ISSN 1342-2804

Research Reports on Mathematical and Computing Sciences

Fluid limit analysis of the FIFO and RR caching for the independent reference model

Naoki Tsukada, Ryo Hirade
and Naoto Miyoshi

July 2011, B–465

Department of
Mathematical and
Computing Sciences
Tokyo Institute of Technology

SERIES B: Applied Mathematical Science

Fluid limit analysis of the FIFO and RR caching for the independent reference model

Naoki Tsukada[a] Ryo Hirade[a,b] Naoto Miyoshi[a*]

[a]Tokyo Institute of Technlogy

[b]IBM Research–Tokyo

Abstract

We study the fluid limit analysis of the random replacement (RR) caching for the independent reference model. Applying the limit theorem for the mean field interaction model, we derive the fluid limit of fault probability in the transient state as well as in the steady state. Since it is known that the stationary fault probability for the RR cache is identical to that for the first-in first-out (FIFO) cache, our result on the stationary fault probability is valid for the FIFO caching. We see that the fluid limit of stationary fault probability, which we obtain, is coincident with the known result by an intuitive approximation; that is, our fluid limit analysis gives a rigorous theoretical foundation to the intuitive approximation.

Keywords: Caching algorithms; first-in first out; random replacement; fluid limit analysis; mean field interaction models.

*Corresponding author: Department of Mathematical and Computing Sciences, Tokyo Institute of Technology, 2-12-1-W8-52 Ookayama, Tokyo 152-8552, Japan. E-mail: miyoshi@is.titech.ac.jp

1 Introduction

In computer systems and communication networks, caching technique is used for high-speed access to a finite subset out of a large number of items by storing the subset in a quickly accessible memory, called a cache. In order for this technique to work well, a replacement rule called a caching algorithm is crucial; that is, which items should be stored in the cache and how they should be updated. The performance of a caching algorithm is often evaluated in terms of the fault probability, that is the probability with which the requested item is not found in the cache.

The least-recently-used (LRU) and first-in first-out (FIFO) caching algorithms are well-known simple algorithms and have been studied in the literature. To keep frequently requested items in the cache, the LRU algorithm works as follows; that is, when the requested item is not in the cache, the least recently requested item in the cache is replaced with the requested one. In the FIFO algorithm, on the other hand, when the requested item is not in the cache, the oldest item in the cache is replaced with the requested one. While the LRU algorithm shows relatively good performance and has been applied in many systems, the FIFO is shown to have higher fault probability than the LRU for the independent reference model, where requests of items are independent and identically distributed (i.i.d.) (see, e.g., Berg & Gandolfi [11]). Nevertheless, since some complex caching algorithms combining the LRU and FIFO, such as the Full2Q by Johnson & Shasha [8] and the Multi-Queue by Zhou *et al.* [12], have been proposed recently, it is still meaningful to study the FIFO algorithm to evaluate the performance of such complex algorithms.

The existing works concerning the analysis of FIFO algorithm are almost done for the independent reference model. King [9] considered a homogeneous Markov chain representing the evolution of cache contents and derived the stationary fault probability for the FIFO cache, as well as for the LRU. The evaluation of fault probability based on King's analysis, however, suffers from the computational complexity when the number of items and/or the capacity of cache are large. Dan & Towsley [3] then presented a computationally efficient approximation evaluating the stationary fault probability for both the FIFO and LRU caches. While it is reported that their approximation has good agreement with simulation results, the argument for derivation is rather intuitive. Also, Gelenbe [4] showed the identity of stationary fault probability for the FIFO cache with that for the random replacement (RR) cache, where an item is chosen uniformly at random in the cache and is replaced with the requested one in case it is not in the cache.

Since the stationary fault probabilities for the FIFO and RR are identical ([4]), we, in this paper, investigate the RR caching algorithm for the independent reference model, instead of the FIFO. We here study the fluid limit analysis of the RR caching. As for the LRU caching, several works consider the fluid limit analysis and present some simple expressions evaluating the fault probability (see, e.g., Jelenković [7], Hirade & Osogami [6] for some complex caching algorithms, and Hattori & Hattori [5] for the related move-to-front list). It is, however, difficult to apply the techniques therein to the RR caching as well as to the FIFO. We thus associate the stochastic model of RR caching with a mean field interaction model studied by Benaïm & Le Boudec [1] (see also Bordenave *et al.* [2]) and apply the limit theorem for it. While the limit theorem for the mean field interaction model in [1] captures the transient behavior of the model in average, it does not yield a satisfiable result on the stationary behavior. For the stationary analysis, we apply the recent result by Le Boudec [10] exploiting the reversibility of the stochastic model under consideration. We can see that the fluid limit of stationary fault probability, which we obtain, is coincident with the approximated fault probability provided in [3], so that, we can say that our fluid limit analysis gives a rigorous theoretical foundation to Dan-Towsley's approximation.

The rest of this paper is organized as follows. In the next section, we describe our stochastic model of caching system and make a brief review on some related existing results in [9], [4] and [3]. The main

results are given in Section 3, where we first derive the fluid limit of empirical measure for the cache contents in the transient state and we then obtain the corresponding fluid limit in the steady state. The fluid limits of fault probabilities are derived from those of the empirical measure. The proofs are provided in Section 4, where we apply the limit theorem for the mean field interaction model in [1] for the proof of transient result and then apply the result of [10] for the steady state. We finally make some concluding remarks in Section 5.

2 Model and related existing results

The model consists of the set of items and a buffer with finite capacity, called a cache. The number of items is N ($\in \mathbb{N} = \{1, 2, \ldots\}$) and the set of items is denoted by $\mathcal{N} = \{1, 2, \ldots, N\}$. We assume that all items are of the same size and the cache has the capacity of K ($< N$) items. An item is requested randomly among $\mathcal{N}$ at each time slot. If the requested item is in the cache, then no change occurs. When the requested item is not in the cache, we call it a cache fault. What happens when a cache fault occurs depends on the caching algorithm. In this paper, we consider the following two algorithms.

First-in first-out (FIFO): When the requested item is not in the cache, the item which stays in the cache for the longest time is replaced with the requested one.

Random replacement (RR): When the requested item is not in the cache, an item is chosen uniformly at random among ones in the cache and it is replaced with the requested one.

Throughout the paper, we consider the independent reference model; that is, requested items are independent and identically distributed for all time slots. The probability with which item i ($\in \mathcal{N}$) is requested is denoted by p_i, satisfying $p_i \geq 0$ for all $i \in \mathcal{N}$ and $\sum_{i=1}^{N} p_i = 1$.

