

# A fluid limit for cache algorithms with general request processes (extended abstract)

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

**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.

## I. INTRODUCTION

Caching data is a widely used technique for scalability and efficiency in today’s 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.

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 the

first  $K$  positions of the MTF list. McCabe [14] 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 [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] 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 [8] studies the miss probability for LRU in the limiting case where the cache size,  $K$ , approaches infinity. In particular, when the request rates,  $\lambda_i$  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 [8] 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 [7] show that the miss probabilities for LRU and the 2Q cache algorithm [12], 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 [10] 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. [13] and Rodrigues [16], respectively, derive the mean and the variance of the stationary position of a requested item in an MTF list,

and Chu and Knott [2] derive an expression for the stationary miss probability for LRU. Coffman and Jelenković [3] 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 [9] and Sugimoto and Miyoshi [18] 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 [3] as  $K \rightarrow \infty$ . Jelenković et al. [11] characterize 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 non-trivial extension of the fluid limits for the independent reference model in [8], [7]. 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 non-stationary. 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  $N \geq 128$  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  $K \rightarrow \infty$ ,  $\bar{p}^{(\infty)}$  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], [18]. 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  $K \rightarrow \infty$  is not known even for the independent reference model. Recall that Jelenković [8] 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 II, we derive an expression for  $\bar{p}$ . In Section III, we define the fluid limit and formally derive a general expression for  $\bar{p}^{(\infty)}$ . In Section IV, 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 V, we show that  $\bar{p}^{(\infty)}$  is asymptotically insensitive to particular dependencies in the request process. Throughout, we omit proofs and details, which are provided in the associated technical report [15].

## II. 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,  $\Psi$ . We first define the model of caching with LRU and state assumptions on  $\Psi$ . We then analyze the average miss probability for LRU, which will be used in Section III to study the fluid limit.

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  $\ell$ -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  $\ell, \ell', i, j$ . 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  $\Psi$  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,  $\Psi$  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  $\ell$ -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 [17] and make additional assumptions about  $\Psi$ . Let  $\theta_t$  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  $\ell$  such that  $t_\ell^{(i)} \leq t$ . Let  $\theta_t \Psi = (\theta_t \Psi_1, \dots, \theta_t \Psi_N)$ . We assume that  $\Psi$  is time-asymptotically stationary, so that there exists a distribution defined by  $\mathbf{P}^*(\Psi \in \mathcal{E}) \equiv \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t \mathbf{P}(\theta_u \Psi \in \mathcal{E}) du$ . Note that a non-stationary  $\Psi$  can be time-asymptotically stationary. For simplicity, we assume that  $\Psi$  is ergodic with respect to  $\mathbf{P}^*$ .

Finally, we assume that the average request rate,  $\lambda_i$ , of  $e_i$  satisfies  $0 < \lambda_i < \infty$  for  $i = 1, \dots, N$ . Formally,  $\lambda_i \equiv$

$\mathbf{E}^*[M^{(i)}(1)]$ , where  $M^{(i)}(1)$  denotes the number of requests for  $e_i$  in  $(0, 1]$ , and  $\mathbf{E}^*$  denotes the expectation with respect to  $\mathbf{P}^*$ . When  $N = \infty$ , we also assume that  $\sum_{i=1}^N \lambda_i < \infty$ .

Under the above assumptions,  $\Psi_i$  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 [17]). Then  $\bar{p}_i$  can be expressed conveniently using  $\mathbf{P}^{0,i}$ :

*Lemma 1:* When  $\Psi$  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.

Lemma 1 can be explained intuitively as follows. We may see  $\mathbf{P}^{0,i}(\mathcal{E})$  as the probability of an event,  $\mathcal{E}$ , when we “randomly observe way out at” [17] 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  $\sum_{j \neq i} I\{t_1^{(j)} < t_1^{(i)}\} \geq K$ .

