A Product-form Model for the Analysis of Systems with Aging Objects

Filippo Cavallin
DAIS - Università Ca' Foscari Venezia
Venezia, Italy
Email: fcavalli@dais.unive.it

Andrea Marin
DAIS - Università Ca' Foscari Venezia
Venezia, Italy
Email: marin@dais.unive.it

Sabina Rossi
DAIS - Università Ca' Foscari Venezia
Venezia, Italy
Email: srossi@dais.unive.it

Abstract—In this paper we propose a new model for the analysis of systems with aging objects such as Time-To-Live cache. We consider a model with an underlying Continuous Time Markov Chain in which objects can be completely or partially rejuvenated. In the former case the object becomes fresh, while in the latter all the objects are simultaneously rejuvenated so that the youngest becomes fresh. We show that under the so-called Independent Reference Model assumption our model is numerically tractable and has a product-form equilibrium distribution. Furthermore, we consider the case in which the object aging stops after a certain threshold and hence the partial rejuvenation introduces a probabilistic behaviour. Also in this case, we can derive a product-form equilibrium distribution under some mild conditions. The models presented in this paper may be interpreted as a new class of G-networks with catastrophes and partial flushing.

I. Introduction

Stochastic models are powerful tools for assessing the non functional quantitative properties of computer networks, communication systems, and software architectures. In many practical applications, Markov processes are the stochastic processes underlying the considered models and their performance evaluation is carried out by using the well-known methods for the analysis of transient or stationary behaviour of Markov processes. In this paper we focus on the analysis of models whose underlying process is a Continuous Time Markov Chain (CTMC).

Since its introduction [8], the theory of product-form solutions has been playing an important role for the practical analysis of models with underlying CTMCs as it allows for an efficient derivation of the stationary performance indices even when the process's state space is huge and the analysis methods based on the solution of the system of global balance equations become computationally prohibitive. Even more interestingly, for a class of product-form models, including the ones studied in this paper, the performance indices can be derived without even generating the joint state space. Successful applications of product-form theory include the BCMP theorem [4], the modelling of neural networks [28], the analysis of systems with fork and join constructs [31], the loss networks [29] and the performance evaluation of wireless networks [7], just to mention a not exhaustive list.

In this paper we focus on modelling and analysing systems with aging objects by means of product-form models. These systems consist of a set of objects which are associated with an age, e.g., the timestamp of their creation or latest access. As time passes, the objects become older. Two types of events can rejuvenate the objects:

- total rejuvenation, i.e., the object timestamp is set to the current time. This event affects only one object.
- partial rejuvenation. In this case the event affects the whole system since the objects are all rejuvenated for the same time interval so that the youngest is associated with the current timestamp.

An example of such a system is a TTL cache in which the total rejuvenation occurs when an object is accessed and the partial rejuvenation can be seen as a method to prevent an under utilisation of the cache memory in case of periods of inactivity. The networking research community has renewed its interest in the performance of caching systems due to the new delivery methods for distributing contents in the networks. The huge number of proxy servers has led to the design of Content Delivery Networks (CDN) which are used by the content providers to deliver information in a large and spread population of users. Caching contents that have the greatest demand closer to the users' locations allows one to improve the client-perceived experience, to reduce the server load and optimize the bandwidth requirements. In this perspective, the caching system plays a fundamental role in the gradual shift from the traditional paradigm of host-to-host communication to the new host-to-content model. Other applications of aging systems are shown in [20], [37] where the failure of nodes in distributed systems are handled by means of checkpoints.

Technical contributions and related work

The main contributions of this work with respect to the literature are the following.
a) We present two models in product-form for the performance evaluation of systems with aging objects. The main difference between the two models is that one allows the object age to grow indefinitely, while the other introduces a maximum age threshold. We discuss the implications on practical applications with some examples. The product-form analysis that we demonstrate is interesting for at least two aspects. The first is that neither the joint CTMC nor the CTMC underlying a single model are reversible as it happens, e.g., in Jackson's queues [27] and G-queues [18]. The second interesting aspect

is that the synchronisations among the objects are not pairwise, i.e., at a given epoch more than two objects can simultaneously change their states. There are few results in this direction in the literature of product-forms. In [19], [16] the authors consider queueing networks in which the departure of a customer from a queue causes a movement of one job from a second queue to a third one, hence causing the simultaneous state change of three components. However, the extension of the result to more than three components is not trivial mainly because the proof technique adopted in those papers is based on solving the system of global balance equations (GBEs). The first model we propose is in the style of G-networks as proposed in [14], while the one with maximum aging is, to the best of our knowledge, very peculiar since very few product-forms are known for finite state space models [1], [2]. The contribution of the unbounded model with respect to [14] is twofold. First, the proof is not based on the solution of the system of global balance equations of the joint model. Secondly, we consider individual jumps of the objects to the zero state. Our proof method is based on the quasi-reversibility property [29] and the Reversed Compound Agent Theorem (RCAT) [22], [3]. Both these results provide a way to elegantly prove the product-form of a CTMC but they consider only pairwise synchronisations and hence they cannot be straightforwardly applied to study our models. We show that they can still be used by introducing a passage to the limit for a transition rate in a similar fashion of what has been done in [9], [23], [31]. Proofs of product-forms based on quasi-reversibility are simple to handle and compositional in the sense that they allow the combination of the models that we study here with others which are known to be quasi-reversible while maintaining the product-form of the equilibrium distribution. As a consequence, heterogeneous networks may be studied without constructing the joint Markov chain.

b) We show how to numerically derive the models' performance indices without constructing the joint CTMCs. This is important because the structures of these chains can be complex since the transitions corresponding to partial rejuvenations depend on the global state of the models. The derivation of the performance indices requires to solve a non-linear system of equations. We propose a fixed point algorithm to tackle this problem and show its efficiency and convergence properties on numerous examples. The system of equations admits a unique positive solution. With respect to [14], [15], we do not require any modification of the network of objects in order to obtain the convergence of the algorithm.

c) As an example we apply our model for the analysis of TTL cache with partial rejuvenation. First we propose an ideal model, whose implementation is very expensive, in which a timer is associated with each object despite the fact that it is inside or outside the cache. We study the performance indices under the Independent Reference Model (IRM) assumptions [17], [30], [38]. Then, we consider a model in which we maintain the timers only for the objects inside the cache. The partial rejuvenation of objects outside the cache has a probabilistic effect, i.e., the object may remain outside or can be copied

inside the cache according to a Bernoulli trial. We prove that it is possible to obtain exactly the same expected performance indices of the ideal model while maintaining the product-form property. We discuss how it is possible to dynamically set the model's parameters to achieve some performance goals. The analyses of TTL caches, often connected to form networks, have been widely addressed in the latest years (see, e.g., [5], [6], [12], [13] and the references therein). In our case study we consider a simpler situation of a single cache as in [35]. Clearly, the analysis becomes challenging because of the partial rejuvenation signals which aim at avoiding the under utilisation of the cache.

Structure of the paper

The paper is structured as follows. Section II gives some theoretical background and introduces the notation. In Section III we present the model with unbounded aging and prove the product-form. The application to the analysis of a TTL cache with partial rejuvenation is shown in Section IV. Section V presents the result about the model with bounded aging and its application is shown in Section VI. Final remarks are given in Section VII.