In the remainder of this section, we make a brief review on related existing results for the independent reference models of the FIFO and RR caches; that is, i) King's exact stationary analysis of the FIFO caching in [9], ii) Gelenbe's identity of the stationary fault probabilities for the FIFO and RR caches in [4] and iii) Dan-Towsley's approximation of the stationary fault probability for the FIFO cache in [3]. Though they also analyzed the LRU caching, we here omit the results on it. For the moment, we assume that $p_i > 0$ for all $i \in \mathcal{N}$ (while we relax this assumption in the next section).

King's exact stationary analysis of the FIFO caching ([9]): For the FIFO caching algorithm, King [9] considered the evolution of cache contents as a homogeneous Markov chain and derived the fault probability in the steady state. Consider the list of K items corresponding to the contents of FIFO cache as follows. When the requested item is found in the list, the list remains unchanged. On the other hand, if the requested item is not in the list, then it is placed at the first position of the list, other items are shifted one position down and the item at the last (Kth) position is pushed out. The item at the first position is then the newest one in the cache and the item at the last position is the oldest. This list forms a homogeneous Markov chain within the state space $\Lambda_{N,K}$, which denotes the set of K-permutations (arrangements of K elements) taken from $\mathcal{N} = \{1, 2, \ldots, N\}$; that is, $\Lambda_{N,K} = \{(i_1, i_2, \ldots, i_K) \in \mathcal{N}^K : i_k \neq i_\ell \text{ for } k \neq \ell\}$. Since this Markov chain is irreducible in the finite state space, the unique stationary distribution exists and is given by

$$\pi_{\mathrm{FIFO}}(i_1, i_2, \ldots, i_K) = \frac{p_{i_1} p_{i_2} \cdots p_{i_K}}{\sum_{(j_1,\ldots,j_K)\in\Lambda_{N,K}} p_{j_1} p_{j_2} \cdots p_{j_K}}, \quad (i_1, i_2, \ldots, i_K) \in \Lambda_{N,K}. \tag{1}$$

The stationary fault probability ρ_{FIFO} is the probability with which the requested item is not found in the list, so that (see [9] for detail),

$$\rho_{\mathrm{FIFO}} = \sum_{A \in \Lambda_{N,K}} \pi_{\mathrm{FIFO}}(A) \sum_{j \in \mathcal{N} \setminus A} p_j.$$

Unfortunately, evaluating the fault probability by using this formula suffers from the computational complexity when the number of items and/or the capacity of cache become large.

Gelenbe's identity of the stationary fault probabilities for the FIFO and RR caches ([4]): For the RR caching algorithm, we can consider a Markov chain whose state represents the set of items in the cache. Let $\Theta_{N,K}$ denote the state space of this Markov chain; that is, $\Theta_{N,K} = \{A \subset \mathcal{N} : |A| = K\}$, the set of subsets of $\mathcal{N}$ with size K, where $|A|$ denotes the size of set A. If the requested item is found in the set A representing the current state, then the state remains unchanged. However, if the requested item is not in the current set A, then an item in A is chosen uniformly at random and it is replaced with the requested one. This Markov chain is also homogeneous and irreducible in the finite state space, so that, the unique stationary distribution exists and is given by

$$\pi_{\mathrm{RR}}(\{i_1, i_2, \ldots, i_K\}) = \frac{p_{i_1} p_{i_2} \cdots p_{i_K}}{\sum_{\{j_1,\ldots,j_K\} \in \Theta_{N,K}} p_{j_1} p_{j_2} \cdots p_{j_K}}, \quad \{i_1, i_2, \ldots, i_K\} \in \Theta_{N,K}. \tag{2}$$

The stationary fault probability ρ_{RR} is then given by

$$\rho_{\mathrm{RR}} = \sum_{A \in \Theta_{N,K}} \pi_{\mathrm{RR}}(A) \sum_{j \in \mathcal{N} \setminus A} p_j.$$

Here, noting in (1) and (2) that

$$\sum_{(i_1,\ldots,i_K) \in \Lambda_{N,K}} p_{i_1} p_{i_2} \cdots p_{i_K} = K! \times \sum_{\{i_1,\ldots,i_K\} \in \Theta_{N,K}} p_{i_1} p_{i_2} \cdots p_{i_K},$$

we can find that $\rho_{\mathrm{FIFO}} = \rho_{\mathrm{RR}}$; that is, the fault probabilities for the RR and FIFO caches are identical in the steady state (see [4] for detail).

Dan-Towsley's approximation of the stationary fault probability for the FIFO cache ([3]): Since the evaluation of fault probability based on the exact analysis suffers from the computational complexity, Dan & Towsley [3] presented a computationally efficient approximation evaluating the stationary fault probability for the FIFO cache. Consider, as in King's analysis, the list of K items representing the contents of FIFO cache. Let $Y_i \in \{0, 1, \ldots, K\}$, $i \in \mathcal{N}$, denote a random variable representing the position of item i in the list in the steady state, where $Y_i = 0$ means that item i is not in the cache. The probability with which item i is not in the cache and is brought in is given by $p_i \,\mathsf{P}(Y_i = 0)$. On the other hand, the probability with which item i is in the cache and is pushed out by a newly requested one is $\sum_{j \in \mathcal{N} \setminus \{i\}} p_j \,\mathsf{P}(Y_i = K, Y_j = 0)$. According to the flow conservation in the steady state, we then have

$$p_i \,\mathsf{P}(Y_i = 0) = \sum_{j \in \mathcal{N} \setminus \{i\}} p_j \,\mathsf{P}(Y_i = K, Y_j = 0), \quad i \in \mathcal{N}. \tag{3}$$

By the stationary distribution (1), we have $\mathsf{P}(Y_i = \ell) = \mathsf{P}(Y_i \neq 0)/K$ for $\ell = 1, 2, \ldots, K$. Thus, taking $\mathsf{P}(Y_i = K, Y_j = 0) \approx K^{-1} \,\mathsf{P}(Y_i \neq 0) \,\mathsf{P}(Y_j = 0)$ and $\mathcal{N} \setminus \{i\} \approx \mathcal{N}$ in (3) approximately,

$$p_i \,\mathsf{P}(Y_i = 0) \approx \frac{\mathsf{P}(Y_i \neq 0)}{K} \,\rho, \quad i \in \mathcal{N},$$

where $\rho = \sum_{j=1}^{N} p_j \, \mathsf{P}(Y_j = 0)$ denotes the stationary fault probability. Some algebraic manipulation yields $\mathsf{P}(Y_i \neq 0) \approx K\, p_i/(\rho + K\, p_i)$, so that, since $\sum_{i=1}^{N} \mathsf{P}(Y_i \neq 0) = \mathsf{E}\big(\sum_{i=1}^{N} \mathbf{1}_{\{Y_i \neq 0\}}\big) = K$, we obtain the approximation of stationary fault probability as the unique solution ρ on $[0, 1)$ to

$$\sum_{i=1}^{N} \frac{K\, p_i}{\rho + K\, p_i} = K. \tag{4}$$

It is reported in [3] that equation (4) is well solved numerically and the approximation is valid in many cases.