### III. 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.

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  $\ell$ -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 [7], [8]. For example, the  $m$ -th scaled system,  $\mathcal{S}^{(m)}$ , of [7] can be seen as a superposition of independent replications of the original system. Specifically, in  $\mathcal{S}^{(m)}$ ,  $\Phi_k$  for  $1 \leq k \leq m$  are independent and stochastically identical to  $\Psi$ . Unfortunately, the dependencies in  $\Psi$  would disappear in  $\mathcal{S}^{(\infty)}$  in the sense that  $\mathcal{S}^{(\infty)}$  with general  $\Psi$  is identical to that when  $\Psi$  is a vector of independent Poisson processes. We formally prove the above observation in [15].

We will define our scaled system as a superposition of *dependent* replications of the original system. Also, in contrast to [7], [8], we will define a sequence of scaled systems for each  $e_i$ , so that the scaled systems for different items have different dependencies in  $\Phi_k$ . 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  $\mathcal{T}_i^{(m)}$ . In  $\mathcal{T}_i^{(m)}$ , we assume that  $\Phi_k$  is stochastically identical to  $\Psi$  (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  $\Phi_k$  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  $\Phi_k$  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  $\mathbf{P}(\Psi_k \in \mathcal{E}_k, \forall k \in \{1, \dots, m\} \mid \Psi_{i,1}) = \prod_{k=1}^m \mathbf{P}(\Psi_k \in \mathcal{E}_k \mid \Psi_{i,1})$ .

To clarify the assumptions on  $\Phi$ , 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  $\Psi_i$  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  $\Psi_j$  for all  $j \neq i$  in the original system in such a way that  $\Psi_i(\omega)$  and  $\Psi_j$  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  $\Psi_j$  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, 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  $\ell$ ,  $t$ , and  $e_i$ .

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  $1 \leq k \leq m$  are not in the cache upon the request. Let  $p_{i,\ell}^{(m)}$  be the probability that the  $\ell$ -th request for  $e_i$  is a miss in  $\mathcal{T}_i^{(m)}$ . We study the average miss probability,  $\bar{p}_i^{(m)} \equiv \lim_{L \rightarrow \infty} \frac{1}{L} \sum_{\ell=1}^L p_{i,\ell}^{(m)}$ , of  $e_i$  as  $m \rightarrow \infty$ .

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

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,  $\mathbf{E} \left[ I\{t_1^{(j)} < t_1^{(i)}\} \mid \Psi_i \right]$ , 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  $\Psi_i$  and by the expected impact that  $\Psi_i$  has on  $\Psi_j$  for  $j \neq i$  via the dependencies between  $\Psi_i$  and  $\Psi_j$  for  $j \neq i$ .

### IV. INHOMOGENEOUS POISSON REQUESTS

In this section, we study the  $\bar{p}_i^{(\infty)}$  derived in Section III in more detail for the particular case when the requests are issued according to inhomogeneous Poisson processes. We first 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. We then study the accuracy of approximating  $\bar{p}_i$  with  $\bar{p}_i^{(\infty)}$ .

*Theorem 2:* Let  $\Lambda_i(t, u) \equiv \int_t^u \lambda_i(v) dv$  and let  $\tau_i(t)$  be the maximum  $u$  such that  $\sum_{j \neq i} (1 - \exp(-\Lambda_j(t, u))) \leq K - \frac{1}{2}$  for  $1 \leq i \leq N$ . In addition to the conditions of Lemma 1, if  $\Psi_i$  is an inhomogeneous Poisson process with rate  $\lambda_i(t)$  at time  $t$  for  $1 \leq i \leq N$ , then  $\bar{p}_i^{(\infty)} = \frac{1}{\lambda_i} \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \exp(-\Lambda_i(t, \tau_i(t))) \lambda_i(t) dt$ , where  $\lambda_i = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \lambda_i(t) dt$ .

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 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.

Figure 1(a) 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(a) 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, Figure 1(b) 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 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 [15], we show the results with other choices of  $\lambda_i(\cdot)$ . We find that the qualitative findings from Figure 1 hold for other settings of  $\lambda_i(\cdot)$ .

