Analysis of large scale interacting systems by mean field method

Andrea Bobbio
University of Piemonte Orientale
Dipartimento di Informatica, Alessandria, Italy, Email: bobbio@mfn.unipmn.it

Marco Gribaudo
Università di Torino
Dipartimento di Informatica, Torino, Italy, Email: marcog@di.unito.it

Miklós Telek
Budapest University of Technology and Economics
Department of Telecommunications, Budapest, Hungary, Email: telek@hit.bme.hu

***Abstract*—Modeling and analysing very large stochastic systems composed of interacting entities is a very challenging and complex task. The usual approach, relying on the generation of the whole state space, is bounded by the state space explosion, even if symmetry properties, often included in the model, allow to apply lumping techniques and building the overall model by means of tensor algebra operations.**

In this paper we resort to the *mean field* theory. The main idea of the mean field theory is to focus on one particular tagged entity and to replace all interactions with the other entities with an average or effective interaction. The reduction of a multi-body problem into an effective one-body problem makes the solution easier while at the same time taking into account the contribution of an averaged interdependence of the whole system on the specific entity. We apply the mean field approach to very large systems of interacting continuous time Markov chains, in which the averaged interaction depends on the distribution of the entity population in each state.

We report several examples of interacting Markovian queues, showing the potentialities of the proposed technique.

Keywords: Continuous time Markov chain, Mean field method, Performance Evaluation.

I. Introduction

Complex systems can usually be disaggregated into interacting parts or components where each part can have a local autonomous behavior that depends on the ensemble of the behaviors of the other parts. In recent years, an enormous amount of literature has been devoted to the study of complex systems in biology, economics, social science, physics, computer and communication systems. In this paper, we focus the attention on very large scale stochastic systems, in which the basic entities evolve according to a CTMC, whose infinitesimal generator depends on current state occupied by all the other entities.

The analysis of large scale stochastic systems composed by interacting objects has been mainly faced in the literature by resorting to the superposition of interacting Markov chains or to fluid models. In the first case, the analysis of the system requires the generation of the global state space, defined as the Cartesian product of the state spaces of the CTMC's describing the individual interacting objects. The explosion of the global state space determines the upper bound for the application of the methodology, even if the explosion is usually mitigated by exploiting the symmetry properties often included in the system definition, that allow to apply lumping techniques and to produce the global transition rate matrix by means of tensor algebra operators applied to the local matrices.

Representative attempts in this direction define the interacting objects directly as Markov chains [5], [7], or as finite state automata [14], [15] or as Petri nets [6], [12]. In [14] the local entity is called automaton and the *Stochastic Automata Network (SAN)* is a system composed by interacting automata. In [1], the states of the individual Markov chains are partitioned in classes and the transition rate of each chain depends on the classes of the other chains. A two layer view is also proposed in social networks in [16] where the local level is a chain that depicts an individual player and the global view models the team action as a whole. The compositional approaches are limited by the explosion of the state space.

A particular model of interacting objects for which a set of exact and approximate analysis methods are available is the queueing network model. In this model the objects communicate via customers which visit the network nodes according to some routing rules. We refer to [3] for a recent survey on the related analysis results. In the most common application of queueing networks the number of objects is finite and the number of states of the objects can be finite and infinite. The case of infinite number of states of finite number of objects can also be approximated with fluid models. Fluid models [11], [8] are able to capture the global behaviour of the system, but they loose the capability of detailing the local behaviour. A continuous approximation to a discrete model is also considered in [9] where components of the same type do not have statistical dependencies but may synchronize on shared activities.

In this paper, we focus the attention on very large scale stochastic systems, whose dimensions exceed the capabilities of all the methods based on the generation of the global state space, even if the basic entities evolve according to a CTMC. Especially, we focus on the case when the number

of interacting objects grows very large and the number of states of these objects is finite and moderately large. We propose an approximation based on *mean field* method [13], [4]. The *mean field* method focusses on a particular tagged entity, and replaces all the interactions with the other entities with an average interaction. In the present case, each entity is a CTMC described by a local infinitesimal generator whose entries depend on the distribution of the other entities in their state space. In this way, we can model the individuality of each entity, but at the same time its interaction with the whole system. Asymptotic results allow us to consider systems in which the number of entities tends to infinity.

The mean field technique is well known and widely applied in many different areas [13]. The main goal of this paper is to present this methodology in a way which allows its use in the performance evaluation community. By this reason, we put more emphasis on how the methodology can find application in stochastic modeling rather than in the theoretical background of the methodology and the practical relevance of the considered examples. [4] and [2] had partially similar goals. The main difference is that in [4] and in [2] the interacting entities are formulated as discrete time Markov models, while in the present paper we take into consideration continuous time Markov models. In some cases the transition form continuous to discrete time Markov models are straight forward, but we believe that it is not immediate in case of the application of the mean field method.

Additionally, [4] and [2] apply a set of strong restrictions on the behavior of the interacting entities and the type of their interactions in order to apply a well established mathematical framework for proving the main convergence results. In this paper, we present examples which are out of the scope of [4] and [2] and still show a nice coincidence with the respective convergence results. These examples suggest that the conditions of [4] and [2] can be relaxed, but the investigation of the most general conditions and their proofs are out of the scope of the present paper.

The paper is organized as follows; Section II introduces the mean field idea for interacting CTMC and provides the main theorem and results. Section III illustrates a simple example of interacting queues and shows how different dependent strategies for accommodating the incoming customers can be modeled and analyzed; the analysis is restricted to identical and indistinguishable entities. Section IV introduces a new variant, by showing that it is possible to consider entities belonging to different types and provides a possible application example. Sections V and VI introduce memory dependencies in mean field analysis and an example of application of queues with memory dependent load, respectively.

II. Mean field method for large CTMC models

There are several efficient methods for constructing and evaluating Markovian models composed by a large finite number of identical interacting entities. The mean field method allows to compute the behaviour of this kind of models when the number of entities tends to infinity and suggests an approximation when the number of entities is large.

Let us assume that we have N identical discrete state entities in the form of CTMC. The state transitions of the CTMCs might depend on the current state of all entities, but cannot depend on the past history of the process and on the state transitions of other entities. This second restriction excludes synchronization between the transitions of entities.

The state of entity ℓ ($\ell = 1, 2, \ldots, N$) at time t is denoted by $X_\ell(t)$ In this section we assume, as an essential property, that all the entities are identical and indistinguishable. With this assumption, the behaviour of entity i does not depend directly on the particular state of a generic entity j, but it may depend on the global number of entities in each state.

Due to the fact that the entities are identical, the state of a randomly chosen (tagged) entity is denoted by $X(t)$. The state space of each entity, S, is composed by $s = |S|$ states, and $N_i(t)$ denotes the number of entities which are in state i ($\forall i \in S$) at time t. The vector composed by $N_i(t)$ is denoted by $\mathbf{N}(t)$ and by this definition, $\sum_{i=1}^{s} N_i(t) = N$.

The global behavior of the set of N entities forms a CTMC over the state space of size s^N. However, due to the fact that the entities are identical and indistinguishable, the state space can be lumped into the aggregate state space S_L of size $\binom{N+s-1}{s-1}$, where a state of the overall CTMC is identified by the number of entities staying in each state of S, i.e., by $\mathbf{N}(t) = (N_1(t), N_2(t), \ldots, N_s(t))$.