3 Fluid limit analysis of random replacement caching

In this section, we consider the fluid limit of RR caching. Since the stationary fault probability for the RR cache is identical to that for the FIFO ([4]), our result on the stationary fault probability is still valid for the FIFO caching and it is shown that the same result as Dan-Towsley's approximation is derived exactly in the fluid limit. We here relax the assumption that $p_i > 0$ for all $i \in \mathcal{N}$, which is imposed in the preceding section, and we write $\mathcal{N}_0 = \{i \in \mathcal{N} : p_i = 0\}$ and $\mathcal{N}_+ = \{i \in \mathcal{N} : p_i > 0\}$.

To derive the fluid limit, we consider scaling the original RR caching model as follows. Let n ($\in \mathbb{N}$) denote a scaling parameter and consider the nth scaled model such that the number of items is $n\,N$ and the capacity of cache is of $n\,K$ items. The set of items are denoted by $\mathcal{N}^{(n)} = \mathcal{N} \times \{1, 2, \ldots, n\}$ and the probability with which item (i, ℓ) ($\in \mathcal{N}^{(n)}$) is requested is p_i/n for all $\ell = 1, 2, \ldots, n$. For $i \in \mathcal{N}$, we refer to an item (i, ℓ), $\ell = 1, 2, \ldots, n$, as an item of class i. Note here that $n = 1$ represents the original (non-scaled) model and the ratio of the number of items and the capacity of cache remains the same as N/K for all $n \in \mathbb{N}$. We define 0-1-random variables $X_{i,\ell}^{(n)}(k)$ for $(i, \ell) \in \mathcal{N}^{(n)}$, $k \in \mathbb{Z}_+$, such that $X_{i,\ell}^{(n)}(k) = 1$ when item (i, ℓ) is in the cache at time k and $X_{i,\ell}^{(n)}(k) = 0$ otherwise. For each $i \in \mathcal{N}$, we also define

$$M_i^{(n)}(k) = \frac{1}{n} \sum_{\ell=1}^{n} X_{i,\ell}^{(n)}(k), \quad k \in \mathbb{Z}_+; \tag{5}$$

that is, $n\, M_i^{(n)}(k)$ represents the number of class i items in the cache at time k and it always holds that $\sum_{i=1}^{N} M_i^{(n)}(k) = K$. Then, $M^{(n)}(k) = (M_1^{(n)}(k), M_2^{(n)}(k), \ldots, M_N^{(n)}(k))$, $k \in \mathbb{Z}_+$, forms a homogeneous Markov chain within a finite state space $\Delta^{(n)} = \big\{(m_1, m_2, \ldots, m_N) \in \{0, 1/n, 2/n, \ldots, 1\}^N : \sum_{i=1}^{N} m_i = K\big\}$ and, when $|\mathcal{N}_+| \geq K$, the unique stationary distribution is directly obtained from (2) as

$$\Pi^{(n)}(m) = (C^{(n)})^{-1} \prod_{i=1}^{N} \binom{n}{n\, m_i} \Big(\frac{p_i}{n}\Big)^{n m_i}, \quad m = (m_1, m_2, \ldots, m_N) \in \Delta^{(n)}, \tag{6}$$

where $C^{(n)} = \sum_{m \in \Delta^{(n)}} \prod_{i=1}^{N} \binom{n}{n\, m_i} \big(p_i/n\big)^{n m_i}$.

We further scale the time and define a continuous-time process $\widehat{M}^{(n)}(t) = M^{(n)}(\lfloor n\, t \rfloor)$, $t \geq 0$, where $\lfloor x \rfloor = \max\{i \in \mathbb{Z} : i \leq x\}$ for $x \in \mathbb{R}$. We then have the following theorems, where $\Delta = \big\{(m_1, m_2, \ldots, m_N) \in [0, 1]^N : \sum_{i=1}^{N} m_i = K\big\}$ and $\|\cdot\|$ denotes the L_2-norm in $\mathbb{R}^N$.

Theorem 1 *Suppose that the initial state $M^{(n)}(0)$ converges in probability to a constant $m \in \Delta$ as $n \to \infty$. Then, for any $T \geq 0$ and any $\epsilon > 0$,*

$$\lim_{n \to \infty} \mathsf{P}\Big(\sup_{t \in [0,T]} \|\widehat{M}^{(n)}(t) - \mu(t)\| > \epsilon\Big) = 0, \tag{7}$$

where $\mu(t) = (\mu_1(t), \mu_2(t), \ldots, \mu_N(t))$, $t \geq 0$, is the solution to the system of differential equations;

$$\frac{\mathrm{d}\mu_i(t)}{\mathrm{d}t} = p_i\,(1 - \mu_i(t)) - \frac{\mu_i(t)}{K}\Big(1 - \sum_{j=1}^{N} p_j\, \mu_j(t)\Big), \quad i \in \mathcal{N},\ t \geq 0, \tag{8}$$

$$\mu(0) = m. \tag{9}$$

The proof of theorem relies on the result for mean field interaction models considered in [1] and is given in the next section. Theorem 1 says that, provided that the initial state $M^{(n)}(0)$ converges in probability to a constant $m \in \Delta$ as $n \to \infty$, then, for each finite $t \geq 0$, $\widehat{M}^{(n)}(t)$ converges in probability to the solution to the differential equation (8) with (9). We refer to this solution $(\mu(t))_{t\geq 0}$ as the fluid limit of $\big(M^{(n)}(k)\big)_{k\in\mathbb{Z}_+}$, $n \in \mathbb{N}$. The theorem, however, does not say the convergence of stationary distribution (6) of $M^{(n)}(k)$, $k \in \mathbb{Z}_+$, as $n \to \infty$ unless, for all initial points $m \in \Delta$, the trajectories of fluid limit converge to a unique stationary point as $t \to \infty$. This is in general hard to verify and we here apply another recent result in [10], which leads to the following.

Theorem 2 *Suppose $|\mathcal{N}_+| \geq K$ and that $M^{(n)}(0)$ follows the stationary distribution (6) for each $n \in \mathbb{N}$. Then, for any $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathsf{P}\big(\|M^{(n)}(0) - m^*\| > \epsilon\big) = 0, \tag{10}$$

where $m^* = (m_1^*, m_2^*, \ldots, m_N^*) \in \Delta$ is given by

$$m_i^* = \begin{cases} \dfrac{K\,p_i}{\rho^* + K\,p_i}, & i \in \mathcal{N}_+, \\ 0, & i \in \mathcal{N}_0, \end{cases} \tag{11}$$

and ρ^* is the unique solution in $[0, 1)$ to

$$\sum_{i\in\mathcal{N}_+} \frac{K\,p_i}{\rho^* + K\,p_i} = K, \tag{12}$$

and satisfies $\rho^* = 1 - \sum_{j=1}^N p_j\, m_j^*$.

