined in [8] and [9], our $\bar{p}_i^{(\infty)}$ differs from those derived in [8] and [9]. However, the only difference between our $\bar{p}_i^{(\infty)}$ and that in [8] is that, in [8], $\tau_i(K)$ is replaced with $\tau(K) = C^{-1}(K)$, where $C^{-1}(\cdot)$ is the inverse function of $C(t) = \sum_{j=1}^{N}(1 - \exp(-\bar{\lambda}_j t))$. The differences between the fluid limits in [8] and [9] are discussed in [8]. We find that these differences are negligible for practical purposes.

4.2. Accuracy of approximation with fluid limit

Now we study the accuracy of approximating $\bar{p}_i$ with $\bar{p}_i^{(\infty)}$. Let $r_i \equiv \bar{\lambda}_i / \sum_{j=1}^{N} \bar{\lambda}_j$ 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 $\bar{p}^{(\infty)} \equiv \sum_{i=1}^{N} r_i \bar{p}_i^{(\infty)}$ evaluated numerically. Recall that $\bar{p}_i^{(\infty)}$ is defined 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 $100|\bar{p}^{(\infty)} - \bar{p}|/\bar{p}$.

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, additional runs are executed 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.

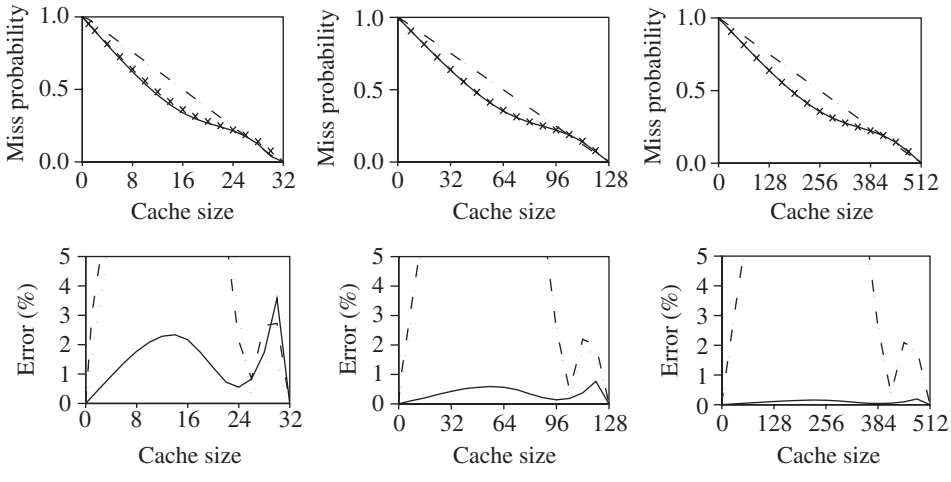
We consider the settings where the value of $\lambda_i(\cdot)$ fluctuates as a trigonometric function, $\lambda_i(t) = 2\sin^2(\pi t/\sigma + \pi i/\nu)$, for each e_i. Observe that, for any e_i, the period of $\lambda_i(\cdot)$ is σ and its average rate is $\bar{\lambda}_i = 1$, so that e_i is expected to be requested σ times in a period. The phase of $\lambda_i(0)$ is chosen depending on $(i \bmod \nu)$. Therefore, items are classified into ν types, and items with different types become popular (requested frequently) in different epochs.

In Figure 1, we set $\sigma = 4$ and $\nu = 8$ for each e_i. In the top row of Figure 1 the solid lines show $\bar{p}^{(\infty)}$ and the crosses show $\bar{p}$. The number of items, N, is set as shown in each column. The horizontal axis represents the cache size, K. Although we have defined $\bar{p}$ and $\bar{p}^{(\infty)}$ only for $1 \le K \le N-1$, Figure 1 shows the range of $0 \le K \le N$. 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 crosses are on top of each other when $N \ge 128$. We can see that $\bar{p}^{(\infty)}$ slightly underestimates $\bar{p}$ for $N = 32$.

This high accuracy of approximation would not be obtained if the dependency of $\lambda_i(t)$ on t was ignored (i.e. each Ψ_i was replaced with a Poisson process having the same $\bar{\lambda}_i$ as Ψ_i). Recall that, in the settings of Figure 1, we have $\bar{\lambda}_i = 1$ for any i. Now, suppose that each item is independently requested with a Poisson process with a common rate. Then, since each item is equally likely to be requested at each moment, the overall miss probability would be $1 - K/N$, which is shown with dotted lines in the top row of Figure 1. In this particular case, the overall miss probability in the fluid limit of [8] would also agree with $1 - K/N$. By Corollary 1, the overall miss probability in our fluid limit would be $1 - (K - \frac{1}{2})/(N-1)$, which is shown with dashed lines in the top row of Figure 1. Observe that the dotted lines and the dashed lines are on top of each other and can be tremendously deviated from the solid line particularly when $K \approx N/2$.