II. PRELIMINARIES

In this section we review some basic notions about the analysis of CTMCs. Let $X(t)$ be a stationary CTMC on the state space $\mathcal{S}$. Its reversed process, denoted $X^R(t)$, is still a stationary CTMC [22], [29] whose transition rates are defined as follows:

$$q^R(s_1, s_2) = \frac{\pi(s_2)}{\pi(s_1)} q(s_2, s_1), \tag{1}$$

where $q(s_2, s_1)$ is the transition rate from state s_2 to s_1 in $X(t)$ and $q^R(s_1, s_2)$ its inverse in $X^R(t)$. $X(t)$ is reversible if it is stochastically indistinguishable from $X^R(t)$. Henceforth we assume that $X(t)$ is ergodic, since the models we are presenting in the following sections are all unconditionally ergodic. The equilibrium distributions π of $X(t)$ and $X^R(t)$ are identical. Moreover, given the forward and the reversed chain, the following generalised Kolmogorov's criteria hold.

Proposition 1 (Kolmogorov's generalised criteria [22]):
Let $X(t)$ be an ergodic CTMC with state space $\mathcal{S}$ and infinitesimal generator $\mathbf{Q}$, then $Y(t)$ with the same state space and infinitesimal generator $\mathbf{Q}'$ is the reversed process $X^R(t)$ if and only if:

- For every state $s_1 \in \mathcal{S}$ we have

$$\sum_{\substack{s_2 \in \mathcal{S} \\ s_2 \neq s_1}} q(s_1, s_2) = \sum_{\substack{s_2 \in \mathcal{S} \\ s_2 \neq s_1}} q'(s_1, s_2),$$

  i.e., the residence times in a state in the forward and in the reversed processes have the same distribution.
- For every finite sequence of states $s_1, s_2, \ldots, s_n \in \mathcal{S}$, it holds that:

$$q(s_1, s_2)q(s_2, s_3) \cdots q(s_n, s_1)$$
$$= q'(s_1, s_n)q'(s_n, s_{n-1}) \cdots q'(s_2, s_1).$$

In the analysis of the model presented in Section III we will use Proposition 1 to derive the rates of the reversed chain of the CTMC underling the model.

We should stress that one can derive the reversed process $X^R(t)$ for any stationary chain $X(t)$ even if this is not reversible. Indeed, we will widely base the product-form analysis presented in Sections III and V on the derivation of the reversed processes of the CTMCs underlying the proposed models even if these chains are *not* reversible.

If one knows the infinitesimal generator of both the forward and the reversed chain, $\mathbf{Q}$ and $\mathbf{Q}^R$, respectively, then we can compute the expression of the equilibrium distribution in a very efficient way. Indeed, it suffices to choose an arbitrary reference state $s_0 \in \mathcal{S}$ and then to compute the equilibrium probability of any state $s \in \mathcal{S}$ with respect to s_0 by finding a path from s_0 to s, e.g.,

$$s_0 \to s_1 \to s_2 \to \cdots \to s_n = s.$$

Then, we have (see, e.g., [32], [33])

$$\pi(s) = \pi(s_0) \frac{\prod_{i=1}^n q^R(s_i, s_{i-1})}{\prod_{i=1}^n q(s_{i-1}, s_i)}. \tag{2}$$

Finally, we recall that quasi-reversibility [9], [29] is a sufficient condition to ensure that the synchronisation of a set of models whose underlying stochastic process is an ergodic CTMC has product-form solution. Informally, we can think that the transition in one of the components can trigger a transition in another by sending a signal. In the terminology used in [29] we say that there is a departure from the component sending the signal and an arrival at the one receiving the signal, while in the process algebraic terminology [25], we say that the component which sends the signal is active and the one receiving it is passive. In the analysis of quasi-reversible models, the receiver cannot prevent the transitions in the sender (but it may ignore them). Fore the sake of simplicity, we assume that a component sends a synchronising signal only to another component (that will be the case for the models we study in the following sections).

Proposition 2 (Quasi-reversibility): Given a cooperating component let $\mathcal{T} \subset \mathcal{S} \times \mathcal{S}$ be the set of transitions that synchronise with another model. Then, the model is quasi-reversible if there exists $x \in \mathbb{R}^+$ such that for all the states $s \in \mathcal{S}$, we have:

$$x = \frac{\sum_{s':(s',s)\in\mathcal{T}} \pi(s')q(s',s)}{\pi(s)}.$$

It is well-known that quasi-reversibility is a sufficient condition for the existence of a product-form equilibrium distribution of the joint model [22], [29], [34].

III. A MODEL FOR AGING OBJECTS

A. *The stochastic process underlying the collection of aging objects.*

We consider a set of K objects whose age at time t is modelled by a stochastic process $Y_k(t)$, $1 \leq k \leq K$,

which takes values in $\mathbb{N}$. Let $\mathbf{Y}(t) = (Y_1(t), \ldots, Y_K(t))$ be the stochastic process associated with all the objects and let $\mathbf{u} = (u_1, \ldots, u_K)$ be its state at time t, and $u_{\min} = \min(u_i, 1 \leq i \leq K)$. We characterise the stochastic process $\mathbf{Y}(t)$, for $h \to 0^+$, as:

$$Pr\{\mathbf{Y}(t+h) = \mathbf{u} - \mathbf{1}u_{\min}|\mathbf{Y}(t) = \mathbf{u}\} = \eta h + o(h)$$
$$Pr\{\mathbf{Y}(t+h) = \mathbf{u}[u_k \leftarrow 0]|\mathbf{Y}(t) = \mathbf{u}\} = \lambda_k h + o(h)$$
$$Pr\{\mathbf{Y}(t+h) = \mathbf{u}[u_k \leftarrow u_k + 1]|\mathbf{Y}(t) = \mathbf{u}\}$$
$$= \gamma_k h + o(h)$$
$$Pr\{\mathbf{Y}(t+h) = \mathbf{u}|\mathbf{Y}(t) = \mathbf{u}\}$$
$$= 1 - \left(\eta + \sum_{k=1}^K \lambda_k + \gamma_k\right) h + o(h)$$

where $\mathbf{u}[u_k \leftarrow val]$ denotes vector $\mathbf{u}$ in which component k takes value val, and $\mathbf{1}$ is a vector of 1s with size K. We say that λ_k is the *refresh rate* for object k, η is the partial rejuvenation rate and γ_k is the aging rate. Clearly, $\mathbf{Y}(t)$ is a CTMC.

We are interested in computing the equilibrium distribution of $\mathbf{Y}(t)$. Notice that the partial rejuvenation events are such that processes $Y_k(t)$ are not stochastically independent. Moreover, $\mathbf{Y}(t)$ is neither reversible nor in product-form, and hence the computation of the equilibrium distribution can be prohibitive for a large number of objects even if truncation is applied.