This theorem is also proved in the next section. Theorem 2 says that, when $|\mathcal{N}_+| \geq K$, $M^{(n)}(0)$ in the steady state converges in probability to the constant $m^* \in \Delta$ given by (11) and (12) as $n \to \infty$; that is, the fluid limit of $M^{(n)}(0)$ in the steady state is m^*. In other words, the stationary distribution (6) converges weakly to the Dirac measure with mass at m^*.

Once Theorems 1 and 2 are provided, the fluid limits of fault probabilities are derived as corollaries of them. We first consider the transient case. Let $\rho^{(n)}(k)$, $k \in \mathbb{Z}_+$, $n \in \mathbb{N}$, denote the fault probability for the nth scaled model at time $k + 1$; that is, the probability with which an item, that is not in the cache at time k, is requested at time $k + 1$. Since the requested item at time $k + 1$ is independent of the cache contents at time k, we have

$$\begin{aligned} \rho^{(n)}(k) &= \sum_{i=1}^N \sum_{\ell=1}^n \frac{p_i}{n}\, \mathsf{P}(X_{i,\ell}^{(n)}(k) = 0) = \sum_{i=1}^N \sum_{\ell=1}^n \frac{p_i}{n} \left(1 - \mathsf{E}(X_{i,\ell}^{(n)}(k))\right) \\ &= 1 - \sum_{i=1}^N p_i\, \mathsf{E}\big(M_i^{(n)}(k)\big), \end{aligned} \tag{13}$$

where the last equality follows from (5). Let also $\widehat{\rho}^{(n)}(t) = \rho^{(n)}(\lfloor n\,t \rfloor)$ for $t \geq 0$. We then have the following.

Corollary 1 *Suppose that the initial state $M^{(n)}(0)$ converges in probability to a constant $m \in \Delta$ as $n \to \infty$. Then, for any $t \geq 0$,*

$$\lim_{n\to\infty} \widehat{\rho}^{(n)}(t) = 1 - \sum_{i=1}^N p_i\, \mu_i(t), \tag{14}$$

where $\mu(t) = (\mu_1(t), \mu_2(t), \ldots, \mu_N(t))$, $t \geq 0$, is the solution to (8) and (9) in Theorem 1.

Proof: Under the assumption that $M^{(n)}(0)$ converges in probability to $m \in \Delta$ as $n \to \infty$, Theorem 1 ensures that $\widehat{M}^{(n)}(t) = M^{(n)}(\lfloor n\,t \rfloor)$ converges in probability to $\mu(t)$ which is the solution to (8) and (9). Hence, (14) follows from (13) since $\widehat{M}_i^{(n)}(t)$ is bounded for $i \in \mathcal{N}$ and $t \geq 0$. □

In the case of steady state, we have the following.

Corollary 2 *Suppose $|\mathcal{N}_+| \geq K$ and that $M^{(n)}(0)$ follows the stationary distribution (6) for each $n \in \mathbb{N}$. Then, we have $\lim_{n\to\infty} \rho^{(n)}(0) = \rho^*$, which is the unique solution in $[0, 1)$ to (12) in Theorem 2.*

Proof: Since $M^{(n)}(0)$ converges in probability to m^* in (11) as $n \to \infty$ by Theorem 2 and $M^{(n)}(0)$ is bounded, we have from (13) that

$$\lim_{n\to\infty} \rho^{(n)}(0) = 1 - \sum_{i=1}^{N} p_i\, m_i^* = \rho^*,$$

where the last equality also follows from Theorem 2. □

Comparing (4) and (12), we find that the fluid limit ρ^* of stationary fault probability is coincident with the approximation of stationary fault probability for the FIFO cache provided in [3]. Since the stationary fault probabilities for the FIFO and RR caches are identical, we can say that our fluid limit analysis gives an exact theoretical foundation to Dan-Towsley's approximation.

4 Proofs of theorems

4.1 Proof of Theorem 1

To prove Theorem 1, we associate our model of RR caching with a mean field interaction model considered in [1]. Consider the nth scaled model; that is, the set of items is $\mathcal{N}^{(n)} = \mathcal{N} \times \{1, 2, \ldots, n\}$ and the cache buffer consists of $n\,K$ cells, each of which holds an item. At each time, an item, say $(i, \ell) \in \mathcal{N}^{(n)}$, is requested with probability p_i/n independently of requests at other times. We here regard the cells of cache as the objects in [1]. Namely, there are $n\,K$ objects and the state of an object is $i \in \mathcal{N}$ when the corresponding cell has an item of class i. This system of objects evolves stochastically as follows. If an item of class i' is requested at time $k+1$ and it is not in the cache at time k, then one of $n\,K$ cells is chosen uniformly at random and the item in the chosen cell is replaced with the requested one. In this case, if the randomly chosen cell has the item of class i at time k, then the corresponding object changes its state from i to i' (including the case of $i = i'$, in which case, the object does note change its state while a cache fault occurs). Let $Y_j^{(n)}(k)$ denote the state of object j ($\in \{1, 2, \ldots, n\,K\}$) at time k. Then, $Y^{(n)}(k) = \big(Y_1^{(n)}(k), Y_2^{(n)}(k), \ldots, Y_{nK}^{(n)}(k)\big)$, $k \in \mathbb{Z}_+$, forms a homogeneous Markov chain satisfying $\sum_{j=1}^{nK} \mathbf{1}_{\{Y_j^{(n)}(k)=i\}} \in \{0, 1, \ldots, n\}$ for each $i \in \mathcal{N}$ and $k \in \mathbb{Z}_+$. Furthermore, due to the RR algorithm, the state transition of $\big(Y^{(n)}(k)\big)_{k\in\mathbb{Z}_+}$ is invariant from the labeling of objects, so that, the process $\big(Y^{(n)}(k)\big)_{k\in\mathbb{Z}_+}$ is thought as a mean field interaction model without a resource (see [1] for more detail).