The bottom row of Figure 1 shows the error (%) of $\bar{p}^{(\infty)}$ with solid lines. Observe that the error of $\bar{p}^{(\infty)}$ is within 5% for $N = 32$ and within 1% for $N \ge 128$. 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 as $N \to \infty$. We find that $\bar{p}^{(\infty)} \le \bar{p}$ for all of the data points in Figure 1. Also, observe that the error of $\bar{p}^{(\infty)}$ is relatively large at $K \approx N/2$ and $K \approx N$. The dashed lines and the dotted lines show the analogous errors when the dependency of the

https://doi.org/10.1239/aap/1282924064 Published online by Cambridge University Press



(a) $N = 32$ (b) $N = 128$ (c) $N = 512$

FIGURE 1: The accuracy of approximating $\bar{p}$ with $\bar{p}^{(\infty)}$ when requests follow inhomogeneous Poisson processes with $\lambda_i(t) = 2\sin^2(\pi t/4 + \pi i/8)$ for each e_i, where N is set as shown in each column. In the top row, the solid lines show $\bar{p}^{(\infty)}$ and the crosses show $\bar{p}$. The bottom row shows the error (%) of $\bar{p}^{(\infty)}$.

request rates on time are ignored. These errors often exceed 5% (sometimes 14%) and are not shown in the range of the figure.

We find that most of the qualitative findings from Figure 1 hold for other settings of $\lambda_i(\cdot)$. In general, as σ becomes larger, $\bar{p}$ becomes smaller, but the error of $\bar{p}^{(\infty)}$ is relatively insensitive to σ. Also, we find that, for a large σ, the error of $\bar{p}^{(\infty)}$ has a single peak at $K \approx N/2$. We find that $\bar{p}$ is less sensitive to ν than to σ, and the sensitivity to ν is hard to characterize. Overall, we find that the error of $\bar{p}^{(\infty)}$ is within 5% for all cases studied with $N \geq 32$.

5. Large cache asymptotics with fluid limit

In this section we study the request processes that are similar to those studied in [3], [10], [12], and [20]. For $1 \leq i \leq N$, let $J(\cdot)$ be a stationary and ergodic semi-Markov chain on a finite state space that determines the request rate for e_i at time t with $\lambda_i(J(t))$. Thus, given $J(\cdot)$, Ψ_i is an inhomogeneous Poisson process with rate $\lambda_i(J(t))$ at time t. Observe that the Ψ_i for $i = 1, \ldots, N$ are conditionally independent given $J(\cdot)$. Note that $\boldsymbol{\Psi}$ is stationary, which is also assumed in [3], [10], [12], and [20], so that the stationary miss probability exists (see Lemma 2.1 of [20]) and agrees with the average miss probability. In Section 5.1 we state the results of our asymptotic analysis. In Section 5.2 we provide proofs of the results.

5.1. Results

We study asymptotic characteristics of 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 of size K as $K \to \infty$. We assume that $N = \infty$ and that $\sum_{j=1}^{N} \bar{\lambda}_j = 1$ (without loss of generality), so that $r_i = \bar{\lambda}_i$. Once again, recall that $\bar{p}^{(\infty)}(K)$ is a formal weighted average of $\bar{p}_i^{(\infty)}$ for $i = 1, 2, \ldots$, where $\bar{p}_i^{(\infty)}$ and $\bar{p}_j^{(\infty)}$ are defined with different fluid limits for $i \neq j$.

https://doi.org/10.1239/aap/1282924064 Published online by Cambridge University Press

We first derive $\bar{p}_i^{(\infty)}$ for the particular request processes under consideration with no assumptions on $\bar{\lambda}_1, \bar{\lambda}_2, \ldots$.

Lemma 2. *Let* $\Lambda_i(u; J) \equiv \int_0^u \lambda_i(J(v))\,\mathrm{d}v$, *and let* $\tau_i(K; J)$ *be the maximum* u *such that* $\sum_{j\neq i}(1-\exp(-\Lambda_j(u; J))) \leq K - \frac{1}{2}$. *In addition to the conditions in Lemma 1, suppose 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 stationary and ergodic semi-Markov chain on a finite state space. Then* $\lim_{m\to\infty} \bar{p}_i^{(m)} = \mathrm{E}[\exp(-\Lambda_i(\tau_i(K; J); J))\lambda_i(J(0))]/\bar{\lambda}_i$.

Now, we consider the case when the distribution of $\bar{\lambda}_i$ over i has a heavy tail. Below, $a \sim_x b$ denotes $\lim_{x\to\infty} a/b = 1$, and $a \lesssim_x b$ denotes $\lim_{x\to\infty} a/b \leq 1$. Also, $\Gamma(z) \equiv \int_0^\infty \mathrm{e}^{-y} y^{z-1}\,\mathrm{d}y$ denotes the gamma function. We find that $\bar{p}^{(\infty)}(K)$ decays with a power law as $K \to \infty$ and is asymptotically insensitive to $J(\cdot)$.

Theorem 3. *In addition to the conditions in Lemma 2, suppose that* $\bar{\lambda}_i \sim_i c/i^\alpha$ *for* $i = 1, 2, \ldots$, *where* $\alpha > 1$ *and* $c > 0$. *Then* $\bar{p}^{(\infty)}(K)$ *is asymptotically insensitive to* $J(\cdot)$ *as* $K \to \infty$, *and it holds that* $\bar{p}^{(\infty)}(K) \sim_K c\alpha^{-1}\Gamma(1-1/\alpha)^\alpha K^{1-\alpha}$.

