Annales Mathematicae et Informaticae
36 *(2009) pp. 143–160*
http://ami.ektf.hu

Investigating the mean response time in finite-source retrial queues using the algorithm by Gaver, Jacobs, and Latouche

Patrick Wüchner[a], János Sztrik[b], Hermann de Meer[a]

[a]Faculty of Informatics and Mathematics, University of Passau, Germany

[b]Faculty of Informatics, University of Debrecen, Hungary

Submitted 18 August 2008; Accepted 1 December 2009

Abstract

In this paper, we discuss the maximum of the mean response time that appears in finite-source retrial queues with orbital search when the arrival rate is varied.

We show that explicit closed-form equations of the mean response time can be derived by exploiting the block-structure of the finite Markov chain underlying the model and using an efficient computational algorithm proposed by Gaver, Jacobs, and Latouche.

However, we also show that already for the discussed relatively simple model, the resulting equation is rather complex which hampers further evaluation.

Keywords: Performance evaluation, Finite-source retrial queues, Closed-form solutions, Orbital search, Block-structured Markov chain, MOSEL-2

1. Introduction

Retrial queues are an important field of study, since in various scenarios, they are able to capture certain behavior of real systems more accurately than classical FCFS queues. Retrial queues are used to model, e.g., telephone traffic in [23], load balancing in multiprotocol label switching (MPLS) networks in [19], Ethernet systems in [2], wireless broadband networks in [20], active queue management of Internet routers in [17], self-organizing peer-to-peer systems in [42], the dynamic host configuration protocol (DHCP) in [24], and mobile communication in [3, 32, 34]. Further application examples are given in [6, 22, 43, 15].

For example, consider a call center scenario with several agents and without a waiting loop installed. If all agents are busy, an additional caller is not able to join a queue, but has to hang up and retry to reach an agent later. Such a retrying caller is said to be in *orbit*.

In addition, consider a call center that is able to log the phone numbers of unserved customers. Then, if an agent gets idle, it may call back unserved orbiting customers. This behavior is called *orbital search*.

In many situations it is unrealistic to assume that the calling population, i.e., the potential number of customers generating requests, is infinitely large. Then, the arrival rate of incoming requests depends on the number of requests already in the system and the arrival process is quasi-random, state-dependent, and non-Poisson. Retrial queues with a finite population size are also known as *finite-source* retrial queues. We are especially interested in models where infinite-source models fail, i.e., models with a small number of sources.

During evaluation of finite-source retrial queues, for some parameter setups a maximum of the mean response time of the system can be identified. Several publications noticed this maximum (e.g., [27, 4, 5, 40]) and gave informal reasons for it (e.g., [40]). Since this maximum should be avoided in real-system configurations by all means, we here try to provide closed-form equations that facilitate the identification of such undesirable configurations during system design.

Our main contribution is the development of novel and explicit closed-form equations for steady-state performance evaluation of the mean response time in finite-source retrial queues with orbital search. To achieve this, we adopt an algorithm introduced by Gaver, Jacobs, and Latouche in [29], which we refer to as *GJL Algorithm*.

The main motivation for this research was to find exact mathematical expressions of the maximum's location in closed form. However, as is discussed later in Section 7, this cannot be achieved, even for the relatively simple model under study.

Previous results on various types of retrial queues are surveyed in [6, 15, 7, 8, 26, 28, 31].

Due to the complexity of finite-source and infinite-source retrial queues, publications on performance measures in closed form are quite rare. Instead, most publications, e.g., the more recent ones [34, 15, 5, 38, 21, 10, 11, 12, 13, 35], employ algorithmic or numerical analysis.

The search for customers immediately on termination of a service was first discussed in the context of classical queues by [33]. More recently, infinite-source retrial queueing systems where the server(s) search for customers after service have been investigated in [21, 16, 25]. We recently introduced and discussed finite-source retrial queues with orbital search by applying numerical analysis in [40] and [41].

There exist several publications discussing infinite-source retrial queues without orbital search and presenting exact results (e.g., [17, 15, 28, 1, 9, 14, 30, 36]), or approximations (e.g., [17, 3, 32, 15, 21]) of performance measures in closed form. Regarding finite-source retrial queues without orbital search, [2] presents closed-form results including phase-type service and multiple servers. However, we are not aware of any publications that present steady-state probabilities and performance measures in closed form applicable to finite-source retrial queues with orbital search.

The remainder of this paper is structured as follows. In Section 2, the investigated model is introduced to fix the notations and preliminary numerical results are presented to state the tackled problem in more detail. Starting from Section 3, we exemplarily focus on the case of three sources and one server. Section 3 discusses the underlying continuous-time Markov chain, and in Section 4, the GJL Algorithm is applied to obtain the steady-state probabilities of the Markov chain in closed form. Based on these equations, in Section 5, mean response time is obtained in closed form and validated in Section 6. In Section 7, we discuss the presented approach with respect to the failure of providing closed-form equations of the maximum's location, and its applicability to derive further performance measures in closed form and for models with a higher number of sources, multiple servers, and phase-type distributed service times. Finally, in Section 8, a conclusion and directions for future work are given.