For the process $\big(Y^{(n)}(k)\big)_{k\in\mathbb{Z}_+}$, we have from (5),

$$\frac{1}{n}\sum_{j=1}^{nK} \mathbf{1}_{\{Y_j^{(n)}(k)=i\}} = \frac{1}{n}\sum_{\ell=1}^{n} X_{i,\ell}^{(n)}(k) = M_i^{(n)}(k), \quad i \in \mathcal{N},\ k \in \mathbb{Z}_+;$$

that is, we can see $M^{(n)}(k) = \big(M_1^{(n)}(k), M_2^{(n)}(k), \ldots, M_N^{(n)}(k)\big)$ as the empirical measure of $Y^{(n)}(k)$ on $\Delta^{(n)} = \big\{(m_1, m_2, \ldots, m_N) \in \{0, 1/n, 2/n, \ldots, 1\}^N : \sum_{i=1}^N m_i = K\big\}$ for each $k \in \mathbb{Z}_+$. To the mean field

interaction model above, we apply the result in [1], which is given as follows with translation into our notations.

Proposition 1 (Theorem 1 of [1]) *We suppose the following.*

(i) There exist a sequence $(\epsilon_n)_{n\in\mathbb{N}}$ *and a function* $f\colon \Delta \to \mathbb{R}^N$ *such that* $\lim_{n\to\infty} \epsilon_n = 0$ *and, for any* $m^{(n)} \in \Delta^{(n)}$, $n \in \mathbb{N}$, *and* $m \in \Delta$ *satisfying* $\lim_{n\to\infty} m^{(n)} = m$,

$$\lim_{n\to\infty} \frac{\mathsf{E}\big(M^{(n)}(k+1) \mid M^{(n)}(k) = m^{(n)}\big) - m^{(n)}}{\epsilon_n} = f(m).$$

(ii) For each $n \in \mathbb{N}$, *there exist a random sequence* $(W^{(n)}(k))_{k\in\mathbb{N}}$ *and a constant* $c > 0$ *such that*

$$\sum_{j=1}^{nK} \mathbf{1}_{\{Y_j^{(n)}(k+1)\neq Y_j^{(n)}(k)\}} \le W^{(n)}(k+1) \quad \textit{and} \quad \mathsf{E}\big(W^{(n)}(k)^2\big) \le c\, n^2\, {\epsilon_n}^2.$$

(iii) There exist a constant $\beta > 0$ *and a continuously differentiable function* $\varphi\colon \Delta \times [0,\beta] \to \mathbb{R}^N$ *such that, for any* $n \in \mathbb{N}$ *and any* $m^{(n)} \in \Delta^{(n)}$,

$$\frac{\mathsf{E}\big(M^{(n)}(k+1) \mid M^{(n)}(k) = m^{(n)}\big) - m^{(n)}}{\epsilon_n} = \varphi\Big(m^{(n)}, \frac{1}{n}\Big).$$

Also, let $\overline{M}^{(n)}(t)$, $t \ge 0$, *denote the linear interpolation of* $M^{(n)}(\lfloor t/\epsilon_n \rfloor)$; *that is,*

$$\overline{M}^{(n)}(t) = \big[M^{(n)}(\lfloor t/\epsilon_n \rfloor + 1) - M^{(n)}(\lfloor t/\epsilon_n \rfloor)\big]\,(t/\epsilon_n - \lfloor t/\epsilon_n \rfloor) + M^{(n)}(\lfloor t/\epsilon_n \rfloor), \quad t \ge 0.$$

Then, for all $T \ge 0$, *there exist constants* $C_{1,T}$, $C_{2,T}$ *and a random variable* $B_T^{(n)}$ *such that*

$$\sup_{t\in[0,T]} \|\overline{M}^{(n)}(t) - \mu(t)\| \le C_{1,T}\left(B_T^{(n)} + \|M^{(n)}(0) - \mu(0)\|\right) \quad \textit{with probability } 1,$$

$$\mathsf{E}\big({B_T^{(n)}}^2\big) \le C_{2,T}\,\epsilon_n,$$

where $\mu(t)$ *is the solution to*

$$\frac{\mathrm{d}\mu(t)}{\mathrm{d}t} = f(\mu(t)), \tag{15}$$

$$\mu(0) = m. \tag{16}$$

Remark 1 Conditions (i), (ii) and (iii) in Proposition 1 respectively correspond to **H2**, **H3** and **H5** in [1]. Since our model has no resource, which is supposed in [1], conditions **H1** and **H4** are irrelevant (that is, automatically satisfied here). Furthermore, the state space in [1] is $\Delta = \{(m_1, m_2, \ldots, m_N) \in [0,\infty)^N : \sum_{i=1}^N m_i = 1\}$; that is, the sum of elements is always unity. However, dividing our $M^{(n)}(k)$ by K and considering $n' = n\,K$ as the scaling parameter, we can easily reduce our model to that in [1].

Proof of Theorem 1: Before we verify that our model satisfies conditions (i)–(iii) in Proposition 1, we now suppose that with $\epsilon_n = 1/n$, $n \in \mathbb{N}$, and $f = (f_1, f_2, \ldots, f_N)$ such as

$$f_i(m) = p_i\,(1 - m_i) - \frac{m_i}{K}\Big(1 - \sum_{j=1}^N p_j\, m_j\Big), \quad i \in \mathcal{N},\ m \in \Delta. \tag{17}$$

Then, applying the proposition under the assumption that $M^{(n)}(0)$ converges in probability to $m \in \Delta$ as $n \to \infty$, we have for any $T \ge 0$ and any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathsf{P}\Big(\sup_{t\in[0,T]} \|\overline{M}^{(n)}(t) - \mu(t)\| > \epsilon\Big) = 0, \tag{18}$$

where $\mu(t)$, $t \geq 0$, is the solution to (15) and (16) with $f = (f_1, f_2, \ldots, f_N)$ given by (17); that is, the differential equation (15) with the initial condition (16) is coincident with (8) with (9). Furthermore, since we take $\epsilon_n = 1/n$, $n \in \mathbb{N}$, we have

$$\|\overline{M}^{(n)}(t) - \widehat{M}^{(n)}(t)\| \leq \|M^{(n)}(\lfloor n\,t \rfloor + 1) - M^{(n)}(\lfloor n\,t \rfloor)\| \leq \frac{\sqrt{2}}{n},$$

so that, with probability 1,

$$\lim_{n\to\infty} \sup_{t\in[0,T]} \|\overline{M}^{(n)}(t) - \widehat{M}^{(n)}(t)\| = 0,$$

which, together with (18), leads to (7).

From the discussion above, it remains to show that our model satisfies conditions (i)–(iii) in Proposition 1 with $\epsilon_n = 1/n$, $n \in \mathbb{N}$, and $f = (f_1, f_2, \ldots, f_N)$ given by (17). First, condition (ii) follows immediately since our model satisfies

$$\sum_{j=1}^{nK} \mathbf{1}_{\{Y_j^{(n)}(k+1) \neq Y_j^{(n)}(k)\}} \leq 1, \quad k \in \mathbb{Z}_+;$$

that is, at most one object (cell of the cache) changes its state at each time slot.