The evolution of the local CTMC is such that there are no synchronous transitions in different entities and the transition rates of a given entity may depend on the global behavior through the actual value of $\mathbf{N}(t)$. With this assumption, the following transition rates govern the evolution of a particular entity

$$\begin{aligned} &K_{ij}(\mathbf{N}(t)) = \\ &\begin{cases} \lim\limits_{\Delta\to 0} \dfrac{1}{\Delta} Pr(X(t+\Delta) = j | X(t) = i, \mathbf{N}(t)) & \text{if } N_i(t) > 0, \\ 0 & \text{if } N_i(t) = 0, \end{cases} \\ &K_{ii}(\mathbf{N}(t)) = -\sum_{j\in S, j\neq i} K_{ij}(\mathbf{N}(t)). \end{aligned} \tag{1}$$

Note that in the first condition of (1), $X(t) = i$ means that $\mathbf{N}(t)$ is such that $N_i(t) \geq 1$, since at least the tagged entity is in state i.

Instead of using $\mathbf{N}(t)$ when N is large, we introduce the normalized vector, $\mathbf{n}(t) = \mathbf{N}(t)/N$, where the entries $\mathbf{n}(t)$, $0 \leq n_i(t) \leq 1$, define the proportion of objects in state i at time t and $\sum_{i\in S} n_i(t) = 1$ and the associated transition rate function:

$$k_{ij}^{(N)}(\mathbf{n}(t)) = K_{ij}(N \cdot \mathbf{n}(t)). \tag{2}$$

(1) and (2) describe the same transition matrix, but (1) defines the transition rates at discrete points of integer valued vectors, $\mathbf{N}(t)$, and (2) defines them at discrete points whose coordinates are integer multiples of $1/N$. For the analysis of a system composed by N entities $K_{ij}(\bullet)$ and $k_{ij}^{(N)}(\bullet)$ are defined for these discrete points.

To investigate the limiting behavior as N tends to infinity we need a $k_{ij}(\mathbf{n}(t))$ function which is defined for all feasible $\mathbf{n}(t)$ vectors and satisfies $k_{ij}(\mathbf{n}(t)) = k_{ij}^{(N)}(\mathbf{n}(t))$ for $\forall N \geq 1$.

The existence and the properties of such $k_{ij}(\bullet)$ functions play an important role in the applicability of the mean field approach. In Section III-A we present a rather simple application example were $k_{ij}(\bullet)$ exists, but it is neither bounded nor continuous and the mean field limit seems to be valid. This suggests us that the practical application of the mean field approach for continuous time Markov chains requires more relaxed conditions than the ones in [4] and [2]. Obviously the problem of unbounded transition rate cannot occur with discrete time Markov chains, because in that case the transition probabilities are upper bounded by one. The discontinuity of $k_{ij}(\mathbf{n}(t))$ in the example of Section III-A is also related to the unboundedness of $k_{ij}(\mathbf{n}(t))$, because $k_{ij}(\mathbf{n}(t))$ is discontinuous at the limit where it tends to infinity. (The comments after (9) and (14) are to emphasize this behaviour through the studied examples.)

In the rest of the paper we use small letters to denote the quantities which are based on $\mathbf{n}(t)$. We define the transition matrix based on Equation (2) as

$$\mathbf{k}(\mathbf{n}(t)) = \{k_{ij}(\mathbf{n}(t))\} \tag{3}$$

The mean field method is based on the following essential theorem.

Theorem 1. *The normalized state vector of the lumped process, $\mathbf{n}(t)$, tends to be deterministic, in distribution, as N tends to infinity and satisfies the following differential equation*

$$\frac{d}{dt}\mathbf{n}(t) = \mathbf{n}(t)\ \mathbf{k}(\mathbf{n}(t)) \tag{4}$$

An individual component out of the set of s equations (4) can be written as:

$$\frac{d}{dt}n_i(t) = \sum_{j\in S} n_j(t)\ k_{ji}(\mathbf{n}(t)), \tag{5}$$

The proof of the theorem requires further investigation on the conditions $\mathbf{k}(\mathbf{n}(t))$ has to fulfill. In [10] we have provided an initial derivation that (at the moment of writing) is currently being refined with respect to these condition. In [4] and [2] very strong conditions where assumed. However in this work we show that the mean-field method can be successfully applied also to cases that do not fulfill such strict requirements.

Theorem 1 provides a formulation that is easy to apply and to compute in the extreme case when N tends to infinity, but in practice it is typically not the case. The following corollary provides an approximation method for the case when the number of entities, N, is finite but sufficiently large.

Corollary 2. *When N is sufficiently large, the normalized state vector of the lumped process, $\mathbf{n}(t)$, is a random vector whose mean can be approximated by the following differential equation*

$$\frac{d}{dt}E(n_i(t)) \approx \sum_{j\in S} E(n_j(t))\ k_{ji}(E(\mathbf{n}(t))) \tag{6}$$

Also for the proof of the Corollary we refer to [10] which is under refinement. While Theorem 1 is exact because $\mathbf{n}(t)$ is deterministic Corollary 2 is approximate, whose accuracy depends on the distribution of $\mathbf{n}(t)$. The closest $\mathbf{n}(t)$ is to deterministic, the better is the approximation of Corollary 2. The speed of the convergence of $\mathbf{n}(t)$ to deterministic as N tends to infinity is also investigated in the subsequent numerical examples.

III. Mean field analysis of dependent queues

To demonstrate the mean field methodology we present a simple example and detail its analysis according to the concepts and quantities discussed in the previous section.

Let us consider a queueing system composed by N identical Markovian queues (entities), each of which has a single server and a buffer of size 1 ($S = \{0, 1, 2\}, s = 3$). Customers arrive at rate $N\lambda$ to this queueing system and their service time is exponentially distributed with parameter μ. The CTMC of a single queue ($N = 1$) is depicted in Figure 1.

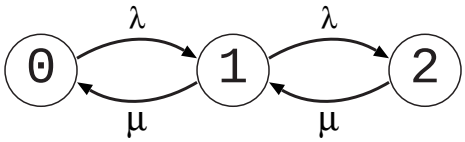

Fig. 1. Markov chain of a single queue in isolation

A. The incoming customer chooses the shortest queue

When $N > 1$ we adopt a policy that the incoming customer chooses the shortest queue and is directed to the queue which has the least number of customers in it. This policy makes the different queues interdependent. When $N = 2$, the CTMC describing this behavior is depicted in Figure 2, where the first number refers to the state of queue 1 and the second to the state of queue 2. We can interpret the transitions of Figure 2 from the view point of queue 1. In this case, the transition rates of the arrivals to queue 1 depend on the state of queue 2 as depicted in Figure 3.