2. Model description and preliminary numerical analysis

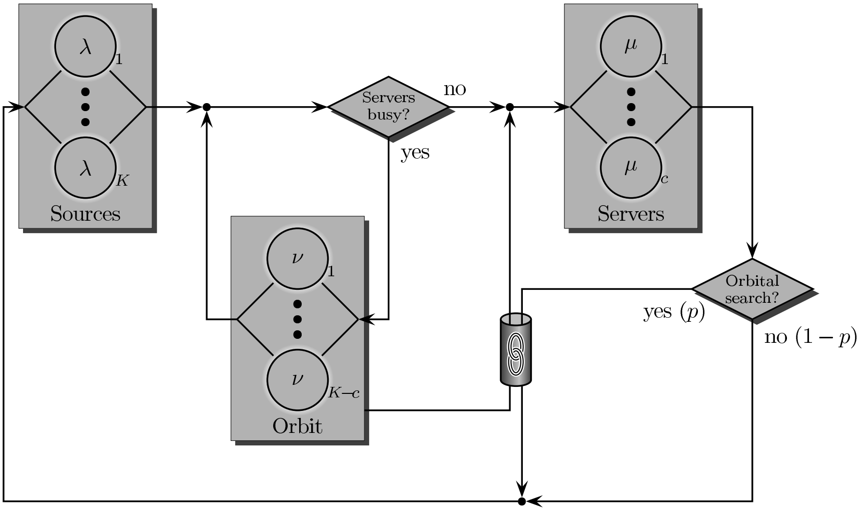

Figure 1: High-level queueing model of finite-source retrial queue with orbital search.

In Fig. 1, a queueing model illustrates the $M/M/c/K/K$ (for Kendall's notation, see [18, p. 242]) finite-source retrial queue with orbital search. All inter-event times involved in the model are assumed to be exponentially distributed. Model extensions by including phase-type distributions are discussed in Section 7.

Each of the K sources is generating primary requests to the retrial queue with rate λ as long as the source is not waiting for a response to an active (i.e., in service or orbiting) request. A primary request first checks whether an idle server is available. If all c identical servers are busy, the primary request enters the orbit instead and retries to get service with rate ν. If a request finds at least one server idle, it starts to receive service with service rate μ. After being serviced, a response is returned to the requesting source. With a probability of p, where $0 \leqslant p \leqslant 1$, at service completion instant, the server carries out orbital search and instantly fetches a request, if available, directly from the orbit (denoted by the link symbol).

The finite-source retrial queue with orbital search can be evaluated numerically quite easily by using the MOSEL-2 performance evaluation tool. The corresponding MOSEL-2 model is shown in Listing 1. The interested reader is referred to [39] for a short introduction to MOSEL-2. In [40, 41] similar models and discussion of MOSEL-2's scalability are presented in the context of finite-source retrial queues.

```
/*** CONSTANTS AND PARAMETERS ************************************************/
CONST      K      := 3;                                  // population size
CONST      mu     := 1;                                  // service rate
PARAMETER lambda := 0.0001, 0.1 .. 1 STEP 0.1;          // request gen. rate
PARAMETER nu     := 0.001, 0.0025, 0.005;               // retrial rate
PARAMETER p      := 1E-8, 0.5, 1-1E-8;                  // search probability

/*** NODES *******************************************************************/
NODE Sources[K]   := K;                                  // the sources
NODE Request[1]   := 0;                                  // primary requests
NODE Server[1]    := 0;                                  // the server
NODE Orbit[K]     := 0;                                  // the orbit
NODE Finished[1]  := 0;                                  // response

/*** RULES *******************************************************************/
FROM Sources    TO Request   RATE Sources*lambda;        // primary requests
FROM Request    TO Server    PRIO 1;                     // to server if idle
FROM Request    TO Orbit     PRIO 0;                     // to orbit  if busy
FROM Orbit      TO Server    RATE Orbit*nu;              // retrials
FROM Server     TO Finished  RATE mu;                    // service
FROM Finished   TO Sources   WEIGHT 1-p;                 // without orb. search
FROM Finished, Orbit TO Server, Sources WEIGHT p;        // with orbital search

/*** RESULTS *****************************************************************/
PRINT rho         := UTIL(Server);                       // server utilization
PRINT M           := MEAN(Orbit)+MEAN(Server);           // mean # active req.
PRINT S           := K-M;                                // mean # active sources
PRINT N           := MEAN(Orbit);                        // mean # orbit
PRINT ml          := S*lambda;                           // mean throughput
PRINT T           := M/ml;                               // mean response time
PRINT To          := N/ml;                               // mean orbit time
PRINT R           := nu*To;                              // mean # retrials
```

Listing 1: MOSEL-2 model of finite-source retrial queue with orbital search.

Fig. 2 shows the mean response time $\overline{T}$ as a function of request generation rate λ for $K = 3$ sources, $c = 1$ server, service rate $\mu = 1$. We chose different values of retrial rate ν and orbital-search probability p. The curves labeled "num" are obtained by using MOSEL-2's numerical analysis.