Theorem 3, which is obtained for the fluid limit, is in agreement with the asymptotic results for the original system derived in [10] and [20]. However, an asymptotic analysis of $\bar{p}^{(\infty)}$, such as Theorem 3, appears to be simpler than the corresponding asymptotic analysis of $\bar{p}$.

Next, we consider the case when the distribution of $\bar{\lambda}_i$ has a light tail. This case has not been fully investigated in the prior work. Jelenković [9] studied asymptotic properties of the overall stationary miss probability in his fluid limit when $\bar{\lambda}_i$ has the 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 find that $\bar{p}^{(\infty)}(K)$ decays exponentially as $K \to \infty$ and is asymptotically insensitive to $J(\cdot)$.

Theorem 4. *In addition to the conditions in Lemma 2, suppose that* $\bar{\lambda}_i \sim_i c\exp(-\xi i^\beta)$ *for* $i = 1, 2, \ldots$, *where* $c, \xi, \beta > 0$. *Then* $\bar{p}^{(\infty)}(K)$ *is asymptotically insensitive to* $J(\cdot)$ *as* $K \to \infty$, *and it holds that*

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

where $\gamma \equiv \int_0^\infty \exp(-y)\ln y\,\mathrm{d}y \approx 0.577$ *is Euler's constant.*

5.2. Proofs

We first prove Lemma 2, which will be used to prove Theorem 3 and Theorem 4.

Proof of Lemma 2. Let $\bar{p}_i^{(m)}(J)$ be the conditional average miss probability for e_i in $\mathcal{T}_i^{(m)}$ given J. Then $\bar{p}_i^{(m)} = \mathrm{E}[\bar{p}_i^{(m)}(J)]$. Since $0 \leq \bar{p}_i^{(m)}(J) \leq 1$, the dominated convergence theorem can be used to show that $\lim_{m\to\infty} \bar{p}_i^{(m)} = \mathrm{E}[\lim_{m\to\infty} \bar{p}_i^{(m)}(J)]$. By Theorem 2 we obtain

$$\bar{p}_i^{(m)} \to \frac{1}{\bar{\lambda}_i}\mathrm{E}\left[\lim_{T\to\infty}\frac{1}{T}\int_0^T \exp(-\Lambda_i(t, \tau_i(t, K; J); J))\lambda_i(J(t))\,\mathrm{d}t\right]$$

as $m \to \infty$, where we define $\Lambda_i(t, u; J) \equiv \int_t^u \lambda_i(J(v))\,\mathrm{d}v$, and $\tau_i(t, K; J)$ is the maximum u such that $\sum_{j\neq i}(1-\exp(-\Lambda_j(t, u; J))) \leq K - \frac{1}{2}$. Because $\Lambda_i(t, \tau_i(t, K; J); J) \geq 0$ for $0 \leq t$, we have $0 \leq \exp(-\Lambda_i(t, \tau_i(t, K; J); J)) \leq 1$ for $0 \leq t$. Also, because $J(\cdot)$ is defined on a finite state space, there exists $\lambda^{(\max)} < \infty$ such that $0 \leq \lambda_i(J(t)) \leq \lambda^{(\max)}$. Therefore, the

https://doi.org/10.1239/aap/1282924064 Published online by Cambridge University Press

dominated convergence theorem can be used to show that

$$\lim_{m\to\infty} \bar{p}_i^{(m)} = \frac{1}{\bar{\lambda}_i} \lim_{T\to\infty} \frac{1}{T} \int_0^T \mathrm{E}[\exp(-\Lambda_i(t, \tau_i(t, K; J); J))\lambda_i(J(t))]\,\mathrm{d}t.$$

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

$$\begin{aligned}\lim_{m\to\infty} \bar{p}_i^{(m)} &= \frac{1}{\bar{\lambda}_i} \lim_{T\to\infty} \frac{1}{T} \int_0^T \mathrm{E}[\exp(-\Lambda_i(0, \tau_i(0, K; J); J))\lambda_i(J(0))]\,\mathrm{d}t \\ &= \frac{1}{\bar{\lambda}_i} \mathrm{E}[\exp(-\Lambda_i(0, \tau_i(0, K; J); J))\lambda_i(J(0))],\end{aligned}$$

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

In our proof of Theorem 3, we will use the following three lemmas.

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

$$\left|\frac{1}{t}\int_0^t \lambda_i(J(u))\,\mathrm{d}u - \bar{\lambda}_i\right| \Big/ \bar{\lambda}_i \to 0$$

almost surely (a.s.) as $t \to \infty$ uniformly for $i = 1, 2, \ldots$, where

$$\bar{\lambda}_i \equiv \lim_{T\to\infty} \frac{1}{T} \int_0^T \lambda_i(t)\,\mathrm{d}t.$$

Lemma 4. *Let $g_i(t) = \sum_{j\neq i}(1 - \exp(-\bar{\lambda}_j t))$. If $\bar{\lambda}_i \sim_i c/i^\alpha$ with $\alpha > 1$ and $c > 0$, then $g_i(t) \sim_t c^{1/\alpha}\Gamma(1 - 1/\alpha)t^{1/\alpha}$.*

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

Lemma 4 is a direct consequence of Corollary 1 of [9]. Note that $g_i(t)$ is asymptotically insensitive to i in Lemma 4, because $1 - \exp(-\bar{\lambda}_i t) \to 0$ as $t \to \infty$ for any i. Lemma 5 can be proved similar to Lemma 3.1 of [20]. We provide proofs of Lemma 3 and Lemma 5 in Appendix A.