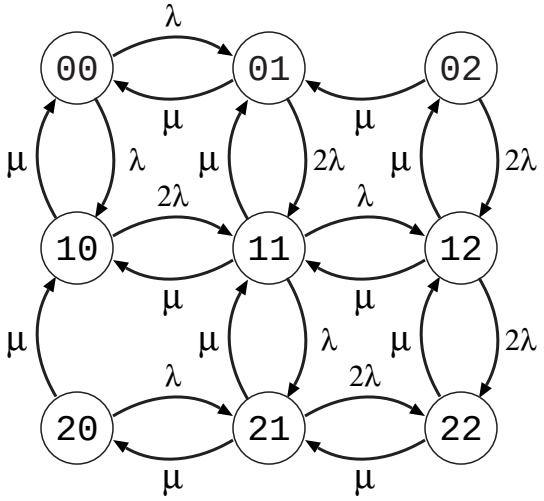

Fig. 2. Markov chain of 2 queues (without lumping)

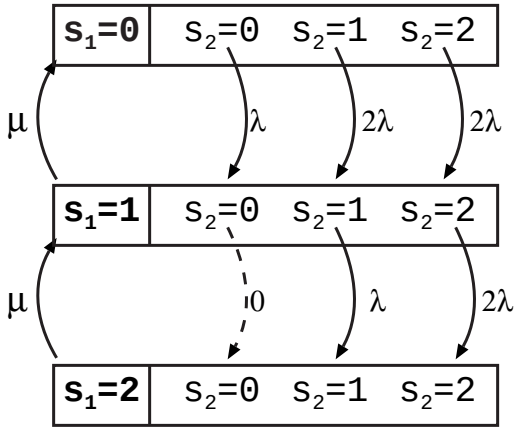

Fig. 3. Dependence of the transitions of queue 1 on the state of queue 2

Since the queues are identical, we can lump the states according to Figure 4 and we obtain the CTMC depicted in Figure 5. The lumped state space is composed by $\binom{2+3-1}{3-1} = 6$ states, $\mathbf{N}(t) \in \{(2,0,0),(1,1,0),(0,2,0),(1,0,1),(0,0,2),(0,1,1)\}$, where the states are identified by the number of queues having a given number of customers in it. E.g., state $(1,1,0)$ means that one of the queues is idle and one of them has 1 customer in it.

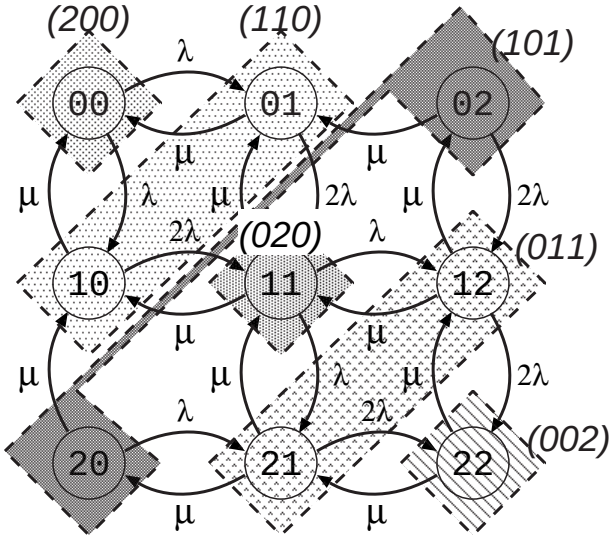

Fig. 4. Lumping the Markov chain of 2 queues

Considering the lumped process, we can interpret the behavior from the point of view of a tagged queue. In this case the arrival rates depend on the states of the lumped CTMC, as it is in Figure 6

Due to the fact that all queues are identical Figure 6 contains all information about the process. Consequently, Figure 6 and Figure 5 give an equivalent description of the process. To keep the system description compact (and independent of N) the description of a single tagged entity (the one in Figure 6) is used in practice.

For example, our queueing system can be described as the

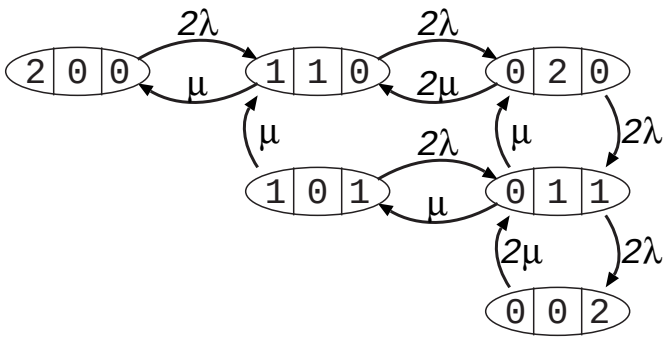

Fig. 5. Markov chain of the overall behaviour

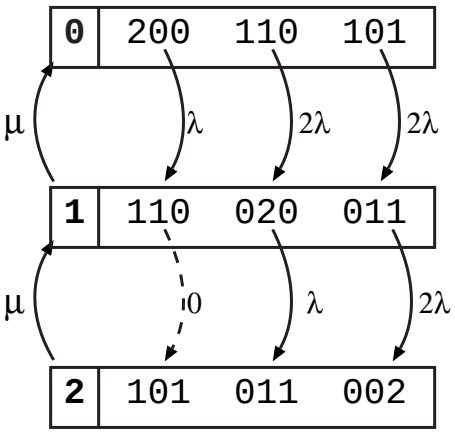

Fig. 6. Dependence of the transitions of the tagged queue on the lumped state

Markov chain in Figure 7, where

$$\Lambda_0(\mathbf{N}(t)) = \begin{cases} \lambda & \text{if } \mathbf{N}(t) = (2,0,0), \\ 2\lambda & \text{if } \mathbf{N}(t) = (1,1,0), \\ 2\lambda & \text{if } \mathbf{N}(t) = (1,0,1). \end{cases}$$

$$\Lambda_1(\mathbf{N}(t)) = \begin{cases} 0 & \text{if } \mathbf{N}(t) = (1,1,0), \\ \lambda & \text{if } \mathbf{N}(t) = (0,2,0), \\ 2\lambda & \text{if } \mathbf{N}(t) = (0,1,1). \end{cases}$$

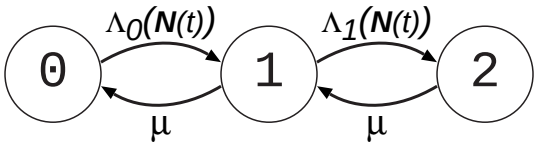

Fig. 7. Markov chain of one of the identical entities

Having this compact system description, the only remaining step is to introduce the normalized occupancy vector $\mathbf{n}(t) = \mathbf{N}(t)/N$. Doing this, we get

$$\mathbf{n}(t) \in \{(1,0,0),(0.5,0.5,0),(0,1,0), (0.5,0,0.5),(0,0,1),(0,0.5,0.5)\}$$

and

$$\lambda_0(\mathbf{n}(t)) = \begin{cases} \lambda & \text{if } \mathbf{n}(t) = (1,0,0), \\ 2\lambda & \text{if } \mathbf{n}(t) = (0.5,0.5,0), \\ 2\lambda & \text{if } \mathbf{n}(t) = (0.5,0,0.5). \end{cases}$$

$$\lambda_1(\mathbf{n}(t)) = \begin{cases} 0 & \text{if } \mathbf{n}(t) = (0.5,0.5,0), \\ \lambda & \text{if } \mathbf{n}(t) = (0,1,0), \\ 2\lambda & \text{if } \mathbf{n}(t) = (0,0.5,0.5). \end{cases}$$

To make the system description independent of N we can rewrite the transition rates as

$$\lambda_0(\mathbf{n}(t)) = \begin{cases} \dfrac{\lambda}{n_0(t)} & \text{if } n_0(t) \neq 0, \\ 0 & \text{if } n_0(t) = 0, \end{cases}$$

$$\lambda_1(\mathbf{n}(t)) = \begin{cases} \dfrac{\lambda}{n_1(t)} & \text{if } n_0(t) = 0 \text{ and } n_1(t) > 0, \\ 0 & \text{if } n_0(t) > 0 \text{ or } n_1(t) = 0, \end{cases} \tag{7}$$

and obtain the CTMC shown in Figure 8. Note that the transitions with fixed rate are the transitions which are independent of the state of the other entities, while the transitions that are function of the occupancy vector represent the dependency of the entities, that is the action of the mean field.