These results show a maximum of the mean response time but they are not detailed enough to estimate the exact location of the maximum (in the following denoted as λ_{peak}). This is achieved more accurately by the dashed curves (labeled "expl") which are, in fact, derived using the closed-form equations developed in

Section 5. Hence, in the following, we aim at finding an explicit equation for the mean response time $\overline{T}$ as a function of λ and further model parameters. Afterwards, we discuss whether this equation can be differentiated with respect to λ and whether is is possible to find the roots of the derivation which would lead to an explicit equation of λ_{peak}.

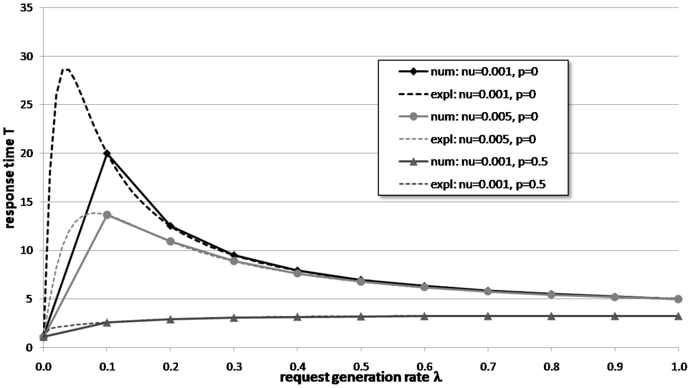

Figure 2: Mean response time $\overline{T}$ over request generation rate λ for service rate $\mu = 1$ and different values of retrial rate ν and orbital-search probability p.

3. Underlying Markov chain

Theorem 3.1. *The behavior of the finite-source retrial queue with orbital search as described in Section 2 can be modeled by a bivariate continuous-time, finite-state Markov chain (CTMC) with state variable $X(t) = (N(t), C(t))$, where variable $N(t)$ is the number of customers in the orbit and variable $C(t)$ is the number of busy servers at time $t \geqslant 0$. Furthermore, this CTMC has a unique steady-state distribution $\pi(i, j)$, with $i = 0, \dots, K - c$, and $j = 0, \dots, c$.*

Proof. Due to the memoryless property of the solely exponentially distributed inter-event times, the sojourn times of $X(t)$ are also exponentially distributed, hence the process is a Markov chain.

It is easy to see that $X(t)$ has a finite number of states and is irreducible for all reasonable (i.e., strictly positive) values of λ, μ, and ν. Hence, the underlying stochastic process is positive recurrent which also implies ergodicity. Ergodicity again implies the existence and uniqueness of steady-state probabilities (see [18, p. 69–70]). □

Note that the order of the variables $N(t)$ and $C(t)$ within $X(t)$ is chosen to reflect the structure (levels and phases) of the underlying Markov chain.

In the following, we restrict our investigation to the case $K = 3$ and $c = 1$ to preserve conciseness and traceability. Directions for $K > 3$ and $c > 1$ are given in Section 7. The state transition diagram of the corresponding CTMC is shown in Fig. 3.

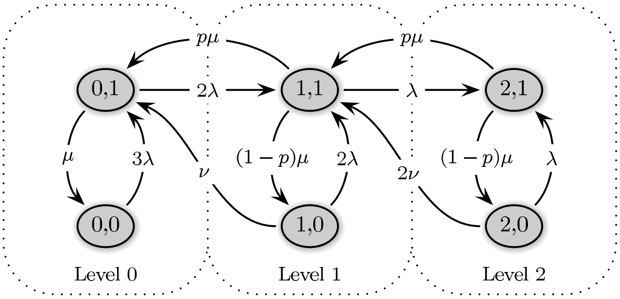

Figure 3: State transition diagram of finite-source retrial queue with orbital search for $K = 3$ and $c = 1$.

Note that for $p = 0$, Fig. 3 reduces to the state transition diagram of the classical $M/M/1/3/3$ finite-source retrial queue without orbital search. On the other hand, for $p = 1$, Fig. 3 reduces to the state transition diagram of the classical $M/M/1/3/3$–First-Come-First-Served (FCFS) queue.

The CTMC shown in Fig. 3 can be structured according to levels reflecting the number of customers in the orbit $N(t)$. Each level consists of two phases indicating the state of the server given by $C(t)$. Moreover, the CTMC constitutes a finite quasi-birth-death process (QBD), which is skip-free in both directions. This structure is also reflected in the block-tridiagonal form of the infinitesimal generator matrix $\mathbf{Q}$ of the CTMC given by

$$\mathbf{Q} = \begin{pmatrix} \mathbf{A}^{(0)} & \mathbf{\Lambda}^{(0)} & \mathbf{0} \\ \mathbf{M}^{(1)} & \mathbf{A}^{(1)} & \mathbf{\Lambda}^{(1)} \\ \mathbf{0} & \mathbf{M}^{(2)} & \mathbf{A}^{(2)} \end{pmatrix}, \tag{3.1}$$