B. *A product-form approximation for* $\mathbf{Y}(t)$

In this section we approximate the CTMC $\mathbf{Y}(t)$ introduced in Section III-A by a CTMC $\mathbf{X}(t)$. We introduce a different semantics for the partial rejuvenation signals that allows us to prove that $\mathbf{X}(t)$ has a product-form equilibrium distribution. Let us introduce the following notation:

$$k^+ = \begin{cases} k+1 & \text{if } k < K \\ 1 & \text{if } k = K \end{cases} \qquad k^- = \begin{cases} k-1 & \text{if } k > 1 \\ K & \text{if } k = 1 \end{cases}.$$

In $\mathbf{X}(t)$, a partial rejuvenation signal iteratively decreases the age of the objects according to their orders until we find a fresh object, i.e., whose age is 0. Formally, the destination state reached by $\mathbf{X}(t)$ immediately after a partial rejuvenation signal is given by Algorithm 1. The aging rate and the refresh rate have the same effects of those described for $\mathbf{Y}(t)$.

Notice that the model does not exactly implement the behaviour described for $\mathbf{Y}(t)$ as shown by the following counterexample. Consider a model with $K = 4$ objects, and suppose their ages are $(2, 1, 3, 4)$ when the reset signal arrives at the first timer. Then, at t_0^+ the state of $\mathbf{Y}(t)$ is $(1, 0, 2, 3)$, however Algorithm 1 performs a first iteration on all the objects and a second one that stops when it reaches the second object leading to a state in t_0^+ which is $(0, 0, 2, 3)$. Notice that the difference between the correct age of an object and the one which is computed by Algorithm 1 can be of at most one unity, therefore we consider the approximation acceptable.

Algorithm 1: Reset of timers upon arrival at time t_0 of a reset signal to the o_κ timer.

Data: κ, $X_i(t_0)$
Result: $X_i(t_0^+)$
for $k \in [1, K]$ **do**
    | $X_k(t_0^+) = X_k(t_0)$;
end
$k \leftarrow \kappa$;
while $X_k(t_0^+) > 0$ **do**
    | $X_k(t_0^+) \leftarrow X_k(t_0^+) - 1$;
    | $k \leftarrow k^+$;
end

C. Stationary analysis of $\mathbf{X}(t)$

In this section we derive the stationary distribution of the model and prove that it is in product-form and unconditionally stable.

Theorem 1: The stochastic process $\mathbf{X}(t)$ has the product-form equilibrium distribution:

$$\pi^X(\mathbf{u}) = \prod_{k=1}^{K} \pi_k^X(u_k) = \prod_{k=1}^{K} (1 - \rho_k)\rho_k^{u_k}, \qquad (3)$$

where ρ_k is the solution of the following non linear system of rate equations:

$$\rho_k = \frac{1}{2(x_k + \eta/K)} \cdot \left(\lambda_k + \frac{\eta}{K} + x_k + \gamma_k \right. \qquad (4)$$

$$\left. - \sqrt{\left(\lambda_k + \frac{\eta}{K} + x_k + \gamma_k\right)^2 - 4\gamma_k\left(x_k + \frac{\eta}{K}\right)} \right)$$

$$x_{k^+} = \rho_k\left(x_k + \frac{\eta}{K}\right) \qquad (5)$$

for all $k = 1, \dots, K$, and u_k is the age of object k. Moreover, the model is unconditionally stable for strictly positive rates λ_k and γ_k.

In principle, one could prove Theorem 1 by substituting Expression (3) in the system of the global balance equations for $\mathbf{X}(t)$ in a similar fashion to what has been done in [14], [15]. Nevertheless, the complexity in the structure of $\mathbf{X}(t)$, which depends on Algorithm 1, makes this way of proving the equilibrium distribution long and prone to errors.

Proof: Let us consider the model for the single object aging depicted in Figure 1-(A). In this model the arrivals of partial rejuvenation signal occur with rate ξ_k. At the arrival of such a signal at object k, this moves from state $u > 0$ to state u' and then propagates the signal to object k^+ while moving to state $u - 1$. The signal propagation becomes instantaneous as required by the definition of $\mathbf{X}(t)$ when $\beta \to \infty$. We prove the product-form by resorting to the quasi-reversibility property, i.e., we prove that the occurrences of the transitions from state i' to state $(i-1)'$, $i > 0$, at t_0 are independent of the occurrences of the same transitions subsequent t_0. These
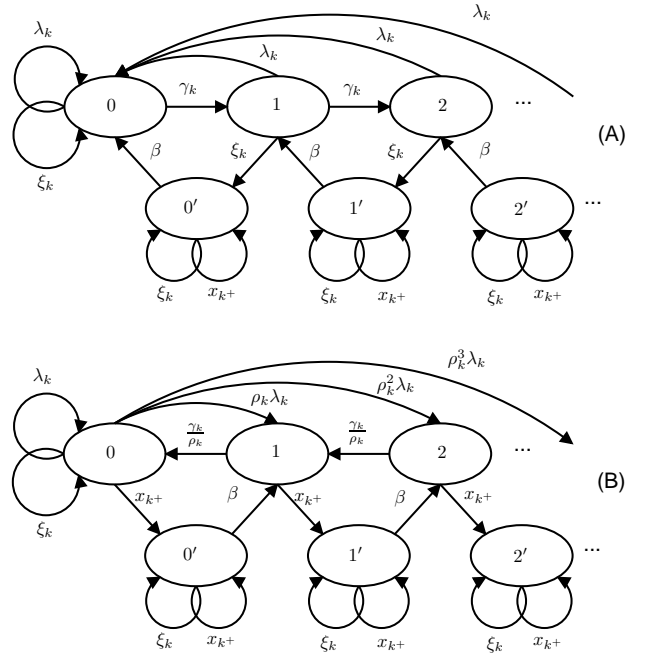

Fig. 1. Quasi reversible model for the single object aging. (A)-Forward process and (B)-Reversed process.

are the transitions that forward the partial rejuvenation signal from object k to k^+.

Claim 1: The reversed CTMC of the process shown in Figure 1-(A) is the process shown in Figure 1-(B) where

$$\rho_k = \frac{1}{2\xi_k}(\lambda_k + \gamma_k + \xi_k - ((\lambda + \gamma_k + \xi_k)^2 - 4\xi_k\gamma_k)^{1/2}$$

and $x_{k^+} = \rho_k\xi_k$.