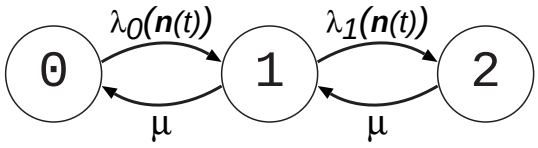

Fig. 8. Markov chain of one of the identical entities

The N independent description of the system in (7) is the key to evaluate the limiting behaviour when N tends to ∞. In this case, the particular form of (5) is

$$\begin{aligned} &\frac{d}{dt}\{n_0(t), n_1(t), n_2(t)\} \\ &= \{n_0(t), n_1(t), n_2(t)\} \\ &\quad \begin{bmatrix} -\lambda_0(\mathbf{n}(t)) & \lambda_0(\mathbf{n}(t)) & 0 \\ \mu & -\mu - \lambda_1(\mathbf{n}(t)) & \lambda_1(\mathbf{n}(t)) \\ 0 & \mu & -\mu \end{bmatrix} \end{aligned} \tag{8}$$

and starting from $\mathbf{n}(0) = \{1,0,0\}$ the transient behaviour of $\mathbf{n}(t)$ can be computed using numerical methods. The limiting behaviour when $t \to \infty$ can also be obtained as the limit of the transient results

$$\lim_{t\to\infty} \mathbf{n}(t) = \begin{cases} \{1 - \frac{\lambda}{\mu}, \frac{\lambda}{\mu}, 0\} & \text{if } \frac{\lambda}{\mu} < 1, \\ \{0,0,1\} & \text{if } 1 \leq \frac{\lambda}{\mu}, \end{cases} \tag{9}$$

which agree with our intuitive understanding on the model behaviour.

Note that $\lambda_0(\mathbf{n}(t))$ is always multiplied by $n_0(t)$ in the rhs of (8). This is why the unboundedness of $\lambda_0(\mathbf{n}(t))$ does not cause problem for the computation of (8). In this example the $n_0(t)\lambda_0(\mathbf{n}(t))$ product is not continuous at $n_0(t) = 0$. It equals to λ when $n_0(t) > 0$ and to 0 when $n_0(t) = 0$. The definition of $\lambda_0(\mathbf{n}(t))$ in (7) provides the required discontinuity of the $n_0(t)\lambda_0(\mathbf{n}(t))$ product at $n_0(t) = 0$. The same situation occurs with $\lambda_1(\mathbf{n}(t))$ and $n_1(t)$.

Figure 9 shows how the limit proposed in equation (9) actually holds, by plotting the mean number of entities in each state as a function of N, where $\lambda = 1.5$ and $\mu = 2$. The results of the figure are computed by numerically solving the stationary distribution of the lumped CTMC of N objects.

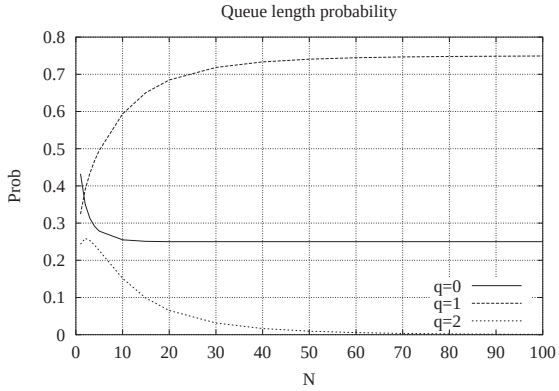

Fig. 9. Queue length probability as function of N

B. Convergence when N tends to infinity

In order to demonstrate the convergence to a deterministic quantity of the occupancy vector $\mathbf{n}(t)$ when N tends to infinity, we have considered a system of N entities where each entity is limited to 2 states (that is, each queue can only be empty or in service).

In this way the complete stationary occupancy vector, $\{n_0, n_1\}$, can be univocally defined by a single random number, n_0. The results of the exact computations over the complete system are reported in Figures 10 and 11, when $\lambda = 1.5$ and $\mu = 2$.

Figure 10 shows how the coefficient of variation of the number of queues in the first state tends to zero as N tends to infinity in both linear and log-log scale. The log-log plot makes it evident that the decreasing behaviour of the coefficient of variation has a slope proportional to $\sqrt{N}$ as stated by the strong law of large numbers. Figure 11 plots instead the whole distribution of the number of entities in the first state, and shows how it tends to become deterministic.

C. The incoming customer chooses the shortest of K queues

Other management policies, introducing different and more complex dependencies among the queues, are also easy to model and analyze with the mean field method. A variant of the previous example is when the new incoming customer randomly selects K queues and it joins the one with the less customers out of the selected K queues. This variant represents the random queue selection (independent case) when $K = 1$ and it represents the shortest queue selection of Section III-A when $K = N$. The subsequent analysis assumes a fixed K independent of N.

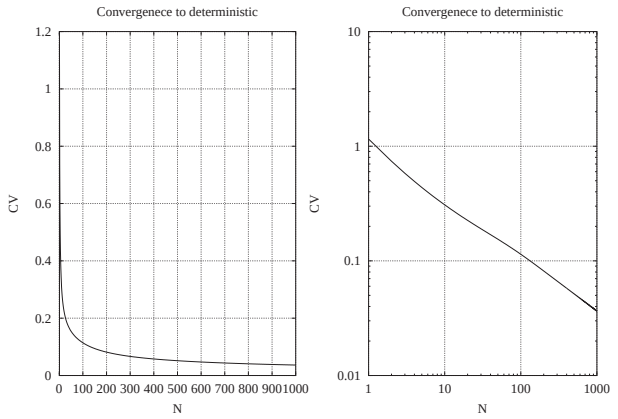

Fig. 10. Cv of the exact system as function of N (linear and logarithmic scale)

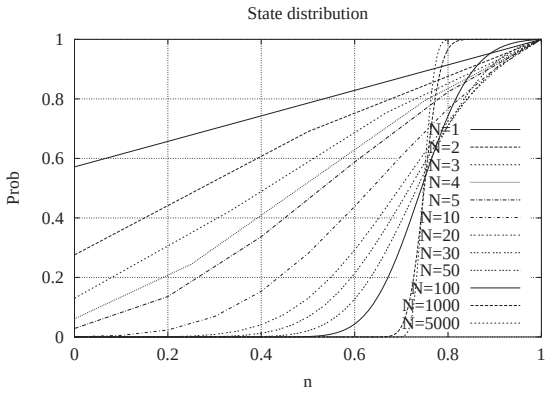

Fig. 11. Distribution of the exact system as function of N

According to this policy, the probability that an arriving customer attends a queue with i customers in it can be computed as follows

$$\begin{aligned} &Pr(\text{new customer goes to queue of length } i) = \\ &Pr(K \text{ selected queues are longer than } i-1 \\ &\quad \text{and at least one selected queue has length } i) = \\ &Pr(K \text{ selected queues are longer than } i-1) - \\ &\quad Pr(K \text{ selected queues are longer than } i) \end{aligned} \tag{10}$$