where the sub-matrices

$$\begin{aligned} \mathbf{A}^{(0)} &= \begin{pmatrix} -3\lambda & 3\lambda \\ \mu & -2\lambda-\mu \end{pmatrix}, & \mathbf{\Lambda}^{(0)} &= \begin{pmatrix} 0 & 0 \\ 0 & 2\lambda \end{pmatrix}, \\ \mathbf{A}^{(1)} &= \begin{pmatrix} -2\lambda-\nu & 2\lambda \\ (1-p)\mu & -\lambda-\mu \end{pmatrix}, & \mathbf{\Lambda}^{(1)} &= \begin{pmatrix} 0 & 0 \\ 0 & \lambda \end{pmatrix}, \\ \mathbf{A}^{(2)} &= \begin{pmatrix} -\lambda-2\nu & \lambda \\ (1-p)\mu & -\mu \end{pmatrix}, & \mathbf{M}^{(1)} &= \begin{pmatrix} 0 & \nu \\ 0 & p\nu \end{pmatrix}, \\ \mathbf{0} &= \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, & \mathbf{M}^{(2)} &= \begin{pmatrix} 0 & 2\nu \\ 0 & p\nu \end{pmatrix}, \end{aligned}$$

can be obtained by inspecting the transition rates given in Fig. 3. Note that the notation of the sub-matrices is chosen in accordance to [29].

4. Application of the GJL algorithm

To obtain the steady-state probabilities in closed form, we apply the computational algorithm proposed in [29]. For this, we exploit the relatively simple structure of the underlying Markov chain as presented in Section 3. For a thorough explanation and proof of the GJL algorithm, we refer the interested reader to [29].

The GJL Algorithm is applied to the structured CTMC (i.e., finite QBD) given in Sect. 3, where $K = 3$:

1. Calculation of $\mathbf{C}_n$ with $0 \leqslant n \leqslant 2$:

$$\mathbf{C}_0 = \mathbf{A}^{(0)} = \begin{pmatrix} -3\lambda & 3\lambda \\ \mu & -2\lambda - \mu \end{pmatrix}, \tag{4.1}$$

$$\begin{aligned} \mathbf{C}_1 &= \mathbf{A}^{(1)} + \mathbf{M}^{(1)} \left(-\mathbf{C}_0^{-1} \boldsymbol{\Lambda}^{(0)}\right) \\ &= \begin{pmatrix} -2\lambda - \nu & 2\lambda + \nu \\ (1-p)\mu & -\lambda - (1-p)\mu \end{pmatrix}, \end{aligned} \tag{4.2}$$

$$\begin{aligned} \mathbf{C}_2 &= \mathbf{A}^{(2)} + \mathbf{M}^{(2)} \left(-\mathbf{C}_1^{-1} \boldsymbol{\Lambda}^{(1)}\right) \\ &= \begin{pmatrix} -\lambda - 2\nu & \lambda + 2\nu \\ (1-p)\mu & -(1-p)\mu \end{pmatrix}. \end{aligned} \tag{4.3}$$

2. Obtaining $\boldsymbol{\pi}_2$: Since the system $\boldsymbol{\pi}_2 \mathbf{C}_2 = (0, 0)$ is linearly dependent, we can replace one equation of the system by the normalization condition $\boldsymbol{\pi}_2 \binom{1}{1} = 1$ and solve

$$\boldsymbol{\pi}_2 \begin{pmatrix} -\lambda - 2\nu & 1 \\ (1-p)\mu & 1 \end{pmatrix} = (0, 1), \tag{4.4}$$

instead. This leads to

$$\boldsymbol{\pi}_2 = \tfrac{1}{\lambda + (1-p)\mu + 2\nu} \left((1-p)\mu \;\, \lambda + 2\nu\right). \tag{4.5}$$

3. Obtaining $\mathbf{P}_n$, $n = 2, 1, 0$, recursively:

$$\begin{aligned} \mathbf{P}_2 &= \boldsymbol{\pi}_2 \\ &= \tfrac{1}{\lambda + (1-p)\mu + 2\nu} \left((1-p)\mu \;\, \lambda + 2\nu\right), \end{aligned} \tag{4.6}$$

$$\begin{aligned} \mathbf{P}_1 &= \mathbf{P}_2 \mathbf{M}^{(2)} \left(-\mathbf{C}_1^{-1}\right) \\ &= \tfrac{1}{\lambda + (1-p)\mu + 2\nu} \\ &\quad \cdot \left(\tfrac{(1-p)\mu^2 (\lambda p + 2\nu)}{\lambda (2\lambda + \nu)} \;\; \tfrac{\mu (\lambda p + 2\nu)}{\lambda} \right), \end{aligned} \tag{4.7}$$

$$\begin{aligned} \mathbf{P}_0 &= \mathbf{P}_1 \mathbf{M}^{(1)} \left(-\mathbf{C}_0^{-1}\right) \\ &= \tfrac{1}{\lambda + (1-p)\mu + 2\nu} \\ &\quad \cdot \left(\tfrac{\mu^3 (\lambda p + 2\nu)(2\lambda p + \nu)}{6\lambda^3 (2\lambda + \nu)} \;\; \tfrac{\mu^2 (2\nu + \lambda p)(2\lambda p + \nu)}{2\lambda^2 (2\lambda + \nu)} \right). \end{aligned} \tag{4.8}$$