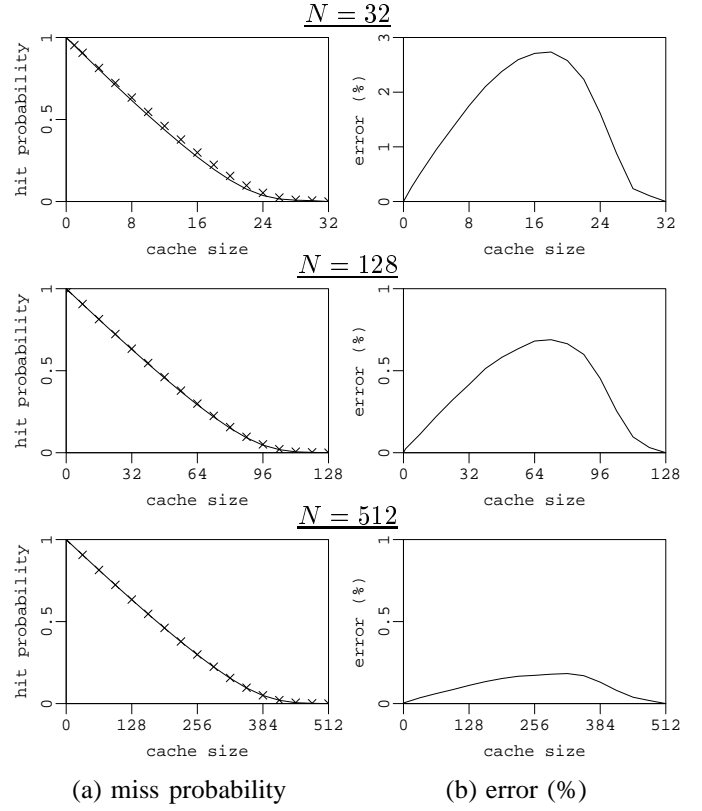


Fig. 1. The accuracy of approximating  $\bar{p}$  with  $\bar{p}^{(\infty)}$  when requests follow inhomogeneous Poisson processes, where  $N$  is set as shown in each row. In Column (a), solid lines show  $\bar{p}^{(\infty)}$ , and  $\times$  marks show  $\bar{p}$ . Column (b) shows the error (%) of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

## V. ASYMPTOTIC ANALYSIS WITH FLUID LIMIT

In this section, we study the request processes that are similar to those studied in [3], [9], [11], [18]. Specifically, 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  $\lambda_i(J(t))$  for  $1 \leq i \leq N$ . Thus, given  $J(\cdot)$ ,  $\Psi_i$  is an inhomogeneous Poisson process with rate  $\lambda_i(J(t))$  at time  $t$ . Observe that  $\Psi_i$  for  $i = 1, \dots, N$  are conditionally independent given  $J(\cdot)$ . Note that  $\Psi$  is stationary, which is also assumed in [3], [9], [11], [18], so that the stationary miss probability exists (see Lemma 2.1 from [18]) and agrees with the average miss probability.

We first derive  $\bar{p}_i^{(\infty)}$  for the particular request processes under consideration.

*Lemma 2:* Let  $\Lambda_i(u; J) \equiv \int_0^u \lambda_i(J(v)) dv$  and  $\tau_i(K; J)$  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  $\Psi_i$  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  $\lim_{m \rightarrow \infty} \bar{p}_i^{(m)} = \frac{1}{\lambda_i} \mathbf{E}[\exp(-\Lambda_i(\tau_i(K; J); J)) \lambda_i(J(0))]$ .

Now, we consider the overall average miss probability in the fluid limit,  $\bar{p}^{(\infty)}(K) \equiv \sum_{i=1}^N r_i \bar{p}_i^{(\infty)}$ , 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 first consider the case when  $\lambda_i$  has a heavy tail. We find that  $\bar{p}^{(\infty)}(K)$  decays with a power law as  $K \rightarrow \infty$  and is asymptotically insensitive to  $J(\cdot)$ . Formally,

*Theorem 3:* In addition to the conditions in Lemma 2, suppose that  $\lambda_i \sim c/i^\alpha$  for  $i = 1, 2, \dots$ , where  $\alpha > 1$ ,  $c > 0$ , and  $a_i \sim b_i$  denotes  $\lim_{i \rightarrow \infty} a_i/b_i = 1$ . Then  $\bar{p}^{(\infty)}(K)$  is asymptotically insensitive to  $J(\cdot)$  as  $K \rightarrow \infty$ , and it holds that  $\bar{p}^{(\infty)}(K) \sim \frac{c}{\alpha} \Gamma(1 - 1/\alpha) K^{1-\alpha}$ , where  $\Gamma(z) \equiv \int_0^\infty e^{-y} y^{z-1} dy$  denotes the gamma function.