To compute these probabilities we introduce the following notation. The number of queues with at least i customers in it is $S_i(t) = \sum_{j=i}^{s} N_j(t)$. The proportion of queues with at least i customers in it is $s_i(t) = S_i(t)/N = \sum_{j=i}^{s} n_j(t)$. Using these notations

$$\begin{aligned} &Pr(K \text{ selected queues are longer than } i-1) = \\ &\frac{S_i(t)}{N} \cdot \frac{S_i(t)-1}{N-1} \cdots \frac{S_i(t)-K+1}{N-K+1} = \frac{\binom{S_i(t)}{K}}{\binom{N}{K}} . \end{aligned} \tag{11}$$

When N tends to infinity we have

$$\begin{aligned} &\lim_{N\to\infty} Pr(K \text{ selected queues are longer than } i-1) \\ &\quad = s_i(t)^K . \end{aligned} \tag{12}$$

Based on (11), the overall arrival rate towards the queues of length i is $\lambda N \dfrac{\binom{S_i(t)}{K} - \binom{S_{i+1}(t)}{K}}{\binom{N}{K}}$ and the arrival rate to one of the queues of length i is

$$\Lambda_i(\mathbf{N}(t)) = \frac{\lambda N}{N_i(t)} \cdot \frac{\binom{S_i(t)}{K} - \binom{S_{i+1}(t)}{K}}{\binom{N}{K}} . \tag{13}$$

Similarly when N tends to infinity, from (12), we have

$$\lambda_i(\mathbf{n}(t)) = \frac{\lambda}{n_i(t)} \left(s_i(t)^K - s_{i+1}(t)^K \right) . \tag{14}$$

Note that, in this case $\lambda_i(\mathbf{n}(t))$ is finite as $n_i(t) \to 0$ because $(s_i(t)^K - s_{i+1}(t)^K)$ contains an $n_i(t)$ factor. Due to this finite limit the $n_i(t)\lambda_i(\mathbf{n}(t))$ product, which appears on the rhs of (5) vanishes as $n_i(t) \to 0$. Consequently, in this case it is indifferent if $\lambda_i(\mathbf{n}(t))$ is set to 0 at $n_i(t) = 0$ (in which case $\lambda_i(\bullet)$ is not continuous) or it is defined to be continuous at $n_i(t) = 0$.

We have implemented the mean field analysis of the above detailed queue selection policy when each queue has at most 3 customers and we have evaluated the system behaviour in two cases:

Case i) - light load, $\rho = \lambda/\mu = 0.5$ ($\lambda = 1, \mu = 2$)
Case ii) - heavy load, $\rho = \lambda/\mu = 2$ ($\lambda = 2, \mu = 1$).

As a result of the mean field analysis, i.e., numerical solution of (4) using Runge-Kutta or Euler elementary steps refined to avoid negative probabilities, we have depicted in Figures 12 and 13 the mean queue length for the light and heavy load, respectively.

We observe different trends in the light and the heavy loaded cases. Under light load (Figure 12), the selection of the shortest queue ($K = N$) means that half of the queues have 1 customer and half of them are idle. Instead, with a random queue selection ($K = 1$) the probability of having some queues with 2 customers is positive and the mean queue length is higher.

In case of heavy load (Figure 13), the selection of the shortest queue ($K = N$) means that all the queues are going to be saturated (i.e., in state 2) with probability 1. Instead in case of random queue selection ($K = 1$) the probability that a significant portion of the queues is not selected for a long time is so high that the probability of having less than 2 customers in a significant portion of the queues is positive. As a result the mean queue length is less in this case.

Another important performance metric that can be computed from the model is the mean loss probability of an incoming customer, that is the probability that a client is routed to a queue which is already full and cannot hold it. This metric is

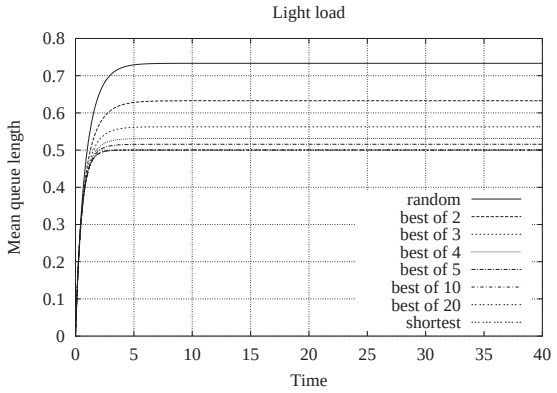

Fig. 12. System behaviour with light load

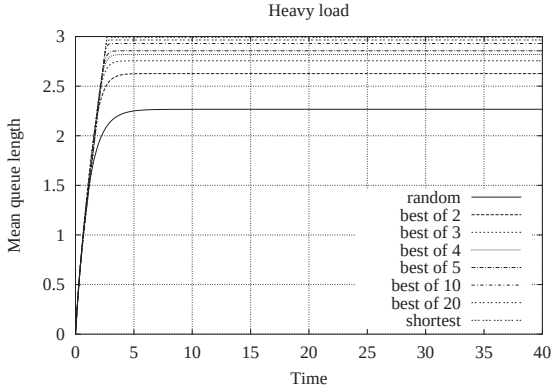

Fig. 13. System behaviour with heavy load

computed as follows:

$$Pr(loss) = \frac{E(\#\text{ incoming customers}) \; - \; E(\#\text{ served customers})}{E(\#\text{ incoming customers})} = \lim_{t\to\infty} \frac{N\lambda - (N - N_0(t))\mu}{N\lambda} = \lim_{t\to\infty} \frac{\lambda - (1 - n_0(t))\mu}{\lambda} \tag{15}$$

Figure 14 shows this quantity for both the light and the heavy loaded cases as a function of K. In both cases, increasing K reduces the loss probability of the system. When the system is light loaded (i.e. $\lambda < \mu$), it is sufficient to have $K \geq 4$ to obtain loss probabilities smaller than the machine precision. For the heavy loaded case (i.e. $\lambda > \mu$), the probability does not tend to 0, but to $\frac{\lambda-\mu}{\lambda}$. In order to better understand how the loss probability reaches this limit, in the logarithmic version of Figure 14, we have plotted $Pr(loss) - \frac{\lambda-\mu}{\lambda}$. As it can be seen, when the system is heavily loaded, K must be increased much more than in the lightly loaded case to reduce the losses.

D. Comparison with finite N systems

In practice systems are finite, and the assumption of N that tends to the infinity is often non-realistic. One of the questions is thus whether the mean field approach is appropriate to approximate finite systems, and up to which extent. In Corollary 2 we have already proven that this approximation is valid for $N \to \infty$. In Figure 15, we elaborate on this problem, focusing on the loss probability of the heavy loaded model of Section III-C, and comparing the loss performance index on a finite system with N queues, using discrete event simulation, with the results obtained from the mean field analysis with $N \to \infty$. The solid line in Figure 15 refers to the mean field computation with $N \to \infty$ and $K = 3$ and the dotted line to the mean field analysis with $K = 2$. It can be seen, that the 95% confidence intervals computed by simulation with $N = 2;\ K = 2$ and with $N = 3;\ K = 3$ are well centered on the asymptotic mean field results. Moreover, for $N \geq 10$ the simulated confidence intervals of the loss probability shrink around the asymptotic mean field values.