4. Re-normalizing vector $\mathbf{P} = (\mathbf{P}_0, \mathbf{P}_1, \mathbf{P}_2)$: For this, we derive the normalization constant P_N as follows:

$$\begin{aligned} P_N &= \mathbf{P}_0\tbinom{1}{1} + \mathbf{P}_1\tbinom{1}{1} + \mathbf{P}_2\tbinom{1}{1} \\ &= \tfrac{1}{\lambda+(1-p)\mu+2\nu}\,\tfrac{1}{6\lambda^3(2\lambda+\nu)}\widetilde{P_N}, \end{aligned} \tag{4.9}$$

where $\widetilde{P_N}$ is given by the term

$$\begin{aligned} \widetilde{P_N} = 2\mu^3\nu^2 &+ 5\mu^3\nu\lambda p + 2\mu^3\lambda^2 p^2 + 6\mu^2\lambda\nu^2 \\ &+ 3\mu^2\lambda^2\nu p + 12\mu^2\lambda^2\nu + 6\mu^2\lambda^3 p\nu \\ &+ 30\mu\lambda^3 + 12\mu\lambda^2\nu^2 + 12\mu\lambda^4 + 12\lambda^5 \\ &+ 30\lambda^4\nu + 12\lambda^3\nu^2. \end{aligned} \tag{4.10}$$

In the following, we denote by $P_{i,j}, i \in \{0,1,2\}, j \in \{0,1\}$, the j-th element of vector $\mathbf{P}_i$ and with $\pi(i,j)$ the steady-state probability of state (i,j), i.e., phase j in level i. With $\mathbf{P}_i$ given by Eqs. (4.6) through (4.8) and P_N given by Eq. (4.9), the desired steady-state probabilities of the CTMC depicted in Fig. 3 can be derived in closed form as follows:

$$\pi(0,0) = \frac{P_{0,0}}{P_N} = \frac{\mu^3(\lambda p + 2\nu)(2\lambda p + \nu)}{\widetilde{P_N}}, \tag{4.11}$$

$$\pi(0,1) = \frac{P_{0,1}}{P_N} = \frac{3\lambda\mu^2(\lambda p + 2\nu)(2\lambda p + \nu)}{\widetilde{P_N}}, \tag{4.12}$$

$$\pi(1,0) = \frac{P_{1,0}}{P_N} = \frac{6\lambda^2(1-p)\mu^2(\lambda p + 2\nu)}{\widetilde{P_N}}, \tag{4.13}$$

$$\pi(1,1) = \frac{P_{1,1}}{P_N} = \frac{6\lambda^2\mu(\lambda p + 2\nu)(2\lambda + \nu)}{\widetilde{P_N}}, \tag{4.14}$$

$$\pi(2,0) = \frac{P_{2,0}}{P_N} = \frac{6\lambda^3(1-p)\mu(2\lambda + \nu)}{\widetilde{P_N}}, \tag{4.15}$$

$$\pi(2,1) = \frac{P_{2,1}}{P_N} = \frac{6\lambda^3(\lambda + 2\nu)(2\lambda + \nu)}{\widetilde{P_N}}. \tag{4.16}$$

5. Mean response time in closed form

In Section 4, closed-form expressions of the steady-state probabilities of the underlying Markov chain were derived. These expressions are now used to obtain the mean response time $\overline{T}$ in closed form.

Mean number of active requests $\overline{M}$: The mean number of requests located in service or in orbit is given by

$$\overline{M} = \pi(0,1) + \pi(1,0) + 2\pi(1,1) + 2\pi(2,0)$$

$$\begin{aligned}&+3\pi(2,1)\\ &=\frac{3\lambda}{\widetilde{P_N}}(\mu^2\lambda p\nu+2\mu^2\nu^2+2\lambda^2\mu^2 p+4\lambda\mu^2\nu\\ &+20\lambda^2\mu\nu+8\lambda\mu\nu^2+8\lambda^3\mu+12\lambda^4\\ &+30\lambda^3\nu+12\lambda^2\nu^2).\end{aligned}\tag{5.1}$$

Mean system throughput $\overline{\lambda}$: The mean throughput of the finite-source retrial queue with orbital search can be obtained from

$$\begin{aligned}\overline{\lambda}&=(K-\overline{M})\lambda\\ &=3\lambda\bigg(1-\frac{\lambda}{\widetilde{P_N}}(\mu^2\lambda p\nu+2\mu^2\nu^2+2\lambda^2\mu^2 p\\ &\quad+4\lambda\mu^2\nu+20\lambda^2\mu\nu+8\lambda\mu\nu^2+8\lambda^3\mu\\ &\quad+12\lambda^4+30\lambda^3\nu+12\lambda^2\nu^2)\bigg)\\ &=\frac{3\lambda\widetilde{\Lambda}}{\widetilde{P_N}},\end{aligned}\tag{5.2}$$

where $\widetilde{\Lambda}$ is defined as follows:

$$\begin{aligned}\widetilde{\Lambda}=\mu(&2\mu^2\lambda^2p^2+5\mu^2\lambda p\nu+2\mu^2\nu^2+4\lambda^3\mu p\\ &+2\lambda^2\mu p\nu+8\lambda^2\mu\nu+4\lambda\mu\nu^2+4\lambda^4\\ &+10\lambda^3\nu+4\lambda^2\nu^2).\end{aligned}\tag{5.3}$$