The claim can be readily verified by applying the generalised Kolmogorov's criteria given in Proposition 1. First, we check that the total rate out of every state in the forward and the reversed process are the same. For states u_k' it is trivial. For state $u_k > 0$ we have:

$$\gamma_k + \lambda_k + \xi_k = \frac{\gamma_k}{\rho_k} + x_{k^+}, \qquad (6)$$

which is satisfied for the definition of ρ_k given in Claim 1. For state 0 we have to prove that $\gamma_k = \sum_{u_k=1}^{\infty} \rho_k^{u_k}\lambda_k + x_{k^+} = \lambda_k\rho_k/(1 - \rho_k) + x_{k^+}$ which is equivalent to Equation (6). Finally, we can easily see that the product of the rates in the forward and reversed processes are the same for every minimal cycle. This is sufficient to prove Claim 1. The model is quasi-reversible (and satisfies RCAT conditions) because every state in the reversed process has an outgoing rate of x_{k^+} associated with the propagation of the partial rejuvenation signal. By the definition of the rates in the reversed process, we can derive the equilibrium distribution of state $u_k > 0$ by using the path from u_k to 0 according to Equation (2). We have $\pi_k^X(u_k) = \pi_k^X(0)\rho_k^{u_k}$ and we easily see that $\pi_k^X(u_k') = \pi_k^X(u_k)\xi_k/\beta$. Notice that neither x_{k^+} nor $\pi_k^X(u_k)$ depends on β. Indeed, for $\beta \to \infty$ we have the instantaneous propagation of the
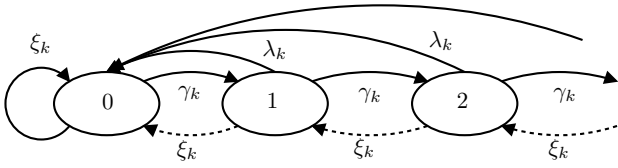
Fig. 2. Simplified version of the model of Figure 1 for proving the quasi-reversibility property.

partial rejuvenation signal as required and the model is quasi-reversible and $\pi_k^X(u_k') \to 0$.

Hence, the theorem is proved by noticing that η_k is the sum of the rates of the partial rejuvenation signal arriving from k^- and from outside, i.e., $\xi_k = x_k + \eta/K$. ∎

The proof method based on the passage to the limit for modelling instantaneous propagation of transitions is inspired by the approach used in [9], [23], [31] for different networks and is alternative to the process algebraic one recently proposed in [24]. Notice that, thanks to the passage to the limit $\beta \to \infty$, proving the product-form of a component as the one shown in Figure 1 can be readily done by considering the simplified model shown in Figure 2, in the sense that if the latter is quasi-reversible also the former is quasi-reversible. In the model depicted in Figure 2, one has just to prove that the reversed rates of the dotted transitions (those propagating the rejuvenation) are constant [22].

D. Solving the system of rate equations

The system of rate equations (4)-(5) does not generally have a symbolic solution and its degree grows very quickly with the number of timers. For this reason we introduce a fixed-point iteration with the aim of solving equations (4)-(5) numerically.

In Algorithm 2 we show the fixed-point iterations. The algorithm initialises the values for x_k randomly according to independent uniform random variables with support $(0, \gamma_k)$. Recall that the x_k may be interpreted as the reversed rates associated with the death transition in the model of Figure 2 and hence cannot exceed γ_k. Then, we iteratively compute ρ_k and x_k according to Equations (4) and (5), respectively, until we reach the desired accuracy, i.e., the L1-norm of the difference between two successive iterations on ρ_k is lower than ε.

Although we do not have a formal proof of the convergence for Algorithm 2, we carried out several tests with random generated models and always observed the computation of a good approximation for ρ_k.

The convergence of Algorithm 2 has been verified on $100,000$ random models with a population ranging from $1,000$ to $100,000$ objects. We have also tested the scalability of the fixed point algorithm with different numbers of objects. Using an Intel Core(TM)2 Duo CPU processor, we have analysed a model with 1000 objects and obtained the solution of the non-linear system in $0.5s$. We have noticed that the convergence time depends on the variance of the set of λ_ks. Indeed, for

---

Algorithm 2: Fixed-point algorithm for the solution of the system of rate equations (5).

Data: $\lambda_k, \gamma_k, \eta, \varepsilon$
Result: ρ_k
Initialise randomly x_k for $k = 1, \ldots K$ with uniform distribution in $(0, \gamma_k)$;
Compute ρ_k using Equation (4);
repeat
  $\rho_k' \leftarrow \rho_k$ for all $k = 1, \ldots, K$;
  Update x_k using ρ_k' by Equation(5);
  Compute ρ_k using Equation (4);
until $|\rho_k' - \rho_k| < \varepsilon$;

$100,000$ objects, the computation time varies between 30 and 160 seconds, where the lowest time corresponds to the lowest variance.

IV. APPLICATION: ANALYSIS OF AN IDEAL TTL CACHE WITH REJUVENATION

In this section we use the results derived in Section III to analyse a class of TTL caches with rejuvenation. TTL caches are attracting the attention of the research community especially in the context of the analysis of Information-Centric Networks [21]. Analyses of TTL cache networks are shown, e.g., in [5], [6], [12], [13]. In TTL caches, each object is associated with a timer. At each object access, the timer is reset and when it expires the corresponding object is evicted from the cache. Here, we propose a rejuvenation mechanism for the timers that reduces their ages in such a way that the last recently used is set to 0. This aims at reducing the problem of under-utilisation of the TTL cache in case of long periods of inactivity. The model we propose here represents an *ideal case* in the sense that we require to maintain a timer for all the objects, including those which are not present in the cache. With the model proposed in Section V we will propose a workaround to this implementation issue.

A. System description

We study the cache model under the Independent Reference Model (IRM) assumptions in continuous time. IRM requires the object requests to be generated according to independent Poisson processes. Although this scenario may be unrealistic for some practical performance analyses, still it is widely used as a benchmark to compare different caching policies as, e.g., in [30], [17], [10], [11], [38]. With reference to the model presented in Section III, we have that K is number of timers associated with the objects, $X_k(t)$ is the state of timer t. Requests of object k occur according to an independent Poisson process with rate λ_k and the partial rejuvenation signal occur with rate η and starts from an object chosen randomly with uniform distribution. At the k-th object request epoch we observe a cache hit if the timer associated with object k is not greater than threshold $T = T_k$, a cache miss otherwise.

B. Stationary performance indices

In this section we derive some performance indices of the model in equilibrium. We will give an expression for the cache hit/miss rates and probabilities and for the expected size of the cache as functions of ρ_k, with $k = 1, \ldots K$. Suppose that object k requires α_k bytes to be stored in the cache. Let us define the following stochastic process that corresponds to the cache size:

$$S(t) = \sum_{k=1}^{K} \alpha_k \mathbb{1}_{X_k(t) \leq T} ,$$

where $\mathbb{1}_{X_k(t) \leq T}$ is 1 if $X_k(t) \leq T$, 0 otherwise. We are interested in the evaluation of the expected cache size in equilibrium, i.e.,

$$\overline{S} = E \left[\lim_{t \to \infty} S(t) \right] .$$

Notice that the limit exists since $\lim_{t \to \infty} X_k(t)$ is the marginal distribution associated with the states of timer k of the ergodic CTMC underlying the model.

Proposition 3 (Expected cache size): In equilibrium, the expected cache size $\overline{S}$ is:

$$\overline{S} = \sum_{k=1}^{K} \alpha_k (1 - \rho_k^{T+1}) \tag{7}$$

Proof: The proof follows from the observation that in equilibrium the timer models behave as if they were independent (product-form). Therefore, the probability that the timer k is not over T is:

$$\lim_{t \to \infty} Pr\{X_k(t) \leq T\} = 1 - \rho_k^{T+1} . \tag{8}$$

The result is readily derived by weighting the stationary probability of finding an object in cache by its size. ∎

The expression for the standard deviation of the cache size can be used combined with Chebyshev's inequality to derive bounds on the probability that the cache size exceeds a given threshold.