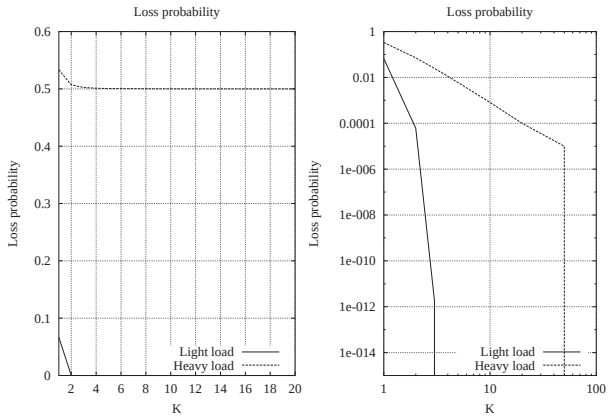

Fig. 14. Loss probability (linear and logarithmic scale)

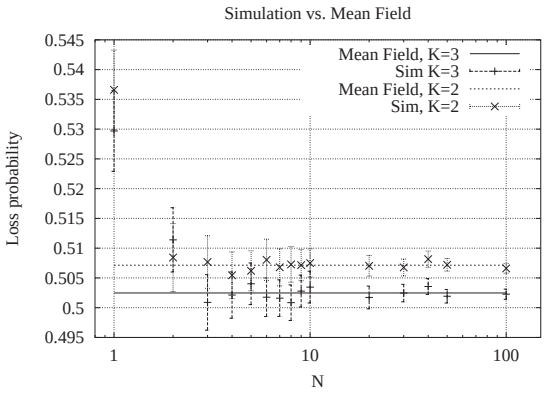

Fig. 15. Comparison of the loss probability between mean field, and simulation of a finite system with N queues

However the quality of the approximation depends also very much on the load of the system. In particular, one of our reviewers recommended to check it at a system load close to 1. It turned out that the convergence to the asymptotic results is slower in this case. For example Figure 16 shows the comparison of loss probability for a system with $K = 2$, with $\rho = 0.99$ and $\rho = 1.01$ for increasing values of N. In this case, in order to have an accurate approximation, we must have $N > 1000$. Figure 17 justifies this conclusion with respect to the mean queue length.

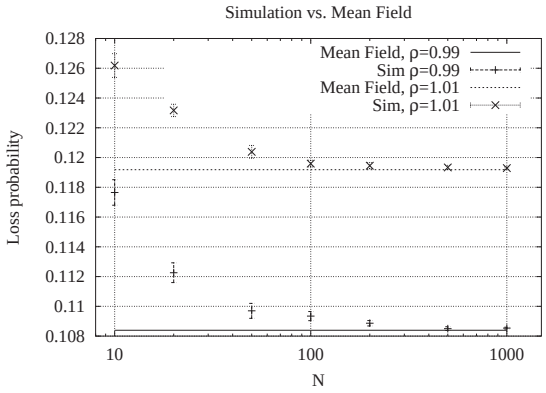

Fig. 16. Comparison of the loss probability between mean field, and simulation of a finite system with N queues, when the total load $\rho \to 1$

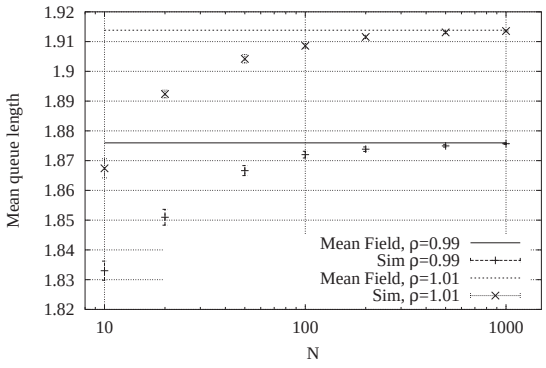

Fig. 17. Comparison of the mean queue length between mean field, and simulation of a finite system with N queues, when the total load $\rho \to 1$

The above results indicate that the solutions computed via mean field analysis, can be considered as a meaningful approximation of the exact performance indices, in some cases, even for relatively small (i.e., $N = 10$) numbers of entities in the system, but in some cases accurate approximation is obtained only at around $N = 100 - 500$.

IV. Mean field method with different kinds of entities

In section II, we considered the mean field analysis of N dependent identical Markovian entities. In this section we extend the analysis to systems composed by more than one type of dependent Markovian entities. Let $N^{(1)}, N^{(2)}, \ldots, N^{(C)}$ be the number of identical entities of type $1, 2, \ldots, C$, respectively. The state space of a type c entity is denoted by $S^{(c)}$ ($c \in \{1, 2, \ldots, C\}$), and is composed by $s^{(c)} = |S^{(c)}|$ states. $N_i^{(c)}(t)$ denotes the number of type c entities which are in state i at time t. We introduce vector $\mathbf{N}^{(c)}(t)$ of size $s^{(c)}$, whose elements are $N_i^{(c)}(t)$ and vector $\mathbf{N}(t)$ of size $s = \sum_{c=1}^{C} s^{(c)}$, whose blocks are $\mathbf{N}^{(c)}(t)$.

In general, the transition matrix of the Markov chain of a type c entity may depend on the whole vector $\mathbf{N}(t)$. The transition rates of a type c entity are

$$K_{ij}^{(c)}(\mathbf{N}(t)) = \lim_{\Delta \to 0} \frac{1}{\Delta} Pr(X^{(c)}(t+\Delta) = j | X^{(c)}(t) = i, \mathbf{N}(t)),$$
$$K_{ii}^{(c)}(\mathbf{N}(t)) = -\sum_{j \in S, j \neq i} K_{ij}^{(c)}(\mathbf{N}(t)).$$

Introducing again, $\mathbf{n}(t) = \mathbf{N}(t)/N$, assuming that $\lim_{N\to\infty} N^{(c)}/N = n^{(c)} \in (0,1)$ and taking the limit $N \to \infty$ we obtain the same differential equation as (5), but this time vector $\mathbf{n}(t)$, contains the proportion of entities of each type in each state.

As a result, the cardinality of the differential equation (5) increases linearly with the number of different types, since matrix $\mathbf{k}(\mathbf{n}(t))$ is of size $s \times s$ (with $s = \sum_{c=1}^{C} s^{(c)}$) and is composed by non-zero blocks of size $s^{(c)} \times s^{(c)}$.

A. Example of a system with regular and spare queues

Let us consider the queueing system of Section III with two types of queues: regular queues and *spare* queues. For each regular queue, there are γ spares. For example $\gamma = 0.5$ means that there is a spare every 2 regular queues. Both regular and spare queues have the same service rate μ, and the same buffer capacity B. Customers arrive at rate λ per regular queue, and are directed to the queue with the lowest occupancy. Spare queues are used only if the mean number of customers in the regular queues exceeds a given threshold β. We call $\alpha = \frac{1}{1+\gamma}$ the fraction of regular queues. Since the arrival rate is expressed *per regular queue*, we compute the total arrival rate as $\tilde{\lambda} = \alpha\lambda$.