We next consider condition (iii). For $m^{(n)} = (m_1^{(n)}, m_2^{(n)}, \ldots, m_N^{(n)}) \in \Delta^{(n)}$, the next states from $m^{(n)}$ are expressed by $m^{(n)} + (e_{i'} - e_i)/n$, $i, i' \in \mathcal{N}$ such as $m_i^{(n)} \neq 0$ and $m_{i'}^{(n)} \neq 1$, where e_i, $i \in \mathcal{N}$, denotes the N-dimensional unit vector such as the ith element is equal to one and others are zero. Since an item of class i' is requested with probability $p_{i'}/n$ independently from the current state $m^{(n)}$, the probability with which one of class i' items outside the cache is requested is $p_{i'}/n \times (n - n\,m_{i'}^{(n)}) = p_{i'}\,(1 - m_{i'}^{(n)})$. On the other hand, since the item which is pushed out by the newly requested one is chosen with probability $1/(n\,K)$ among ones in the cache, the probability with which one of class i items in the cache is pushed out is $(n\,m_i^{(n)})/(n\,K) = m_i^{(n)}/K$. Namely, for $m^{(n)} \in \Delta^{(n)}$ and $k \in \mathbb{Z}_+$,

$$\mathsf{P}\Big(M^{(n)}(k+1) = m^{(n)} + \frac{e_{i'} - e_i}{n} \;\Big|\; M^{(n)}(k) = m^{(n)}\Big) = \begin{cases} \dfrac{m_i^{(n)}}{K}\, p_{i'}\,(1 - m_{i'}^{(n)}), & i \neq i' \\ \dfrac{1}{K}\displaystyle\sum_{i=1}^{N} p_i\, m_i^{(n)}\,(1 - m_i^{(n)}) + \sum_{i=1}^{N} p_i\, m_i^{(n)}, & i = i', \end{cases} \tag{19}$$

where, in the case of $i = i'$, the first term denotes the probability with which a cache fault occurs but the requested and pushed-out items are of the same class while the second term denotes the probability with which a cache fault does not occur, noting that $p_i\, m_i^{(n)} = p_i/n \times n\, m_i^{(n)}$ represents the probability with which one of class i items stored in the cache is requested. Thus, we have

$$\begin{aligned} \mathsf{E}\big(M^{(n)}(k+1) \mid M^{(n)}(k) = m^{(n)}\big) - m^{(n)} &= \sum_{i=1}^{N}\sum_{i'=1}^{N} \frac{e_{i'} - e_i}{n} \frac{m_i^{(n)}}{K}\, p_{i'}\,(1 - m_{i'}^{(n)}) \\ &= \frac{1}{n}\sum_{i'=1}^{N} e_{i'}\, p_{i'}\,(1 - m_{i'}^{(n)}) - \frac{1}{n\,K}\sum_{i=1}^{N} e_i\, m_i^{(n)}\Big(1 - \sum_{i'=1}^{N} p_{i'}\, m_{i'}^{(n)}\Big), \end{aligned}$$

where we use $\sum_{i=1}^{N} m_i^{(n)} = K$ in the second equality. The ith element of the above is given by

$$\mathsf{E}\big(M_i^{(n)}(k+1) \mid M^{(n)}(k) = m^{(n)}\big) - m_i^{(n)} = \frac{1}{n}\, p_i\,(1 - m_i^{(n)}) - \frac{m_i^{(n)}}{n\,K}\Big(1 - \sum_{j=1}^{N} p_j\, m_j^{(n)}\Big).$$

Therefore, applying $\epsilon_n = 1/n$, we have

$$\frac{\mathsf{E}\big(M_i^{(n)}(k+1) \mid M^{(n)}(k) = m^{(n)}\big) - m_i^{(n)}}{1/n} = p_i\,(1 - m_i^{(n)}) - \frac{m_i^{(n)}}{K}\Big(1 - \sum_{j=1}^{N} p_j\, m_j^{(n)}\Big), \tag{20}$$

so that, condition (iii) holds with $\varphi = (\varphi_1, \varphi_2, \ldots, \varphi_N)$ such as $\varphi_i(m, \alpha) = f_i(m)$ in (17) for any $\alpha \in [0, \beta]$ with an arbitrary $\beta > 0$. Finally, when $m^{(n)} \to m \in \Delta$ as $n \to \infty$, condition (i) follows from (20) with $f = (f_1, f_2, \ldots, f_N)$ given in (17). Hence, the proof is completed. □

Remark 2 By the discussion deriving (19), we have $p_j\, m_j^{(n)}$ as the conditional probability with which one of class j items stored in the cache is requested given the current state $m^{(n)} = (m_1^{(n)}, m_2^{(n)}, \ldots, m_N^{(n)})$; that is, the conditional fault probability for the nth scaled model given the current state $m^{(n)}$ is $1 - \sum_{j=1}^N p_j\, m_j^{(n)}$. Since the fluid limit of $M^{(n)}(k)$, $k \in \mathbb{Z}_+$, is deterministic function $\mu(t)$, $t \geq 0$, we see that (14) is also derived from this observation.

4.2 Proof of Theorem 2

To prove Theorem 2, we apply the recent result in [10], exploiting the reversibility of stochastic process to show the convergence of its stationary version to the stationary point of fluid limit, which is given as follows also with translation into our notations.

Proposition 2 (Corollary 1 of [10]) *We suppose the following.*

(i) $\big(\widehat{M}^{(n)}(t)\big)_{t\geq 0}$ *is reversible under a probability measure* $\widehat{\Pi}^{(n)}$ *on* Δ *such that* $\widehat{\Pi}^{(n)}(\Delta^{(n)}) = 1$ *in the sense that, for every* $t \geq 0$ *and any bounded and continuous function* h*:* $\Delta^2 \to \mathbb{R}$*,*

$$\int_\Delta \mathsf{E}\big(h(m, \widehat{M}^{(n)}(t)) \mid \widehat{M}^{(n)}(0) = m\big)\, \widehat{\Pi}^{(n)}(\mathrm{d}m) = \int_\Delta \mathsf{E}\big(h(\widehat{M}^{(n)}(t), m) \mid \widehat{M}^{(n)}(0) = m\big)\, \widehat{\Pi}^{(n)}(\mathrm{d}m). \tag{21}$$