Proposition 4 (Standard deviation of the cache size): The standard deviation of the cache size in equilibrium is:

$$\sigma_S = \sqrt{\sum_{k=1}^{K} \alpha_k^2 \rho_k^{T+1} \left(1 - \rho_k^{T+1}\right)} .$$

Proof: The proof follows after simple algebraic simplifications of the expression:

$$\sigma_S^2 = \sum_{k=1}^{K} \alpha_k^2 (1 - \rho_k^{T+1}) - \sum_{k=1}^{K} \alpha_k^2 (1 - \rho_k^{T+1})^2 .$$

∎

Proposition 5 (Hit and miss probability): In equilibrium, the probability of observing a cache hit (h_k) or a cache miss (m_k) event for a request of object k is:

$$h_k = 1 - \rho_k^{T+1} , \qquad m_k = \rho_k^{T+1} .$$

Proof: According to the PASTA property (see, e.g., [36]), in equilibrium, an event that occurs according to an independent Poisson process sees the stationary distribution. Then, the proof follows the lines of that of Proposition 3. ∎

The following proposition gives an expression for the rate at which we observe a cache hit or a cache miss event when the model is in equilibrium.

Proposition 6 (Total hit/miss rate): In equilibrium, the total cache hit rate is $H = \sum_{k=1}^{K} h_k \lambda_k$ and the miss rate is $M = \sum_{k=1}^{K} m_k \lambda_k$.

Finally, we give an expression for the probability of a request to generate a cache hit or a cache miss:

Proposition 7 (Total hit/miss probability): In equilibrium, the probability that an object request causes a cache hit event (P_H) or a cache miss event (P_M) is given by the following expressions:

$$P_H = \frac{\sum_{k=1}^{K} \lambda_k (1 - \rho_k^{T+1})}{\sum_{k=1}^{K} \lambda_k} ,$$

$$P_M = 1 - P_H = \frac{\sum_{k=1}^{K} \lambda_k \rho_k^{T+1}}{\sum_{k=1}^{K} \lambda_k} .$$

We observe that in a perfectly symmetric system, i.e., when all the objects are requested with the same rate and occupy the same space, the cache hit probability and rate are directly proportional to the expected space dedicated to the cache. This should be not surprising since the choice of evicting one particular object at a given time is arbitrary under the IRM assumption given that all the object requests occur with the same rate (recall that IRM inherits the memoryless property of the exponential distribution).

C. Experiments

In this section we present some experiments in order to evaluate the performances of the TTL cache with reset signals. Considering 1000 objects (and so 1000 timers) with dimension 1, $T = 20$, $\gamma = \gamma_k = 30$ and homogeneous $\lambda_k = 0.5$ for $1 \leq k \leq K$, we tested the effect of η on the hit probability. In this experiment, η varies from 0.001 to 5.0 with a step of 0.1. In Figure 3 we can see how η affects the probability of finding an object in the cache.

We study the influence of the variance of λ_k on the hit probability and the average space occupation of the cache. In Figure 4 we use sets of λ_k with growing variance and we notice that the fraction of objects in the cache is lower than the hit probability for higher variance. Intuitively, this
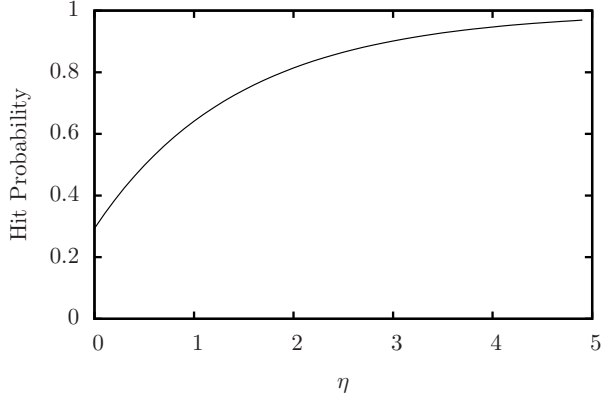
Fig. 3. Cache hit probability vs. η with homogeneous λ_k.

happens because in the TTL cache, the highly frequently required objects are present with very high probability and this causes a good cache hit probability even with small cache sizes. As a consequence, the benefits on the cache hit probability with higher values of η are lower.

Finally, we study the differences between the TTL model with resets and the model with the FIFO policy studied in [30]. In this case, we took $\gamma = \gamma_k = 80$, the $\lambda_k = (1, \ldots, 1, 10, 10)$ and $\eta = 0.005$. We varied T to get the same space occupation for the two models (for TTL model we consider the average space occupation). With low space occupation, the FIFO model works slightly better but, between 10% and 80% of cached objects, the TTL model with resets shows a higher cache hit probability (up to 10% of improvement) due to the fact that the reset signals avoid the removal of some objects from the cache.
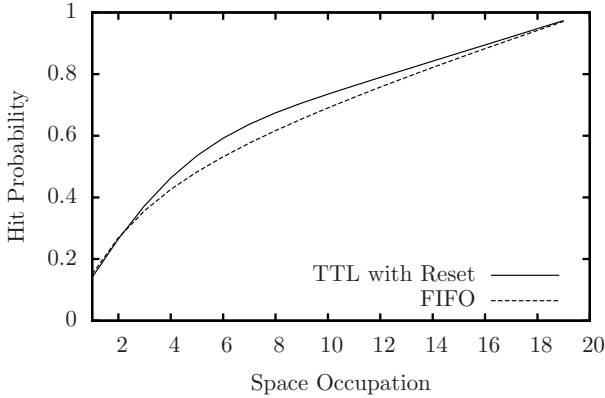

Fig. 5. Hit probability of TTL cache with reset and FIFO cache.

V. A MODEL FOR AGING OBJECTS WITH MAXIMUM THRESHOLD

In this section we consider a model which is similar to that described in Section III but in which the aging of objects has a maximum threshold. Once this threshold is reached, the
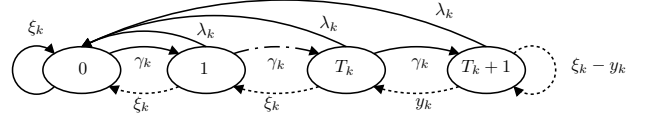

Fig. 6. Simplified model for an aging object with maximum age.

object stops its aging. The motivation to study this type of aging objects is that when aging is handled by timers and the proportion of *young* objects is small with respect to the object population, it is too computationally expensive to handle all the timers. So according to this idea we use one state to denote that an object is *very old* and hence we do not handle its timer any more. Clearly, we need to specify how the partial rejuvenation signal affects the objects in their terminal state. Intuitively, in order to get around the problem of the lack of knowledge of the true age of the objects in their terminal state we introduce a probabilistic behaviour, i.e., object k in its terminal state is partially rejuvenated with a certain probability p_k or it is left in its terminal state with probability $1 - p_k$. In both cases, the partial rejuvenation signal is transmitted immediately to object k^+.