We can apply mean field analysis to this system considering regular queues (identified by vector $\mathbf{n}^{(R)}$), and spare queues (identified by vector $\mathbf{n}^{(S)}$) separately. In particular, if we consider $B = 1$ for sake of simplicity, we have:

$$\begin{aligned} \mathbf{n}(t) &= \{n_0^{(R)}(t), n_1^{(R)}(t), n_0^{(S)}(t), n_1^{(S)}(t)\} \\ \mathbf{n}(0) &= \{\alpha, 0, 1-\alpha, 0\} \end{aligned}$$

$$\mathbf{k}(\mathbf{n}(t)) = \begin{bmatrix} -\lambda_R(\mathbf{n}(t)) & \lambda_R(\mathbf{n}(t)) & 0 & 0 \\ \mu & -\mu & 0 & 0 \\ 0 & 0 & -\lambda_S(\mathbf{n}(t)) & \lambda_S(\mathbf{n}(t)) \\ 0 & 0 & \mu & -\mu \end{bmatrix}$$

where:

$$\lambda_R = \begin{cases} \dfrac{\tilde{\lambda}}{n_0^{(R)}(t)} & \text{if } n_1^{(R)}(t) \leq \alpha\beta \\ \dfrac{\tilde{\lambda}}{n_0^{(R)}(t) + n_0^{(S)}(t)} & \text{if } n_1^{(R)}(t) > \alpha\beta \end{cases}$$

$$\lambda_S = \begin{cases} 0 & \text{if } n_1^{(R)}(t) \leq \alpha\beta \\ \dfrac{\tilde{\lambda}}{n_0^{(R)}(t) + n_0^{(S)}(t)} & \text{if } n_1^{(R)}(t) > \alpha\beta \end{cases}$$

Figure 18 shows some results for $\gamma = 0.5$, varying both the load of the system $\rho = \frac{\lambda}{\mu}$, and the switching point β. The

introduction of the spare queues allows the system to respond to load greater than 1 (up to $\rho = 1+\gamma$). Spare queues are used only if the total load produces a mean queue length larger than β. After this threshold, the mean queue length of the regular queues remains constant, until it is reached by the mean length of the spare queues. From this point on both regular and spare queues grow with the same slope. If we consider a fixed load $\rho < 1$, we can see that large values of β reduce spare queues usage, while small values of β improve the response time by allowing shorter queues.

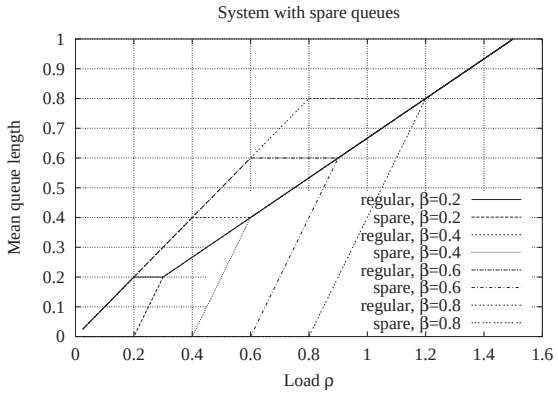

Fig. 18. Mean queue length in a system with 50% spare queues

V. Mean field method with memory dependency

The mean field framework allows to consider the dependency of the individual entities on a set of reward variables. The reward variables represent a kind of memory on the past evolution of the process, hence the reward variables are interchangeably referred to as memory variables.

We introduce vector $\mathbf{M}(t)$, whose d-th element $M_d(t)$ ($d \in \{1, 2, \ldots, D\}$) denotes the accumulated reward of type d at time t. When an entity is in state i it accumulates reward of type d at rate $r_i^{(d)}$ and the overall accumulation rate of type d reward (generated by all entities) is

$$\frac{d}{dt}\ M_d(t) = \sum_{i \in S} r_i^{(d)}\ N_i(t)\ .$$

In general, as N tends to infinity $M_d(t)$ tends to infinity as well. By this reason we introduce $m_d(t) = M_d(t)/N$ and the associated vector $\mathbf{m}(t) = \mathbf{M}(t)/N$. Using this notation

$$\frac{d}{dt}\ m_d(t) = \sum_{i \in S} r_i^{(d)}\ n_i(t)\ .$$

If the $\mathbf{k}(\bullet)$ transition rate matrix depends not only on $\mathbf{n}(t)$, but also on $\mathbf{m}(t)$, the behavior of the system is characterized by the following theorem.

Theorem 3. *The normalized state vector, $\mathbf{n}(t)$, and the normalized reward vector, $\mathbf{m}(t)$, of the process tends to be deterministic, in distribution, as N tends to infinity and they satisfy the following differential equations*

$$\begin{aligned} \frac{d}{dt}\mathbf{n}(t) &= \mathbf{n}(t)\ \mathbf{k}(\mathbf{n}(t), \mathbf{m}(t)), \\ \frac{d}{dt}\ \mathbf{m}(t) &= \mathbf{n}(t)\ \mathbf{R}, \end{aligned} \tag{16}$$

where the i, d element of matrix $\mathbf{R}$ is $R_{id} = r_i^{(d)}$.

Introducing vector $\mathbf{v}(t) = [\mathbf{n}(t), \mathbf{m}(t)]$ and matrix $\mathbf{H}(\mathbf{v}(t)) = \begin{array}{|c|c|} \hline \mathbf{k}(\mathbf{n}(t), \mathbf{m}(t)) & \mathbf{R} \\ \hline 0 & 0 \\ \hline \end{array}$, (16) can be rewritten to the following form which is similar to the differential equation in (4)

$$\frac{d}{dt}\mathbf{v}(t) = \mathbf{v}(t)\ \mathbf{H}(\mathbf{v}(t)). \tag{17}$$

Similar to Theorem 1, the proof of the Theorem 3 requires a detailed investigation of the properties of $\mathbf{k}(\bullet)$, which we neglect here.

VI. Mean field analysis of queues with memory dependent load

Let us consider the queueing system of Section III but with two types of queues of different buffer length $B^{(1)}$ and $B^{(2)}$ such that σ portion of the queues are of size $B^{(1)}$ and $1-\sigma$ portion of them of size $B^{(2)}$. I.e., $S^{(1)} = \{0, 1, \ldots, B^{(1)}\}, s^{(1)} = B^{(1)} + 1$, and $S^{(2)} = \{0, 1, \ldots, B^{(2)}\}, s^{(2)} = B^{(2)} + 1$.

The goal is to set the traffic served by the different type of queues to a predefined value, ξ. To this end, we can apply a reward based queue selection policy: we define two memory variables $m_1(t)$ and $m_2(t)$ accumulating the amount of traffic served by type 1 and 2 queues. The associated type 1 reward rates are

$$\frac{d}{dt}\ m_1(t) = \sum_{i=0}^{B^{(1)}} r_i^{(1)}\ n_i^{(1)}(t) = \sum_{i=1}^{B^{(1)}} \mu\ n_i^{(1)}(t)\ .$$

I.e., $r_i^{(1)}$ is μ if $i \geq 1$. The reward rates associated with $m_2(t)$ are defined similarly.