Theorem 3, which is obtained for the fluid limit, is in agreement with the asymptotic results for the original system derived in [9], [18]. 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  $\lambda_i$  has a light tail. This case has not been fully investigated in the prior work. Jelenković [8] studies asymptotic properties of the overall stationary miss probability in his fluid limit when  $\lambda_i$  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  $K \rightarrow \infty$  and is asymptotically insensitive to  $J(\cdot)$ . Formally,

*Theorem 4:* In addition to the conditions in Lemma 2, suppose that  $\lambda_i \sim c \exp(-\xi i^\beta)$  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.

Finally, we discuss the case when  $J(\cdot)$  is a constant (i.e., each  $\Psi_i$  is an independent Poisson process) to gain further insights into our fluid limit. The following corollary can be compared against the stationary miss probabilities in the fluid limits obtained in [8], [7].

*Corollary 1:* If  $\Psi_i$  is an independent Poisson process with rate  $\lambda_i$  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  $C_i(t) \equiv \sum_{j \neq i} (1 - \exp(-\lambda_j t))$ .

The corollary can be understood as follows. Suppose that  $e_{i,1}$  is requested and move 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,  $1 - \exp(-\lambda_j t)$ , 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  $1 \leq k \leq m$  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  $t = \tau_i(K)$ . The probability that the next request for  $e_{i,1}$  is issued after  $t = \tau_i(K)$  is  $\exp(-\lambda_i \tau_i(K))$ . 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.

Since our fluid limit differs from the fluid limits defined in [8], [7], our  $\bar{p}_i^{(\infty)}$  differs from those derived in [8], [7]. However, the only difference between our  $\bar{p}_i^{(\infty)}$  and that in [7] is that, in [7],  $\tau_i(K)$  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 [8] and [7] are discussed in [7]. We find that these differences become negligible when we study the asymptotic characteristics.

## VI. CONCLUDING REMARK

We hope 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 [7] show that, in a fluid limit, the 2Q cache algorithm [12] 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] P. J. Burville and J. F. C. Kingman. On a model for storage and search. *Journal of Applied Probability*, 10(3):697–701, 1973.
- [2] J.-H. Chu and G. D. Knott. A new method for computing page-fault rates. *SIAM Journal on Computing*, 22(6):1319–1330, 1993.
- [3] 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.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] R. Hirade and T. Osogami. Analysis of page replacement policies in the fluid limit. Technical Report RT0768 (submitted for publication), IBM Research, 2007.
- [8] 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.
- [9] P. R. Jelenković and A. Radovanović. Least-recently-used caching with dependent requests. *Theoretical Computer Science*, 326(1-3):293–327, 2004.
- [10] P. R. Jelenković and A. Radovanović. The persistent-access-caching algorithm. *Random Structures & Algorithms*, in press, 2008.
- [11] 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.
- [12] 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.
- [13] K. Lam, M. Leung, and M. Siu. Self-organizing files with dependent accesses. *Journal of Applied Probability*, 21(2):343–359, 1984.
- [14] J. McCabe. On serial files with relocatable records. *Operations Research*, 13(4):609–618, 1965.
- [15] T. Osogami. A fluid limit for cache algorithms with general request processes. Technical Report RT0813, IBM Research, 2008 (available at <http://www.research.ibm.com/tr/people/osogami/paper/rt0813.pdf>).
- [16] 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.
- [17] K. Sigman. *Stationary Marked Point Processes: An Intuitive Approach*. Chapman & Hall, New York, NY, 1995.
- [18] 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.