In this case the analogue of the simplified model of Figure 2 is shown in Figure 6. We use $T_k + 1$ to denote the terminal state of aging object k. When the model is in state $T_k + 1$ and a partial rejuvenation signal arrives it may either jump to state T_k (rate y_k) or stay in state $T_k + 1$ (rate $\xi_k - y_k$), where $y_k < \xi_k$. We now study the conditions on y_k that gives the product-form. In order for the model of Figure 6 to be quasi-reversible, we must prove that all the dotted transitions have the same reversed rate [22]. A necessary condition to satisfy this constraint is that $\pi_k^H(u)/\pi_k^H(u-1) = \rho_k$, for all $u = 1, \ldots T_k$. Let us start by writing down the global balance equation associated with state u, with $0 < u < T_k$:

$$\pi_k^H(u)(\lambda_k + \xi_k + \gamma_k) = \pi_k^H(u+1)\xi_k + \pi_k^H(u-1)\gamma_k \,,$$

which can be rewritten as:

$$\lambda_k + \xi_k + \gamma_k = \rho_k \xi_k + \frac{\gamma_k}{\rho_k} \,. \tag{9}$$

Moreover, we have to satisfy the GBEs of states T_k and T_k+1:

$$\begin{cases} \pi_k^H(T_k)(\lambda_k + \xi_k + \gamma_k) = \pi_k^H(T_k+1)y_k + \pi_k^H(T_k-1)\gamma_k \\ \pi_k^H(T_k+1)(y_k + \lambda_k) = \pi_k^H(T_k)\gamma_k \end{cases} \tag{10}$$

The last condition we need to satisfy is that the reversed rate of the transitions with rates y_k and $\xi_k - y_k$ must be equal to the reversed rate of the transitions with rate ξ_k, i.e.:

$$\frac{\pi_k^H(T_k+1)}{\pi_k^H(T_k)}y_k = \xi_k - y_k = \rho_k \xi_k \,. \tag{11}$$

Notice that we have only one free variable, y_k, to satisfy Equations (9), (10), (11).

Lemma 1: The model depicted in Figure 6 is quasi reversible if:

$$y_k = \frac{1}{2}\left(\xi_k - \gamma_k + \sqrt{(\lambda_k + \gamma_k + \xi_k)^2 - 4\xi_k\gamma_k}\right) \,. \tag{12}$$
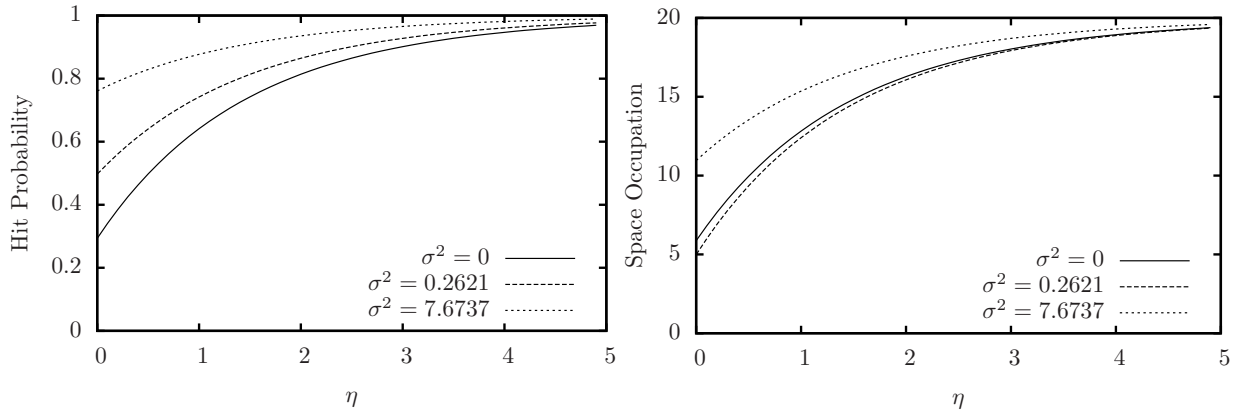
Fig. 4. Comparison between different heterogeneous λ_k with varying η.

In this case we have the following equilibrium distribution:

$$\pi_k^H(u) = \begin{cases} \pi_k^H(0)\rho_k^u & \text{if } u \le T_k \\ \pi_k^H(0)\rho_j^{T_k} \frac{\gamma_k}{\lambda_k + y_k} & \text{if } u = T_k + 1 \end{cases} \quad (13)$$

with

$$\rho_k = \frac{\lambda_k + \gamma_k + \xi_k - \sqrt{(\lambda_k + \gamma_k + \xi_k)^2 - 4\xi_k\lambda_k}}{2\xi_k} \quad (14)$$

and:

$$\pi_k^H(0) = (1 - \rho_k).$$

Proof: First we derive ρ_k from Equation (9) which gives:

$$\rho_k = \frac{\lambda_k + \gamma_k + \xi_k \pm \sqrt{(\lambda_k + \gamma_k + \xi_k)^2 - 4\xi_k\lambda_k}}{2\xi_k}.$$

We choose the solution given by Equation (14) because it is the only one that admits a positive solution for y_k.

From the second equation of System (10) we derive $\pi_k^H(T_k + 1)/\pi_k^H(T_k) = \gamma_k/(\lambda_k + y_k)$. If we divide both hand sides of the first equation for $\pi^H(T_k)$ and substitute the expression of $\pi_k^H(T_k+1)/\pi_k^H(T_k) = \gamma_k/(\lambda_k+y_k)$, we obtain:

$$\lambda_k + \xi_k + \gamma_k = \frac{\gamma_k y_k}{\lambda_k + y_k} + \frac{\gamma_k}{\rho_k}, \quad (15)$$

that can be reduced to a linear equation in y_k whose solution is given by Expression (12). We can prove that y_k is positive when all the other rates are also positive. At this point we have proved that for this choice of y_k we have:

- $\pi_k^H(u)/\pi_k^H(u-1) = \rho_k$ for all $0 < u \le T_k$
- $\pi_k^H(T_k + 1)/\pi_k^H(T_k) = \gamma_k/(\lambda_k + y_k)$

and hence the equilibrium distribution of Equation (13) can be readily derived. However, in order to prove that the model is quasi-reversible we still have to check the reversed rates of the dotted transitions of Figure 6. Notice that the reversed rates of the transitions with forward rate ξ_k from state u to $u-1$ is $\rho_k\xi_k$. This is equal to the reversed rate of the transition with forward rate y_k which is $\gamma_k y_k/(y_k + \lambda_k)$ as can be seen by comparing Equation (15) and the GBE (9). It remains to

prove that $\xi_k - y_k$ is positive and that its reversed rate, which is equal to its forward rate, is $\rho_k\xi_k$. The verification of this equality is purely algebraic. Finally, we derive the expression of $\pi_k^H(0)$ by normalising the stationary probabilities, i.e.:

$$\sum_{u=0}^{T_k} \pi_k^H(0)\rho_k^u + \pi_k^H(0)\rho_k^{T_k} \frac{\gamma_k}{\lambda_k + y_k} = 1.$$

After some algebraic reductions one obtains:

$$\pi_k^H(0) = \left[\frac{1 - \rho_k^{T_k+1}}{1 - \rho_k} + \frac{\rho_k^{T_k}\gamma_k}{\lambda_k + y_k} \right]^{-1} = (1 - \rho_k),$$

as required. This concludes the proof. ■

Theorem 2 gives the product-form equilibrium distribution for this model. Let $\mathbf{H}(t)$ be the CTMC underlying a set of aging objects with maximum age and partial rejuvenation.