Proof of Theorem 3. We first study $C_i(t; J) \equiv \sum_{j\neq i}(1 - \exp(-\Lambda_j(t; J)))$ as $t \to \infty$. Lemma 3 implies that, for any ε, there exists t_0 such that, for all $t > t_0$, we have

$$\sum_{j\neq i}(1 - \exp(-(1-\varepsilon)\bar{\lambda}_j t)) \le C_i(t; J) \le \sum_{j\neq i}(1 - \exp(-(1+\varepsilon)\bar{\lambda}_j t)) \tag{9}$$

a.s. for any e_i. Hence, Lemma 4 suggests that

$$(1-\varepsilon)^{1/\alpha}c^{1/\alpha}\Gamma\left(1 - \frac{1}{\alpha}\right)t^{1/\alpha} \lesssim_t C_i(t; J) \lesssim_t (1+\varepsilon)^{1/\alpha}c^{1/\alpha}\Gamma\left(1 - \frac{1}{\alpha}\right)t^{1/\alpha}$$

a.s. uniformly for $i = 1, 2, \ldots$. Taking $\varepsilon \to 0$, we obtain

$$C_i(t; J) \sim_t c^{1/\alpha}\Gamma\left(1 - \frac{1}{\alpha}\right)t^{1/\alpha} \tag{10}$$

a.s. uniformly for $i = 1, 2, \ldots$.

https://doi.org/10.1239/aap/1282924064 Published online by Cambridge University Press

Recall that $\tau_i(K; J)$ is the maximum t such that $C_i(t; J) \leq K - \frac{1}{2}$. Because $C_i(t; J)$ is continuous with respect to t, we have $K \sim_K C_i(\tau_i(K; J); J)$. Also, because $\tau_i(K; J) \to \infty$ a.s. as $K \to \infty$, (10) implies that $C_i(\tau_i(K); J) \sim_K c^{1/\alpha}\Gamma(1 - 1/\alpha)(\tau_i(K; J))^{1/\alpha}$ a.s. Thus, we have

$$\tau_i(K; J) \sim_K \tau(K) \equiv \frac{K^\alpha}{c\Gamma(1 - 1/\alpha)^\alpha} \tag{11}$$

a.s. uniformly for $i = 1, 2, \ldots$.

Finally, we consider $\bar{p}^{(\infty)}$. The uniform convergence of (11) and Lemma 3 imply that, for any ε, there exists K_0 such that, for all $K > K_0$, we have

$$(1 - \varepsilon)\bar{\lambda}_i\tau(K) \leq \Lambda_j(\tau_i(K; J); J) \leq (1 + \varepsilon)\bar{\lambda}_i\tau(K) \tag{12}$$

a.s. for $i = 1, 2, \ldots$. Now, Lemma 2 and inequality (12) imply that, for all $K > K_0$, we have

$$\begin{aligned}
\bar{p}^{(\infty)}(K) &\leq \sum_{i=1}^{\infty} \mathrm{E}[\exp(-(1 - \varepsilon)\bar{\lambda}_i\tau(K))\lambda_i(J(0))] \\
&= \sum_{i=1}^{\infty} \mathrm{E}[\lambda_i(J(0))] \exp(-(1 - \varepsilon)\bar{\lambda}_i\tau(K)) \\
&= \frac{1}{1 - \varepsilon} \sum_{i=1}^{\infty} (1 - \varepsilon)\bar{\lambda}_i \exp(-(1 - \varepsilon)\bar{\lambda}_i\tau(K)),
\end{aligned} \tag{13}$$

where the last equality (specifically, $\bar{\lambda}_i = \mathrm{E}[\lambda_i(J(0))]$) follows from the stationality of $J(\cdot)$. By Lemma 5 we obtain $\bar{p}^{(\infty)}(K) \lesssim_K (1 - \varepsilon)^{1/\alpha - 1}c^{1/\alpha}\alpha^{-1}\Gamma(1 - 1/\alpha)t^{-1+1/\alpha}$. Similarly, we obtain an asymptotic lower bound, $\bar{p}^{(\infty)}(K) \gtrsim_K (1 + \varepsilon)^{1/\alpha - 1}c^{1/\alpha}\alpha^{-1}\Gamma(1 - 1/\alpha)t^{-1+1/\alpha}$. Taking $\varepsilon \to 0$, we complete the proof of the theorem.

Our proof of Theorem 4 follows a slightly different procedure than that of Theorem 3. In Theorem 3, an asymptotic expression for $\tau_i(K; J)$ in (11) is obtained from an asymptotic expression for $C_i(t; J)$ in (10). If we were to follow the same procedure as the proof of Theorem 3, the asymptotic upper bound would not match the asymptotic lower bound in Theorem 4. This difference stems from the fact that $C_i(t; J)$ is asymptotically polynomial in Theorem 3 and asymptotically poly-logarithmic in Theorem 4. 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 of [9]. Our proof relies on Lemma 3 and the following two lemmas.