Mean response time $\overline{T}$: The mean time spent by each request in the orbit and the server can also be calculated by applying Little's Law as follows:

$$\begin{aligned}\overline{T}&=\frac{\overline{M}}{\overline{\lambda}}\\ &=\frac{1}{\widetilde{\Lambda}}(\mu^2\lambda p\nu+2\mu^2\nu^2+2\lambda^2\mu^2p+4\lambda\mu^2\nu\\ &\quad+12\lambda^4+20\lambda^2\mu\nu+8\lambda\mu\nu^2+8\lambda^3\mu\\ &\quad+30\lambda^3\nu+12\lambda^2\nu^2)\end{aligned}\tag{5.4}$$

In Fig. 4, we exemplarily plot the mean response time $\overline{T}$ as a function of request generation rate λ and retrial rate ν for service rate $\mu = 1$ and orbital-search probability $p = 0.5$ by employing Eq. (5.4). The presented closed-form equations facilitate the retrieval of fine-grained results since, in general, they can be implemented more efficiently than numerical analysis.

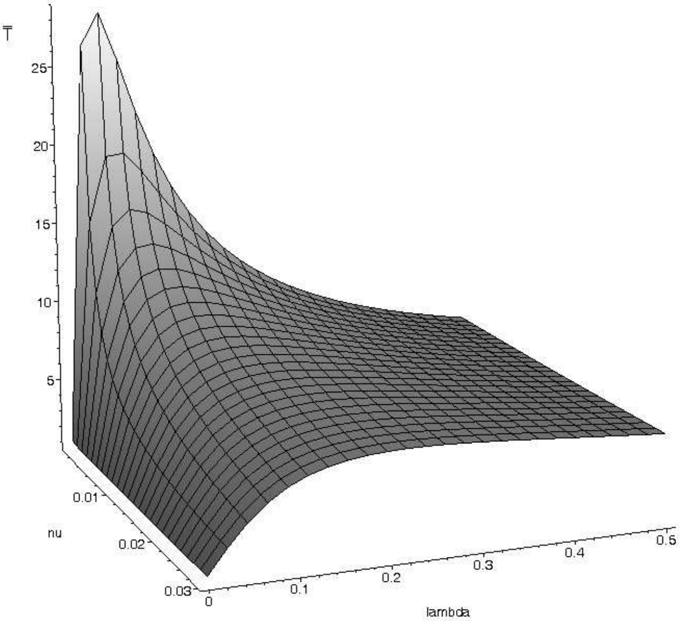

Figure 4: Mean response time $\overline{T}$ (z-axis) over request generation rate λ (x-axis) and retrial rate ν (y-axis) for service rate $\mu = 1$ and orbital-search probability $p = 0.5$.

6. Validation of closed-form equations

In this section, the closed-form equations derived in Section 5 are validated against numerical results and against well-known closed-form equations of $M/M/1/K/K$–$FCFS$ queueing systems.

6.1. Comparison to numerical results

Table 1 compares results obtained from numerical analysis using MOSEL-2 (see Section 2) to results obtained by using the closed-form expressions presented in Section 5 for $\lambda = 0.1$, $\nu = 0.0025$, $\mu = 1$, and $p = 0.5$. It can be seen that the numerical results are very close to the closed-form results.

6.2. Comparison to $M/M/1/K/K$–$FCFS$ system

As already mentioned in Section 3, the state-transition diagram given in Fig. 3 takes the form of the CTMC underlying an $M/M/1/3/3$–$FCFS$ queueing system for $p = 1$. Such finite-population FCFS systems are also known as *Machine Repairman Models* (see [18, p. 252]), for which performance measures are available in

Perf. Measure	Num. Analysis	Closed-Form Expression
ρ	0.239555	0.2395547133
$\overline{M}$	0.604453	0.6044528670
$\overline{S}$	2.39555	2.395547133
$\overline{N}$	0.364898	0.3648981538
$\overline{\lambda}$	0.239555	0.2395547133
$\overline{T}$	2.52324	2.523235126
$\overline{T}_O$	1.52324	1.523235126
$\overline{R}$	0.00380809	0.003808087814

Table 1: Model results for $\lambda = 0.1$, $\nu = 0.0025$, $\mu = 1$, and $p = 0.5$.

closed form.