Theorem 2: The stochastic process $\mathbf{H}(t)$ in which the rates satisfy the condition of Lemma 1 has the product-form equilibrium distribution:

$$\pi^H(\mathbf{u}) = \prod_{k=1}^{K} \pi_k^H(u_k) \quad (16)$$

where ρ_k is the solution of the non linear system of equations (4)-(5) and π_k^H is given by Equation (13) where $\xi_k = x_k + \eta/K$.

Proof: The proof follows the same steps of that of Theorem 1 given Lemma 1 stating the quasi-reversibility of the model in Figure 6. ■

Corollary 1 (Partial rejuvenation probability): The model has a product-form if at the arrival of a partial rejuvenation signal at an object in state $T_k + 1$, the probability of changing its state to T_k is:

$$p_k = \frac{\xi_k - \gamma_k - \lambda_k + \sqrt{(\gamma_k + \lambda_k + \xi_k)^2 + 4\gamma_k\xi_k}}{2\xi_k}, \quad (17)$$

where $\xi_k = x_k + \eta/K$.

Proof: The expression is obtained by simply computing y_k/ξ_k, since ξ_k is the rate at which the partial rejuvenation signal arrives at the object. ∎

The following result is important for understanding the connections between the model studied in Section III with unbounded aging and the one considered here. The result is in some sense surprising since it basically states that with the definition of rate y_k as specified by Equation (12) we have two enjoyable properties: the first is that the equilibrium distribution is in product-form and hence analytically tractable, and the second is that the stationary probability of observing a state u, with $1 \leq u \leq T_k$ is the same in the model with unbounded aging and that with bounded aging. Therefore, in a practical situation in which the object agings are handled by timers, one can replace the ideal model with unbounded aging with the one proposed in this section while maintaining the same performance indices, provided that the probability of rejuvenating an object k in state $T_k + 1$ upon the arrival of a partial rejuvenation signal is set according to Equation (17).

Corollary 2 (Equivalence corollary): Given the CTMCs $\mathbf{X}(t)$ and $\mathbf{H}(t)$ and let π^X and π^H be their equilibrium distribution functions. Then

$$\pi^H(\mathbf{u}) = \prod_{i=1}^{K} g_k(u_k),$$

where $\mathbf{u} = (u_1, \ldots, u_K)$ and

$$g_k(u_k) = \begin{cases} \pi_k^X(u_k) & \text{if } 0 \leq u_k \leq T_k \\ \sum_{u'=T_k+1}^{\infty} \pi_k^X(u') & \text{if } u_k = T_k + 1 \end{cases}$$

Therefore, we can see $\mathbf{H}(t)$ as an aggregation of $\mathbf{X}(t)$ that preserves the equilibrium distribution and the product-form.

It is worth of notice that $\mathbf{H}(t)$ is not a lumping (neither strong nor exact [26], [33]) of $\mathbf{X}(t)$. Moreover, we observe that the choice of y_k can be interpreted as the conditional transition rate from state $T_k + 1$ to state T_k of object k considered in isolation:

$$y_k = \frac{\pi_k^X(T_k + 1)}{\sum_{u'=T_k+1}^{\infty} \pi_k^X(u')} \xi_k.$$

VI. APPLICATION: REVISITING THE TTL CACHE WITH PARTIAL REJUVENATION

In Section IV we have shown an application of our theoretical findings in the analysis of a single TTL cache with partial rejuvenation. We called the model *ideal* since its actual implementation would be prohibitive because it would require to associate a timer with each object, including those which are not in cache. However, we overcome this problem by using the model presented in this section. For each object k we have a timer that is handled only when the object is in cache, i.e., its state u_k is $0 \leq u_k \leq T_k$. When the object is evicted from the cache, its state is $T_k + 1$. The partial rejuvenation signal

has a probabilistic effect on the objects outside the cache, i.e., they are rejuvenated with probability p_k as specified by Equation (17) or they stay in state $T_k + 1$ with probability $1 - p_k$. In both cases the object instantaneously propagates the signal to the following object k^+. By Corollary 2 the performance measures are the same as derived in Section IV.

We now address the inverse problem, i.e., finding γ_k and T_k such that the probability of copying in the cache an object which is outside, i.e., in state $T_k + 1$, is q_k and the expected eviction time is R_k. From a theoretical point of view, the problem has a simple solution:

$$\gamma_k = \frac{(1 - q_k)(\lambda + q_k \xi_k)}{q_k}, \tag{18}$$

then we may choose T_k in such a way that the expected eviction time is R_k, i.e., $\lceil T_k = \gamma_k R_k \rceil$. Nevertheless, the evaluation of Equation (18) requires the solution of the non-linear system (4)-(6) which may be computationally expensive to be performed dynamically. A dynamic adjustment of γ_k in order to obtain the desired probability q_k can be formulated by observing that p_k in monotonic on γ_k in Equation (17). Indeed, we have:

$$\frac{dp_k}{d\gamma_k} = \frac{1}{2\xi} \left(-1 + \frac{\gamma_k + \lambda_k - \xi_k}{\sqrt{(\gamma_k + \lambda_k + \xi_k)^2 - 4\gamma_k \xi_k}} \right)$$

which is always negative for positive transition rates. In other words, the dynamical control of p_k in order to reach the desired value q_k can be performed by simply augmenting γ_k (and reducing T_k) if the current value of p_k is higher than q_k, and by reducing it when it is lower.

VII. CONCLUSION

In this paper we propose two novel product-form models that can be applied for the analysis of systems with aging objects. These two models differ because in one the objects have an unbounded age while in the other there is a maximum threshold for the age. The approach that we propose can be applied also for the analysis of heterogeneous systems, i.e., systems in which some objects have a maximum threshold for the age and some others have not.

Informally, we can say that the peculiarity of these models is that the transitions are not "local" -as in most of product-form models- i.e., they may change the states of all the objects instantaneously and the effect of a partial rejuvenation event depends on the global state of the model.

We show that the CTMC underlying the model of objects with limited aging can be seen as an aggregation of the chain of the model with unbounded aging. This allowed us to prove some equivalence results on the expected performance indices. The consequences of these equivalences are important for practical applications especially if the system must maintain a timer for the objects. Indeed, by introducing a maximum age, we avoid to maintain the timers of all the objects that reach this age thus reducing the computational effort required for monitoring the object agings. We used as case-study the

analysis of a cache with TTL policy and partial rejuvenation of the objects.

The results can be extended in order to include the partial rejuvenation of clusters of objects (instead of all) and more sophisticated interactions among them in a similar fashion of what is considered in [14]. Indeed, our model may be seen as a queueing network with external independent Poisson arrival streams. The queues can only be partially flushed (i.e., the customers in the network are reduced as computed by Algorithm 1) or totally flushes, i.e., the number of customers in a single queue is set to 0. The exponential distribution of the equilibrium distributions stated by Theorem 1 allows the introduction of state-independent probabilistic customer routing.