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

Lemma 7. *Let* $f(t) \equiv \int_0^\infty \bar{\lambda}_x \exp(-\bar{\lambda}_x t)\,\mathrm{d}x$. *If* $\bar{\lambda}_x \sim_x c\exp(-\xi x^\beta)$, *where* $c, \xi, \beta > 0$, *then* $f(t) \sim_t (\ln(ct))^{1/\beta - 1}\xi^{-1/\beta}\beta^{-1}t^{-1}$.

Lemma 6 is a trivial extension of Lemma 6 of [9], and Lemma 7 is equivalent to Lemma 3 of [9] by Equation (7.49) of [9].

Proof of Theorem 4. We first study $\tau_i(K; J) = C_i^{-1}(K - \frac{1}{2}; J)$, where $C_i^{-1}(\cdot; J)$ is the inverse function of $C_i(t; J) \equiv \sum_{j \neq i}(1 - \exp(-\Lambda_j(u; J)))$. Observe that inequality (9) remains valid for $t > t_0$ when $\bar{\lambda}_i$ has a light tail. Let

$$C(t, \varepsilon) \equiv \sum_{j=1}^{\infty} (1 - \exp(-(1 + \varepsilon)\bar{\lambda}_j t)) = \int_0^\infty (1 - \exp(-(1 + \varepsilon)\bar{\lambda}_x t))\,\mathrm{d}x,$$

https://doi.org/10.1239/aap/1282924064 Published online by Cambridge University Press

where we extend the domain of $\bar{\lambda}_i$ to nonnegative real numbers, so that $\bar{\lambda}_x = \bar{\lambda}_{\lceil x \rceil}$. Note that $\bar{\lambda}_x \sim_x c \exp(-\xi x^\beta)$. Let $D(t,\varepsilon) \equiv C(t,-\varepsilon) - 1$. Then inequality (9) implies that $D(t,\varepsilon) \le C_i(t;J) \le C(t,\varepsilon)$ a.s. for any i and $t > t_0$. Let $C^{-1}(\cdot,\varepsilon)$ and $D^{-1}(\cdot,\varepsilon)$ respectively denote the inverse functions of $C(\cdot,\varepsilon)$ and $D(\cdot,\varepsilon)$. Since $C_i(\tau_i(K;J);J) = K - \frac{1}{2}$, there exists K_0 such that

$$C^{-1}\big(K-\tfrac{1}{2},\varepsilon\big) \le \tau_i(K;J) \le D^{-1}\big(K-\tfrac{1}{2},\varepsilon\big) \quad \text{a.s. for all } K > K_0.$$

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

The uniform convergence of $\tau_i(K;J)$ and Lemma 3 imply that inequality (13) with $\tau(K)$ replaced with $\hat{\tau}(K)$ remains valid when $\bar{\lambda}_i$ has a light tail. Hence, Lemma 7 implies that

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

Substituting $\hat{\tau}(K)$ into the above inequality, we obtain

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

Similarly, we obtain an asymptotic lower bound:

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

Now the theorem follows by taking $\varepsilon \to 0$.

6. Conclusion

We have introduced and demonstrated the usefulness of a fluid limit of a stochastic model for LRU with possibly nonstationary 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 nonstationary request processes in the fluid limit. To this end, Hirade and Osogami [8] showed that, in their fluid limit, the 2Q cache algorithm [13] 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 was 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 nonstationary request processes.

https://doi.org/10.1239/aap/1282924064 Published online by Cambridge University Press

Appendix A. Technical lemma and proofs

Proof of Lemma 1. Since the ℓth request for e_i is a miss if and only if at least K distinct items are requested in $(t_{\ell-1}^{(i)}, t_\ell^{(i)})$, we have

$$p_{i,\ell} = \mathrm{P}\left(\sum_{j\neq i} I\{\text{there exist } \kappa \text{ such that } t_{\ell-1}^{(i)} < t_\kappa^{(j)} < t_\ell^{(i)}\} \geq K\right).$$

Let $\mathcal{E}_i = \{\boldsymbol{\Psi} \mid \sum_{j\neq i} I\{t_1^{(j)} < t_1^{(i)}\} \geq K\}$. Since there exists κ such that $t_{\ell-1}^{(i)} < t_\kappa^{(j)} < t_\ell^{(i)}$ if and only if the first request for e_j after time $t_{\ell-1}^{(i)}$ is before $t_\ell^{(i)}$, we obtain $p_{i,\ell} = \mathrm{P}(\theta_{t_{\ell-1}^{(i)}}\boldsymbol{\Psi} \in \mathcal{E}_i)$. 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: $\bar{p}_i = \lim_{L\to\infty}(1/(L-1))\sum_{\ell=2}^{L}\mathrm{P}(\theta_{t_{\ell-1}^{(i)}}\boldsymbol{\Psi} \in \mathcal{E}_i) = \mathrm{P}^{0,i}(\mathcal{E}_i)$, which completes the proof of the lemma.