(ii) The sequence $(\widehat{\Pi}^{(n)})_{n\in\mathbb{N}}$ *is tight.*

(iii) For any $m^{(n)} \in \Delta^{(n)}$*,* $n \in \mathbb{N}$*, and* $m \in \Delta$ *satisfying* $\lim_{n\to\infty} m^{(n)} = m$*, there exists a semi-flow* ϕ_t*:* $\Delta \to \Delta$*,* $t \geq 0$*, such that* $\phi_{s+t} = \phi_s \circ \phi_t$ *for* s*,* $t \geq 0$*,* $\phi_0(m) = m$ *and, for every* $t \geq 0$*, the conditional law of* $\widehat{M}^{(n)}(t)$ *given* $\widehat{M}^{(n)}(0) = m^{(n)}$ *converges weakly to the constant* $\phi_t(m)$ *as* $n \to \infty$*; that is, for all bounded and continuous function* h*:* $\Delta \to \mathbb{R}$*,*

$$\lim_{n\to\infty} \mathsf{E}\big(h(\widehat{M}^{(n)}(t)) \mid \widehat{M}^{(n)}(0) = m^{(n)}\big) = h(\phi_t(m)).$$

(iv) The semi-flow ϕ_t*,* $t \geq 0$*, has a unique stationary point* m^* *such that* $m^* = \phi_t(m^*)$ *for all* $t \geq 0$*.*

Then, the sequence $(\widehat{\Pi}^{(n)})_{n\in\mathbb{N}}$ *converges weakly to the Dirac measure with mass at* m^* *as* $n \to \infty$*.*

Remark 3 It should be noted, and is also noted in [10], that our state space Δ is a subset of $\mathbb{R}^N$, so that the semi-flow ϕ_t, $t \geq 0$, can be a differential equation of the form of (15). In this case, the stationary points of ϕ_t are the solutions to $f(m) = 0$.

Proof of Theorem 2: Since the state space Δ in our model is compact, condition (ii) in Proposition 2 necessarily holds. Also, given that $\widehat{M}^{(n)}(0) = M^{(n)}(0) = m^{(n)} \in \Delta^{(n)}$ and $m^{(n)} \to m \in \Delta$ as $n \to \infty$, Theorem 1 implies that $\widehat{M}^{(n)}(t)$ converges in probability to $\mu(t)$ given by (8) and (9) for any $t \geq 0$, so that, condition (iii) in Proposition 2 holds for $\phi_t(m) = \mu(t)$ with $\mu(0) = m$. Therefore, if we can verify that our model further satisfies conditions (i) and (iv) in Proposition 2 with $\widehat{\Pi}^{(n)}(\{m\}) = \Pi^{(n)}(m)$ in (6) for $m \in \Delta^{(n)}$ and with $m^* \in \Delta$ given by (11) and (12), then we can apply the proposition and obtain (10).

We first show that, when $|\mathcal{N}_+| \geq K$, condition (iv) holds in our model with $m^* \in \Delta$ given by (11) and (12). Since the semi-flow ϕ_t, $t \geq 0$, is given as the solution $\mu(t)$ to the differential equation

$\mathrm{d}\mu(t)/\mathrm{d}t = f(\mu(t))$, $t \geq 0$, its stationary points are the solutions to $f(m) = 0$; that is, it suffices to show that, when $|\mathcal{N}_+| \geq K$, the system of equations

$$p_i\,(1 - m_i) - \frac{m_i}{K}\Big(1 - \sum_{j=1}^{N} p_j\, m_j\Big) = 0, \quad i \in \mathcal{N}, \tag{22}$$

has a unique solution $m^* = (m_1^*, m_2^*, \ldots, m_N^*) \in \Delta$, which is given by (11) and (12). In order for (22) to hold for $i \in \mathcal{N}_0 = \{i \in \mathcal{N} : p_i = 0\}$, either of the following must be true for $m = (m_1, m_2, \ldots, m_N) \in \Delta$;

(a) $\sum_{i \in \mathcal{N}_0} m_i = 0$,

(b) $\sum_{i \in \mathcal{N}_0} m_i > 0$ and $\sum_{j=1}^{N} p_j\, m_j = 1$.

If we assume (b) above, (22) leads to $m_i = 1$ for $i \in \mathcal{N}_+ = \{i \in \mathcal{N} : p_i > 0\}$. Then, since $|\mathcal{N}_+| \geq K$ and $\sum_{i=1}^{N} m_i = K$ for $m \in \Delta$, we have $\sum_{i \in \mathcal{N}_0} m_i = K - \sum_{i \in \mathcal{N}_+} m_i = K - |\mathcal{N}_+| \leq 0$, which contradicts the assumption of (b). Therefore, (a) is true and (11) is valid for $i \in \mathcal{N}_0$. For $i \in \mathcal{N}_+$, arranging (22) with taking $\rho = 1 - \sum_{j=1}^{N} p_j\, m_j$, we have

$$m_i = \frac{K\, p_i}{\rho + K\, p_i}, \quad i \in \mathcal{N}_+. \tag{23}$$

Since $\sum_{i=1}^{N} m_i = K$ and $\sum_{i \in \mathcal{N}_0} m_i = 0$ from (a), (23) yields

$$\sum_{i \in \mathcal{N}_+} \frac{K\, p_i}{\rho + K\, p_i} = K.$$

Let us see the left-hand side above as a function $g(\rho)$ of ρ. Then, $g(\rho)$ is decreasing in $\rho \geq 0$ with $g(0) = |\mathcal{N}_+| \geq K$ and $g(1) = \sum_{i \in \mathcal{N}_+} [K\, p_i\,(1 + K\, p_i) - K^2\, {p_i}^2]/(1 + K\, p_i) = K - \sum_{i \in \mathcal{N}_+} K^2\, {p_i}^2/(1 + K\, p_i) < K$, so that, $g(\rho) = K$ has a unique solution on $[0, \infty)$ and this solution ρ^* lies in $[0, 1)$. Applying this ρ^*, we have m_i^* in (11) for $i \in \mathcal{N}_+$, which is uniquely determined by (23).

We next show that our model satisfies condition (i) in Proposition 2 with $\widehat{\Pi}^{(n)}(\{m\}) = \Pi^{(n)}(m)$ for $m \in \Delta^{(n)}$. To this end, we first show the reversibility of Markov chain $\big(M^{(n)}(k)\big)_{k \in \mathbb{Z}_+}$. For $i \neq i'$, by the stationary distribution (6) and the transition probability (19), we have

$$\Pi^{(n)}(m)\, \mathsf{P}\Big(M^{(n)}(1) = m + \frac{e_{i'} - e_i}{n} \;\Big|\; M^{(n)}(0) = m\Big) = (C^{(n)})^{-1} \prod_{j=1}^{N} \binom{n}{n\, m_j} \Big(\frac{p_j}{n}\Big)^{n m_j} \frac{m_i}{K}\, p_{i'}\,(1 - m_{i'}).$$