An incoming customer is directed to that type of queues which served less customers than the predefined ratio and among the queues of this type it attends the shortest one. I.e., a customer attends a type 1 queue, if $\frac{m_1(t)}{m_2(t)} < \frac{\xi}{1-\xi}$, a type 2 queue if if $\frac{m_1(t)}{m_2(t)} > \frac{\xi}{1-\xi}$, and attends the shortest of all queues if $\frac{m_1(t)}{m_2(t)} = \frac{\xi}{1-\xi}$.

The structure of the transition matrix of type 1 and 2 entities remains the same as the one in (8) with $B^{(1)}$ and $B^{(2)}$ customers, but the arrival rates depend also on the memory

variables

$$\lambda_i^{(1)}(\mathbf{n}(t),\mathbf{m}(t)) =$$

$$\begin{cases} 0 & \text{if } \frac{m_1(t)}{m_2(t)} > \frac{\xi}{1-\xi} \text{ or } n_i^{(1)}(t)=0 \text{ or} \\ & \left(\frac{m_1(t)}{m_2(t)} = \frac{\xi}{1-\xi} \text{ and} \right. \\ & \left. \exists k<i \text{ s.t. } n_k^{(1)}(t)>0 \right), \\ \dfrac{\lambda}{n_i^{(1)}(t)} & \text{if } \frac{m_1(t)}{m_2(t)} < \frac{\xi}{1-\xi}, \\ \dfrac{\lambda}{n_i^{(1)}(t)+n_i^{(2)}(t)} & \text{if } \left(\frac{m_1(t)}{m_2(t)} = \frac{\xi}{1-\xi} \text{ and} \right. \\ & \left. \forall k<i, n_k^{(1)}(t)=n_k^{(2)}(t)=0 \right), \end{cases} \tag{18}$$

The arrival rates to type 2 entities are symmetric with (18).

Starting from $\mathbf{n}^{(1)}(0)=\{\sigma,0,\ldots,0\}$ and $\mathbf{n}^{(2)}(0)=\{1-\sigma,0,\ldots,0\}$ this memory based queue selection policy results in

$$\lim_{t\to\infty}\mathbf{n}^{(1)}(t) = \begin{cases} \{\sigma-\xi\frac{\lambda}{\mu},\xi\frac{\lambda}{\mu},0,\ldots,0\} & \text{if } \frac{\lambda}{\mu}<\frac{\sigma}{\xi}, \\ \{0,0,\ldots,0,\sigma\} & \text{if } \frac{\sigma}{\xi}\le\frac{\lambda}{\mu}, \end{cases}$$

$$\lim_{t\to\infty}\mathbf{n}^{(2)}(t) = \begin{cases} \{1-\sigma-(1-\xi)\frac{\lambda}{\mu},(1-\xi)\frac{\lambda}{\mu},0,\ldots,0\} & \text{if } \frac{\lambda}{\mu}<\frac{1-\sigma}{1-\xi}, \\ \{0,0,0,\ldots,1-\sigma\} & \text{if } \frac{1-\sigma}{1-\xi}\le\frac{\lambda}{\mu}. \end{cases}$$

and

$$\lim_{t\to\infty}\frac{\mathbf{m}(t)}{t} = \{\lambda\xi,\lambda(1-\xi)\}\ .$$

Note that the memory variables tends to infinity as time increases.

VII. Conclusion

Mean field theory aims at representing a multi-body problem constituted by a large number of interacting identical entities with a one-body problem where the interdependencies are still considered. We have shown how the mean field method can be applied to performance problems where the interacting entities are represented by Continuous Time Markov Chains (CTMC). We have applied this powerful method to study different dependent policies for feeding Markovian queues with a finite buffer. The mutual interaction is modeled by defining the transition rates of a tagged entity as a function of the proportion of queues in each state and solving a differential equation defined over the normalized state occupancies. Various performance indices are computed and the behaviour of the interdependent policies is compared with the independent case.

Acknowledgments

The authors thank the valuable comments of the reviewers. These comments made us to reevaluate the derivations and the statements of the paper with respect to the submitted version. Unfortunately we could not answer all involved questions in the available short time, but we made efforts to compose a correct and still meaningful version of the paper.

The research presented in this paper was partially supported by the EU projects CRUTIAL ((http://crutial.cesiricerca.it/) and NAPA-WINE (www.napa-wine.eu) and partially by OTKA grant no. K61709.

References

[1] F. Ball, R.K. Milne, I.D. Tame, and G.F. Yeo. Superposition of interacting aggregated continuous-time Markov chains. *Advances in Applied Probability*, 29:56–91, 1997.

[2] M. Benaim and J-Y. Le Boudec. A class of mean field interaction models for computer and communication systems. *Performance Evaluation*, doi:10.1016/j.peva.2008.03.005, 2008.

[3] G. Bolch, S. Greiner, H. de Meer, and K.S. Trivedi. *Queueing Networks and Markov Chains*. Wiley, II Edition, 2006.

[4] J.Y. Le Boudec, D. McDonald, and J. Mundinger. A generic mean field convergence result for systems of interacting objects. In *4th Int Conf on Quantitative Evaluation of Systems - QEST2007*, 2007.

[5] P. Buchholz. Hierarchical Markovian models -symmetries and aggregation. *Performance Evaluation*, 22:93–110, 1995.

[6] P. Buchholz. Hierarchical structuring of superposed GSPNs. *IEEE Transactions Software Engineering*, 25:166–181, 1999.

[7] P. Buchholz and T. Dayar. Comparison of multilevel methods for Kronecker based Markovian representations. *Computing*, 73:349–371, 2004.

[8] M. Gribaudo, C.-F. Chiasserini, R. Gaeta, M. Garetto, D. Manini, and M. Sereno. A spatial fluid-based framework to analyze large-scale wireless sensor networks. In *IEEE International Conference on Dependable Systems and Networks, DSN2002*, 2005.

[9] J. Hillston. Fluid flow approximation of PEPA models. In *2nd International Conference on Quantitative Evaluation of Systems - QEST*, 2005.

[10] M. Gribaudo, M. Telek, and A. Bobbio. Mean field methods in performance analysis. Technical report, Dip Informatica - Università Piemonte Orientale; www.di.unipmn.it/Tecnical-R/index.htm, 2008.

[11] J.M. Kelif and E. Altman. Downlink fluid model of CDMA networks. In *IEEE 61th Vehicular Technology Conference (VTC 2005)*, 2005.

[12] P. Kemper. Transient analysis of superposed GSPNs. *IEEE Trans Soft Engineering*, 25:182–193, 1999.

[13] M. Opper and D. Saad. *Advanced Mean Field Methods: Theory and Practice*. MIT University Press, 2001.

[14] B.D. Plateau and K. Atif. Stochastic automata network for modeling parallel systems. *IEEE Transactions on Software Engineering*, 17:1093–1108, 1991.

[15] B.D. Plateau and J.M. Fourneau. A methodology for solving Markov models of parallel systems. *Journal of Parallel and Distributed Computing*, 12:370–387, 1991.

[16] D. Zhang, D. Gatica-Perez, S. Bengio, and D. Roy. Learning influence among interacting Markov chains. In *Adv Neural Information Processing Systems (NIPS)*, 2005.