Proof of Lemma 3. Let $\{1,\ldots,U\}$ be the state space of J. For $1 \leq u \leq U$, let $V_u(t;J)$ be the time that J spends at state u by time t. Then, for any i, we have $\int_0^t \lambda_i(J(u))\,\mathrm{d}u = \sum_{u=1}^{U}\lambda_i(u)V_u(t,J)$. Let π_u be the stationary probability that J is at state u. Then $\bar{\lambda}_i = \sum_{u=1}^{U}\lambda_i(u)\pi_u$ for any i. Therefore, we have

$$\begin{aligned}\left|\frac{1}{t}\int_0^t \lambda_i(J(u))\,\mathrm{d}u - \bar{\lambda}_i\right| &= \left|\frac{1}{t}\sum_{u=1}^{U}\lambda_i(u)V_u(t,J) - \sum_{u=1}^{U}\lambda_i(u)\pi_u\right| \\ &\leq \sum_{u=1}^{U}\lambda_i(u)\left|\frac{V_u(t,J)}{t} - \pi_u\right|. \end{aligned} \tag{14}$$

Since J is an ergodic semi-Markov chain, $V_u(t;J)/t \to \pi_u$ a.s. as $t\to\infty$. Now the lemma follows from inequality (14), since U is finite.

Proof of Lemma 5. We first consider an asymptotic upper bound for the special case where $\bar{\lambda}_i = c/i^\alpha$. Let $\eta(x) = cx^{-\alpha}\exp(-ctx^{-\alpha})$. Then $\eta(\cdot)$ is increasing in $[0,(ct)^{1/\alpha})$ and decreasing in $[(ct)^{1/\alpha},\infty)$, so that $\eta(x) \leq \eta((ct)^{1/\alpha}) = 1/(\mathrm{e}t)$. Let $i_0 = \lceil (ct)^{1/\alpha}\rceil$. Then

$$\begin{aligned} f(t) &\leq \int_0^{i_0} cx^{-\alpha}\exp(-ctx^{-\alpha})\,\mathrm{d}x + \frac{1}{\mathrm{e}t} + \int_{i_0}^{\infty} cx^{-\alpha}\exp(-ctx^{-\alpha})\,\mathrm{d}x \\ &= \int_0^{\infty} cx^{-\alpha}\exp(-ctx^{-\alpha})\,\mathrm{d}x + \frac{1}{\mathrm{e}t}. \end{aligned}$$

Changing the variable with $y = ct/x^\alpha$, we obtain

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

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

$$f(t) \lesssim_t \frac{c^{1/\alpha}\Gamma(1-1/\alpha)t^{-1+1/\alpha}}{\alpha}. \tag{15}$$

Next, we consider the general case, where $\bar{\lambda}_i \sim_i c/i^\alpha$. For any ε, there exists j_0 such that $(1-\varepsilon)c/i^\alpha < \bar{\lambda}_i < (1+\varepsilon)c/i^\alpha$ for all $i > j_0$. Hence,

$$f(t) \leq \sum_{i=1}^{j_0}\bar{\lambda}_i\exp(-\bar{\lambda}_i t) + \sum_{i=j_0+1}^{\infty}(1+\varepsilon)ci^{-\alpha}\exp\left(-\frac{(1-\varepsilon)ct}{i^\alpha}\right).$$

https://doi.org/10.1239/aap/1282924064 Published online by Cambridge University Press

Let $\lambda^\star \equiv \min(\bar\lambda_1, \dots, \bar\lambda_{j_0})$. Since $\bar\lambda_i \le c$ for any i, we obtain

$$f(t) \le j_0 c \exp(-\lambda^\star t) + \frac{1+\varepsilon}{1-\varepsilon} \sum_{i=1}^{\infty} (1-\varepsilon) c i^{-\alpha} \exp\left(-\frac{(1-\varepsilon)ct}{i^\alpha}\right).$$

By inequality (15) we obtain

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

By taking $\varepsilon \to 0$ we obtain $f(t) \lesssim_t c^{1/\alpha}\alpha^{-1}\Gamma(1-1/\alpha)t^{-1+1/\alpha}$.

The corresponding asymptotic lower bound can be proved similarly, but also follows from a simpler argument. Since $e^{-x} \ge 1-x$ for any x, we have $f(t) \ge \sum_{i=1}^{\infty} \bar\lambda_i (1-\bar\lambda_i)^t$. Now Lemma 3.1 of [20] can be used to derive the asymptotic lower bound, which completes the proof of the lemma.

Lemma 8. *Let* $\zeta(m) \equiv (\sum_{j=0}^{M} c_j \exp(-j\chi/m))^m$, *where* $\sum_{j=0}^{M} c_j = 1$, $|\chi| < \infty$, *and* $0 \le M \le \infty$. *Then* $\zeta(m) \to \exp(-\chi \sum_{r=0}^{M} j c_j)$ *as* $m \to \infty$.

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

$$\lim_{m\to\infty} \ln \zeta(m) = \chi \lim_{x \downarrow 0} \frac{\ln(\sum_{j=0}^{M} c_j \exp(-jx))}{x}.$$

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

$$\lim_{m\to\infty} \ln \zeta(m) = -\chi \lim_{x\downarrow 0} \frac{\sum_{j=0}^{M} j c_j \exp(-jx)}{\sum_{j=0}^{M} c_j \exp(-jx)} = -\chi \sum_{j=0}^{M} j c_j,$$

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

Appendix B. Fluid limit defined with independent replications

In this section we prove that the dependencies in $\mathbf{\Psi}$ would disappear in the fluid limit defined with $\mathcal{S}^{(m)}$. Recall that, in $\mathcal{S}^{(m)}$, the $\mathbf{\Phi}_k$ for $1 \le k \le m$ are independent and identically distributed (i.i.d.) as $\mathbf{\Psi}$. Specifically, the following proposition holds for $\mathcal{S}^{(\infty)}$.