According to [18], the mean response time $\overline{T}$ of an $M/M/1/K/K$–$FCFS$ queue is given by

$$\overline{T}_{\text{FCFS}} = \frac{K}{\mu(1-\pi_0)} - \frac{1}{\lambda}, \tag{6.1}$$

where the steady-state probability of an idle server π_0 is given by

$$\pi_0 = \frac{1}{\sum\limits_{k=0}^{K} \left(\frac{\lambda}{\mu}\right)^k \frac{K!}{(K-k)!}}. \tag{6.2}$$

In the current scenario, where $K = 3$, Eq. (6.1) can be rewritten as

$$\begin{aligned}\overline{T}_{\text{FCFS}} &= \frac{3}{\mu(1-\pi_0)} - \frac{1}{\lambda} \\ &= \frac{6\lambda^2 + 4\mu\lambda + \mu^2}{(\lambda^2 + 2\mu\lambda + \mu^2)\mu}.\end{aligned} \tag{6.3}$$

When setting $p = 1$ in Eq. (5.4), we equivalently get

$$\begin{aligned}\overline{T} = \Big(&5\mu^2\lambda\nu + 2\mu^2\nu^2 + 2\mu^2\lambda^2 + 20\lambda^2\mu\nu \\ &+ 8\lambda\mu\nu^2 + 8\mu\lambda^3 + 12\lambda^4 + 30\lambda^3\nu + 12\lambda^2\nu^2\Big) \\ &\Big/ \Big(\mu\big(5\mu^2\lambda\nu + 2\mu^2\lambda^2 + 2\mu^2\nu^2 + 4\mu\lambda^3 \\ &+ 10\lambda^2\mu\nu + 4\lambda\mu\nu^2 + 4\lambda^4 + 10\lambda^3\nu \\ &+ 4\lambda^2\nu^2\big)\Big) \\ = \;& \frac{6\lambda^2 + 4\mu\lambda + \mu^2}{(\lambda^2 + 2\mu\lambda + \mu^2)\mu} \\ = \;& \overline{T}_{\text{FCFS}}.\end{aligned} \tag{6.4}$$

Hence, the two closed forms match in the case of $p = 1$. For the sake of completeness, we compare the mean response time $\overline{T}_p$ of an $M/M/1/3/3$ retrial queue with orbital search ($\mu = 1$, $\nu = 0.001$, $p = 0.1 \dots 0.9$) with the mean response time $\overline{T}_{\text{FCFS}}$ of an $M/M/1/3/3$–$FCFS$ queue ($\mu = 1$) in Fig. 5.

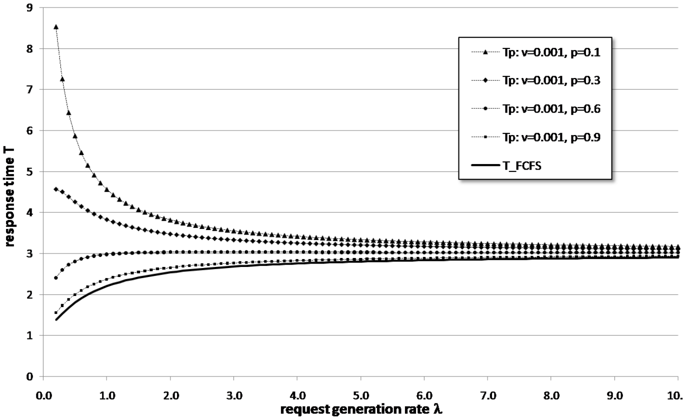

Figure 5: Mean response times $\overline{T}_p$ and $\overline{T}_{\text{FCFS}}$ over request generation rate λ.

As expected, $\overline{T}_p$ gets close to $\overline{T}_{\text{FCFS}}$ for $p \approx 1$. It can also be seen that all curves get close to each other for high values of the request generation rate λ. High generation rates lead to high server utilization. The server is then kept busy by primary requests even if the orbital search probability p is low. Also for high values of ν (compared to μ), the behavior of finite-source retrial queues with orbital search should be close to the behavior of an $M/M/1/K/K$–$FCFS$ queueing system. This statement is confirmed by Fig. 6, where the mean response times $\overline{T}_\nu$ of an $M/M/1/3/3$ retrial queue with orbital search ($\mu = 1$, $\nu = 0.1 \dots 20$, $p = 0.1$), and $\overline{T}_{\text{FCFS}}$ of the $M/M/1/3/3$–$FCFS$ queue ($\mu = 1$) are compared.

It can be seen that for high values of ν, $\overline{T}_\nu$ gets close to $\overline{T}_{\text{FCFS}}$. Again, for high server utilization, $\overline{T}_\nu$ becomes independent of ν.

Note that all results of Figs. 5 and 6 are obtained using Eqs. (5.4) and (6.3).

7. Discussion of approach

7.1. Location of maximum

To find an equation for λ_{peak}, i.e., the arrival rate of the maximum mean response time, in closed form, we need to find the roots of equation $\frac{d\overline{T}}{d\lambda}$. However, according to Eq. (5.4), $\overline{T}$ is quite complex already for this simple model. The resulting equation derived from $\frac{d\overline{T}}{d\lambda}$ is a ratio of high order polynomials for which the roots could be found numerically, but unfortunately not in a closed form.

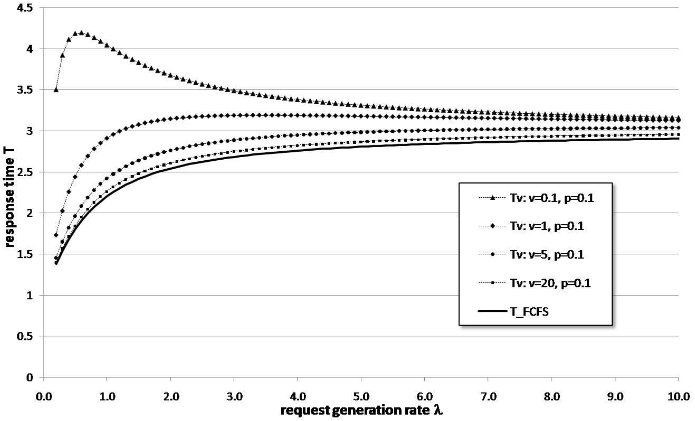

Figure 6: Mean response times $\overline{T}_\nu$ and $\overline{T}_{\mathrm{FCFS}}$ over request generation rate λ.