On the other hand, we have from (6),

$$\Pi^{(n)}\Big(m + \frac{e_{i'} - e_i}{n}\Big) = (C^{(n)})^{-1} \prod_{j=1}^{N} \binom{n}{n\, m_j} \Big(\frac{p_j}{n}\Big)^{n m_j} \frac{n\, m_i}{n - n\, m_i + 1} \frac{n - n\, m_{i'}}{n\, m_{i'} + 1} \Big(\frac{p_i}{n}\Big)^{-1} \Big(\frac{p_{i'}}{n}\Big),$$

and from (19),

$$\mathsf{P}\Big(M^{(n)}(1) = m \;\Big|\; M^{(n)}(0) = m + \frac{e_{i'} - e_i}{n}\Big) = \frac{p_i}{K}\Big(1 - m_i + \frac{1}{n}\Big)\Big(m_{i'} + \frac{1}{n}\Big).$$

Therefore, it holds that

$$\begin{aligned} &\Pi^{(n)}(m)\, \mathsf{P}\Big(M^{(n)}(1) = m + \frac{e_{i'} - e_i}{n} \;\Big|\; M^{(n)}(0) = m\Big) \\ &= \Pi^{(n)}\Big(m + \frac{e_{i'} - e_i}{n}\Big)\, \mathsf{P}\Big(M^{(n)}(1) = m \;\Big|\; M^{(n)}(0) = m + \frac{e_{i'} - e_i}{n}\Big). \end{aligned} \tag{24}$$

This is, of course, the case of $i = i'$ and the Markov chain $\big(M^{(n)}(k)\big)_{k\in\mathbb{Z}_+}$ is reversible. By the induction from (24), we have for any $k \in \mathbb{Z}_+$ and any $m, m' \in \Delta^{(n)}$,

$$\Pi^{(n)}(m)\,\mathsf{P}\big(M^{(n)}(k) = m' \mid M^{(n)}(0) = m\big) = \Pi^{(n)}(m')\,\mathsf{P}\big(M^{(n)}(k) = m \mid M^{(n)}(0) = m'\big).$$

Hence, taking $\widehat{\Pi}^{(n)}(\{m\}) = \Pi^{(n)}(m)$ for $m \in \Delta^{(n)}$, we have for any $t \ge 0$ and any bounded and continuous function $h\colon \Delta^2 \to \mathbb{R}$,

$$\begin{aligned}
&\int_\Delta \mathsf{E}\big(h(m, \widehat{M}^{(n)}(t)) \mid \widehat{M}^{(n)}(0) = m\big)\,\widehat{\Pi}^{(n)}(\mathrm{d}m)\\
&= \sum_{m,m'\in\Delta^{(n)}} h(m,m')\,\mathsf{P}\big(M^{(n)}(\lfloor n\,t\rfloor) = m' \mid M^{(n)}(0) = m\big)\,\Pi^{(n)}(m)\\
&= \sum_{m,m'\in\Delta^{(n)}} h(m,m')\,\mathsf{P}\big(M^{(n)}(\lfloor n\,t\rfloor) = m \mid M^{(n)}(0) = m'\big)\,\Pi^{(n)}(m')\\
&= \int_\Delta \mathsf{E}\big(h(\widehat{M}^{(n)}(t), m') \mid \widehat{M}^{(n)}(0) = m'\big)\,\widehat{\Pi}^{(n)}(\mathrm{d}m');
\end{aligned}$$

that is, (21) holds and the proof is completed. □

Remark 4 In order for (22) to have a unique solution in Δ, the condition $|\mathcal{N}_+| \ge K$ is necessary. Suppose $|\mathcal{N}_+| < K$ and consider case (a) in the proof of Theorem 2. Then, since $m_i \in [0,1]$, we have $\sum_{i=1}^N m_i = \sum_{i\in\mathcal{N}_+} m_i \le |\mathcal{N}_+| < K$, which contradicts $\sum_{i=1}^N m_i = K$ for $m \in \Delta$. Therefore, (b) is true in this case. From (22), it must be $m_i = 1$ for $i \in \mathcal{N}_+$ and there are uncountably many solutions m_i for $i \in \mathcal{N}_0$ satisfying $\sum_{i\in\mathcal{N}_0} m_i = K - \sum_{i\in\mathcal{N}_+} m_i = K - |\mathcal{N}_+| > 0$.

5 Concluding remarks

In this paper, to evaluate the performance of FIFO caching, we have instead studied the RR caching and have derived the fluid limit of transient behavior by applying the limit theorem for mean field interaction models in [1]. The fluid limit in the steady state has also been obtained by applying the result in [10]. As a result, we have seen that the fluid limit of stationary fault probability is coincident with the approximation of that for the FIFO caching provided in [3]. Though we have considered the independent reference model, we may extend some results to a dependent reference model by associating it with the mean field interaction model with a resource (see [1] and [2] for detail). To investigate more complex caching algorithms by the fluid limit approach, we may have to further develop the limit theory of mean field models with interactions.

Acknowledgements

The second author (RH)'s work was supported in part by JSPS (Japan Society for the Promotion of Science) Global COE (Centers of Excellence) Program "Computation as a Foundation for the Sciences." The third author (NM)'s work was supported by JSPS Grant-in-Aid for Scientific Research (C) 22510142.

References

[1] M. Benaïm and J.-Y. Le Boudec. A class of mean field interaction models for computer and communication systems. *Performance Evaluation*, 65 (2008) 823–838.

[2] C. Bordenave, D. McDonald and A. Proutière. A particle system in interaction with a rapidly varying environment: Mean field limits and applications. *Networks and Heterogeneous Media*, 5 (2010) 31–62.

[3] A. Dan and D. Towsley. An approximate analysis of the LRU and FIFO buffer replacement schemes. *ACM SIGMETRICS Performance Evaluation Review*, 18 (1990) 143–152.

[4] E. Gelenbe. A unified approach to the evaluation of a class of replacement algorithms. *IEEE Transactions on Computers*, 22 (1973) 611–618.

[5] K. Hattori and T. Hattori. Hydrodynamic limit of move-to-front rules and search cost probabilities. arXiv: 0908.3222v1 (2009).

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

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

[8] T. Johnson and D. Shasha. 2Q: A low overhead high performance buffer management replacement algorithm. *Proc. 20th International Conference on VLDB* (1994) 439–450.

[9] W. F. King. Analysis of paging algorithms. *Proc. IFIP Congress* (1971) 697–701.

[10] J.-Y. Le Boudec. The stationary behaviour of fluid limits of reversible processes is concentrated on stationary points. arXiv: 1009.5021v2 (2010).

[11] J. van den Berg and A. Gandolfi. LRU is better than FIFO under the independent reference model. *Journal of Applied Probability*, 29 (1992) 239–243.

[12] Y. Zhou, J. Phillbin and K. Li. The multi-queue replacement algorithm for second level buffer caches. *Proc. General Track: 2001 USENIX Annual Technical Conference* (2001) 91–104.