Proposition 1. *Suppose that* $\hat\Psi_j = \{\hat t_\ell,\ \ell \in \mathbb{Z}\}$ *is identically distributed as* Ψ_j *for* $1 \le j \le N$ *and that* $\hat\Psi_1, \dots, \hat\Psi_N$ *and* Ψ_i *are mutually independent. Let* $q_{i,k,\ell}$ *be the miss probability of the* ℓ*th request for* $e_{i,k}$ *in* $\mathcal{S}^{(m)}$. *Then*

$$\bar q_{i,k}^{(m)} \equiv \lim_{L\to\infty} \frac{1}{L} \sum_{\ell=1}^{L} q_{i,k,\ell}^{(m)} \to \mathrm{P}^{0,i}\left(\sum_{j=1}^{N} \mathrm{E}[I\{\hat t_1^{(j)} < t_1^{(i)}\} \mid t_1^{(i)}] \ge K\right)$$

as $m \to \infty$ *for any* k.

Observe that the expression of $\bar q_{i,k}^{(\infty)}$ is independent of the dependencies between Ψ_i and Ψ_j for $i \ne j$.

https://doi.org/10.1239/aap/1282924064 Published online by Cambridge University Press

Proof of Proposition 1. Since the $\mathbf{\Phi}_k$ for $1 \le k \le m$ are i.i.d., we consider the miss probability for $e_{i,1}$. In $\mathcal{S}^{(m)}$, let $\tilde{C}_{i,\ell}^{(m)}$ be the total size of distinct items that are requested in $(t_{\ell-1}^{(i,1)}, t_\ell^{(i,1)})$. Note that $q_{i,1,\ell}^{(m)} = \mathrm{P}(\tilde{C}_{i,\ell}^{(m)} \ge K)$.

We study the convergence of $\tilde{C}_{i,\ell}^{(m)}$, given $(t_{\ell-1}^{(i,1)}, t_\ell^{(i,1)})$, as $m \to \infty$ by showing the convergence of its Laplace transform, $\tilde{\psi}_{i,\ell}^{(m)}(s) \equiv \mathrm{E}[\exp(-sC_{i,\ell}) \mid t_{\ell-1}^{(i,1)}, t_\ell^{(i,1)}]$ for $0 \le s < \infty$. Let $\tilde{I}_\ell(j,k)$ be the indicator random variable such that $\tilde{I}_\ell(j,k) = 1$ if and only if $e_{j,k}$ is requested in $(t_{\ell-1}^{(i,1)}, t_\ell^{(i,1)})$. Note that $\tilde{I}_\ell(i,1) = 0$. Then

$$\begin{aligned}\tilde{\psi}_{i,\ell}^{(m)}(s) &= \mathrm{E}\left[\exp\left(-\frac{s}{m}\sum_{j=1}^{N}\sum_{k=1}^{m}\tilde{I}_\ell(j,k)\right) \,\middle|\, t_{\ell-1}^{(i,1)}, t_\ell^{(i,1)}\right] \\ &= \mathrm{E}\left[\exp\left(-\frac{s}{m}\sum_{j=1}^{N}\tilde{I}_\ell(j,2)\right) \,\middle|\, t_{\ell-1}^{(i,1)}, t_\ell^{(i,1)}\right]^{m-1} \\ &\quad \times \mathrm{E}\left[\exp\left(-\frac{s}{m}\sum_{j=1}^{N}\tilde{I}_\ell(j,1)\right) \,\middle|\, t_{\ell-1}^{(i,1)}, t_\ell^{(i,1)}\right],\end{aligned}$$

where the last equality holds since the $\mathbf{\Phi}_k$ for $1 \le k \le m$ are i.i.d. Similar to (3), we can show that

$$\lim_{m\to\infty}\mathrm{E}\left[\exp\left(-\frac{s}{m}\sum_{j=1}^{N}\tilde{I}_\ell(j,2)\right) \,\middle|\, t_{\ell-1}^{(i,1)}, t_\ell^{(i,1)}\right]^{m-1} = \exp\left(-s\,\mathrm{E}\left[\sum_{j=1}^{N}\tilde{I}_\ell(j,2) \,\middle|\, t_{\ell-1}^{(i,1)}, t_\ell^{(i,1)}\right]\right).$$

Also, by the dominated convergence theorem, we can exchange the limit and the expectation to obtain

$$\lim_{m\to\infty}\mathrm{E}\left[\exp\left(-\frac{s}{m}\sum_{j=1}^{N}\tilde{I}_\ell(j,1)\right) \,\middle|\, t_{\ell-1}^{(i,1)}, t_\ell^{(i,1)}\right] = 1.$$

Therefore, as $m \to \infty$, we have $\psi_{i,\ell}^{(m)}(s) \to \exp(-s\,\mathrm{E}[\sum_{j=1}^{N}\tilde{I}_\ell(j,2) \mid t_{\ell-1}^{(i,1)}, t_\ell^{(i,1)}])$.