The proposed model could also be further developed with the possibility of a probabilistic insertion of an object in the cache, i.e. when an object is not in the cache and it is requested, it can enter the cache with a probability q or remain outside with probability $1-q$. Moreover, we aim at overcoming the limitations of the IRM by allowing state-dependent request rates so that the last requested objects will be associated with an higher rate thus incorporating in the analysis the time-locality property of the network traffic.

REFERENCES

[1] I. F. Akyildiz. Exact product form solution for queueing networks with blocking. *IEEE Trans. on Comp.*, C-36-1:122–125, 1987.

[2] S. Balsamo, P. G. Harrison, and A. Marin. A unifying approach to product-forms in networks with finite capacity constraints. In *Proc. of SIGMETRICS*, pages 25–36, 2010.

[3] S. Balsamo and A. Marin. Performance engineering with product-form models: efficient solutions and applications. In *Proc. of ICPE 2011*, pages 437–448, 2011.

[4] F. Baskett, K. M. Chandy, R. R. Muntz, and F. G. Palacios. Open, closed, and mixed networks of queues with different classes of customers. *J. ACM*, 22(2):248–260, 1975.

[5] D.S. Berger, P. Gland, S. Singla, and F. Ciucu. Exact analysis of TTL cache networks. *Perf. Eval.*, 79:2–23, 2014.

[6] D.S. Berger, P. Gland, S. Singla, and F. Ciucu. Exact analysis of TTL cache networks: the case of caching policies driven by stopping times. In *SIGMETRICS 2014*, pages 595–596, 2014.

[7] R. Block and B. van Houdt. Spatial fairness in multi-channel CSMA line networks. In *Proc. of Valuetools 2014*, pages 1–8, 2014.

[8] P. J. Burke. The output of a queueing system. *Oper. Res.*, 4(6):699–704, 1956.

[9] X. Chao, M. Miyazawa, and M. Pinedo. *Queueing networks: customers, signals and product form solutions*. Wiley, 1999.

[10] A. Dan and D. Towsley. An approximate analysis of the LRU and FIFO buffer replacement schemes. In *Proc. of SIGMETRICS*, pages 143–152, 1990.

[11] P. Flajolet, D. Gardy, and L. Thimonier. Birthday paradox, coupon collectors, caching algorithms and self-organizing search. *Discrete Applied Mathematics*, 39:207–229, 1992.

[12] N.C. Fofack, P. Nain, G. Neglia, and D. Towsley. Analysis of TTL-based cache networks. In *Proc. of VALUETOOLS 2012*, pages 1–10, 2012.

[13] N.C. Fofack, P. Nain, G. Neglia, and D. Towsley. Performance evaluation of hierarchical TTL-based cache networks. *Computer Networks*, 65:212–231, 2014.

[14] J.-M. Fourneau, L. Kloul, and D. Verchre. Multiple class g-networks with list-oriented deletions. *European J. of Operational Research*, 126(2):250 – 272, 2000.

[15] J. M. Fourneau and F. Quessette. Computing the steady-state distribution of G-Networks with synchronized partial flushing. In *ISCIS, 21th International Symposium*, pages 887–896, Istanbul, Turkey, 2006.

[16] J. M. Fourneau and D. Verchere. G-networks with triggered batch state-dependent movement. In *MASCOTS '95: Proc. of the Third Int. Workshop on Modeling, Analysis, and Simulation On Computer and Telecommunication Systems*, pages 33–37, 1995.

[17] E. Gelenbe. A unified approach to the evaluation of a class of replacement algorithms. *IEEE Trans. on Computers*, C-22(6):611–618, 1973.

[18] E. Gelenbe. Product form networks with negative and positive customers. *J. of Appl. Prob.*, 28(3):656–663, 1991.

[19] E. Gelenbe. G-networks with triggered customer movement. *J. of Appl. Prob.*, 30:742–748, 1993.

[20] E. Gelenbe and M. Hernández. Optimum checkpoints with age dependent failures. *Acta Inf.*, 27(6):519–531, 1990.

[21] A. Ghodsi, S. Shenker, T. Koponen, A. Singla, B. Raghavan, and J. Wilcox. Information-centric networking: Seeing the forest for the trees. In *Proc. of the 10th ACM Workshop on Hot Topics in Networks*, HotNets-X, pages 1:1–1:6, New York, NY, USA, 2011. ACM.

[22] P. G. Harrison. Turning back time in Markovian process algebra. *Theoretical Computer Science*, 290(3):1947–1986, 2003.

[23] P. G. Harrison and C. M. Lladó. Hierarchically constructed Petri-nets and product-forms. In *5th International ICST Conference on Performance Evaluation Methodologies and Tools Communications, VALUETOOLS '11*, pages 101–110, 2011.

[24] P. G. Harrison and A. Marin. Product-forms in multi-way synchronizations. *Comput. J.*, 57(11):1693–1710, 2014.

[25] J. Hillston. *A Compositional Approach to Performance Modelling*. Cambridge University Press, 1996.

[26] J. Hillston, A. Marin, S. Rossi, and C. Piazza. Contextual lumpability. In *7th International Conference on Performance Evaluation Methodologies and Tools, ValueTools '13*, pages 194–203, 2013.

[27] J. R. Jackson. Jobshop-like queueing systems. *Management Science*, 10:131–142, 1963.

[28] E. Gelenbe J.M. Fourneau. Random neural networks with multiple classes of signals. *Neural Computation*, 11(4):953–963, 1999.

[29] F. Kelly. *Reversibility and stochastic networks*. Wiley, New York, 1979.

[30] W. F. King. Analysis of paging algorithms. In *Proc. of IFIP congress*, pages 485–490, 1971.

[31] A. Marin, S. Balsamo, and P.G. Harrison. Analysis of stochastic Petri nets with signals. *Perf. Eval.*, 85(7):1520–1539, 2012.

[32] A. Marin and S. Rossi. Autoreversibility: exploiting symmetries in Markov chains. In *Proc. of the IEEE 21st Int. Symp. on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS'13)*, pages 151–160, 2013.

[33] A. Marin and S. Rossi. On the relations between lumpability and reversibility. In *Proc. of the IEEE 22nd Int. Symp. on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS'14)*, pages 427–432, 2014.

[34] A. Marin and M.G. Vigliotti. A general result for deriving product-form solutions of Markovian models. In *Proc. of First Joint WOSP/SIPEW Int. Conf. on Perf. Eng.*, pages 165–176, San Josè, CA, USA, 2010. ACM.

[35] V. Martina, M. Garetto, and E. Leonardi. A unified approach to the performance analysis of caching systems. In *Proc. of INFOCOM 2014*, pages 2040–2048, 2014.

[36] B. Melamed and W. Whitt. On arrivals that see time average. *Operations Research*, 38(1), 1990.

[37] S. K. Tripathi, D. Finkel, and E. Gelenbe. Load sharing in distributed systems with failures. *Acta Inf.*, 25(6):677–689, 1988.

[38] N. Tsukada, R. Hirade, and N. Miyoshi. Fluid limit analysis of FIFO and RR caching for independent reference models. *Perf. Eval.*, 69:403–412, 2012.