Now, the continuity theorem and the linearity of expectation imply that, as $m \to \infty$, we have $\tilde{C}_{i,\ell}^{(m)} \xrightarrow{\mathrm{D}} \sum_{j=1}^{N}\mathrm{E}[\tilde{I}_\ell(j,2) \mid t_{\ell-1}^{(i,1)}, t_\ell^{(i,1)}]$. Since $\mathbf{\Phi}_1$ and $\mathbf{\Psi}_2$ are independent in $\mathcal{S}^{(m)}$, it follows that $\tilde{C}_{i,\ell}^{(\infty)}$ does not depend on the dependencies in the $\mathbf{\Psi}$. In particular, $\tilde{I}_\ell(j,2)$ is identically distributed as $I\{\text{there exists } \kappa \text{ such that } t_{\ell-1}^{(i)} < \hat{t}_\kappa^{(j)} < t_\ell^{(i)}\}$ given $(t_{\ell-1}^{(i)}, t_\ell^{(i)})$. Therefore, we have

$$\begin{aligned}q_{i,k,\ell}^{(m)} &= \mathrm{P}(\tilde{C}_{i,\ell}^{(m)} \ge K) \\ &\to \mathrm{P}\left(\sum_{j=1}^{N}\mathrm{E}[I\{\text{there exists } \kappa \text{ such that } t_{\ell-1}^{(i)} < \hat{t}_\kappa^{(j)} < t_\ell^{(i)}\} \mid t_{\ell-1}, t_\ell] \ge K\right).\end{aligned}$$

Now the proposition can be proved in the same way as Lemma 1.

Acknowledgements

We thank Naoto Miyoshi for his helpful comments on technical aspects of the paper and Shannon Jacobs for his English rewriting on an earlier version of the paper. A five-page extended abstract of a preliminary version of the paper appears in [17].

References

[1] Burville, P. J. and Kingman, J. F. C. (1973). On a model for storage and search. *J. Appl. Prob.* **10,** 697–701.

[2] Chu, J.-H. and Knott, G. D. (1993). A new method for computing page-fault rates. *SIAM J. Comput.* **22,** 1319–1330.

[3] Coffman, E. G., Jr. and Jelenković, P. (1999). Performance of the move-to-front algorithm with Markov-modulated request sequences. *Operat. Res. Lett.* **25,** 109–118.

[4] Fill, J. A. (1996). Limits and rates of convergence for the distribution of search cost under the move-to-front rule. *Theoret. Comput. Sci.* **164,** 185–206.

[5] Fill, J. A. and Holst, L. (1996). On the distribution of search cost for the move-to-front rule. *Random Structures Algorithms* **8,** 179–186.

[6] Flajolet, P., Gardy, D. and Thimonier, L. (1992). Birthday paradox, coupon collectors, caching algorithms and self-organizing search. *Discrete Appl. Math.* **39,** 207–229.

[7] Fristedt, B. and Gray, L. (1997). *A Modern Approach to Probability Theory*. Birkhäuser, Boston, MA.

[8] Hirade, R. and Osogami, T. (2010). Analysis of page replacement policies in the fluid limit. *Operat. Res.* **58,** 971–984

[9] Jelenković, P. R. (1999). Asymptotic approximation of the move-to-front search cost distribution and least-recently-used caching fault probabilities. *Ann. Appl. Prob.* **9,** 430–464.

[10] Jelenković, P. R. and Radovanović, A. (2004). Least-recently-used caching with dependent requests. *Theoret. Comput. Sci.* **326,** 293–327.

[11] Jelenković, P. R. and Radovanović, A. (2008). The persistent-access-caching algorithm. *Random Structures Algorithms* **33,** 219–251.

[12] Jelenković, P. R., Radovanović, A. and Squillante, M. S. (2006). Critical sizing of LRU caches with dependent requests. *J. Appl. Prob.* **43,** 1013–1027.

[13] Johnson, T. and Shasha, D. (1994). 2Q: low overhead high performance buffer management replacement algorithm. In *Proc. 20th Internat. Conf. Very Large Data Bases*, Morgan Kaufmann, San Francisco, CA, pp. 439–450.

[14] Lam, K., Leung, M. Y. and Siu, M. K. (1984). Self-organizing files with dependent accesses. *J. Appl. Prob.* **21,** 343–359.

[15] McCabe, J. (1965). On serial files with relocatable records. *Operat. Res.* **13,** 609–618.

[16] Nelson, R. (1995). *Probability, Stochastic Processes, and Queueing Theory*. Springer, New York.

[17] Osogami, T. (2009). A fluid limit for cache algorithms with general request processes. In *Proc. IEEE INFOCOM 2009*, pp. 2836–2840.

[18] Rodrigues, E. R. (1995). The performance of the move-to-front scheme under some particular forms of Markov requests. *J. Appl. Prob.* **32,** 1089–1102.

[19] Sigman, K. (1995). *Stationary Marked Point Processes: An Intuitive Approach*. Chapman & Hall, New York.

[20] Sugimoto, T. and Miyoshi, N. (2006). On the asymptotics of fault probability in least-recently-used caching with Zipf-type request distribution. *Random Structures Algorithms* **29,** 296–323.