7.2. Further performance measures

Unfortunately, our forseen goal to provide a closed-form equation for the maximum's location cannot be achieved. However, by using the steady-state probabilities presented in Section 4, further performance measures of the discussed retrial queue can be derived in closed form.

For example, Eqs. (4.11) through (4.16) can be readily used together with the equations provided in our previous work [40, Sec. 2.3] to obtain steady-state performance measures like the server utilization, the mean number of orbiting customers, the mean waiting time, etc. in closed form.

7.3. Model generalization

While in Sections 3 through 6, for the sake of clearness, the investigation is restricted to $K = 3$ and $c = 1$ to show the principles, we now discuss the applicability of the method for a higher number of sources and servers as well as phase-type service.

7.3.1. Increasing the number of sources

The GJL Algorithm employed in Section 4 can be applied in principle also for higher values of K. If K is increased, the number of levels of the underlying CTMC (recall Fig. 3) increases, but the number of phases in each level stays the same, i.e., two. As a consequence, matrix $\mathbf{Q}$ (recall Eq. (3.1)) will grow by one additional column and one additional row of 2×2 sub-matrices per each additional source. The number of the matrices $\mathbf{C}_n$ increases ($0 \leqslant n \leqslant K-1$) but not their size ($2 \times 2$). This results in additional iteration steps in Steps 1 and 3 of the GJL Algorithm but the matrices $\mathbf{C}_n$ can still be inverted explicitly in a relatively compact way.

7.3.2. Increasing the number of servers

If the number of servers is increased, then also the size of square matrices $\mathbf{C}_n$ increases. By using, e.g., Eq. (7.1) (cf. [37]):

$$\mathbf{C}_n^{-1} = \frac{1}{det(\mathbf{C}_n)} adj(\mathbf{C}_n), \tag{7.1}$$

the matrices $\mathbf{C}_n$ can still be inverted explicitly. This, however, increases the effort and leads to even more complex closed-form equations.

7.3.3. Increasing the number of service phases

The method can also be used in case of a single server which conducts phase-type service with a finite number of service phases. Comparable to Section 7.3.2, this results in additional phases within the Markov chain and in larger $\mathbf{C}_n$ matrices, which can still be inverted explicitly. The proposed method cannot be applied directly to multiple-server retrial queues with phase-type service, since this implies higher-dimensional Markov chains.

8. Conclusion and future work

In this paper, we present steady-state probabilities and the mean response time of single-server finite-source retrial queues with orbital search and three sources in closed form. The equations are derived by adopting an algorithm introduced in [29]. The results are validated against results obtained by numerical analysis and against closed-form equations well-known for $M/M/1/K/K\text{–}FCFS$ queueing systems.

It could be shown that due to the high complexity of the derived equations, it is not possible to derive the location of the mean response time's maximum in closed form. However, using the derived closed-form equations of the steady-state probabilities gives raise to other interesting performance measures in closed-form as well.

Our planned future work includes applying the algorithm to a higher number of sources and servers, phase-type service, and unreliable servers. It may also be worthwhile to study approximate solutions for higher numbers of sources, servers, and service phases.

9. Acknowledgments

This research is partially supported by the German-Hungarian Intergovernmental Scientific Cooperation, HAS-DFG, 436 UNG 113/197/0-1, by the Hungarian Scientific Research Fund, OTKA K60698/2006, by the AutoI project (STREP, FP7 Call 1, ICT-2007-1-216404), by the ResumeNet project (STREP, FP7 Call 2, ICT-2007-2-224619), by the SOCIONICAL project (IP, FP7 Call 3, ICT-2007-3-231288), and by the EuroNF Network of Excellence (FP7, IST 216366).

The first author is especially grateful to the participants of the Dagstuhl Seminar on “Numerical Methods for Structured Markov Chains” (07461) for very fruitful discussions.

References

[1] Aissani, A., Artalejo, J., “On the single server retrial queue subject to breakdowns,” *Queueing Systems*, vol. 30, (1998) 309–321.

[2] Alfa, A.S., Isotupa, K.S., “An M/PH/k retrial queue with finite number of sources,” *Computers and Operations Research*, vol. 31, (2004) 1455–1464.

[3] Alfa, A.S., Li, W., “PCS networks with correlated arrival process and retrial phenomenon,” *IEEE Transactions on Wireless Communications*, vol. 1, no. 4, (October 2002) 630–637.

[4] Almasi, B., Bolch, G., Sztrik, J., “Heterogeneous finite-source retrial queues,” University of Erlangen-Nuremberg, Erlangen, Germany, Tech. Rep. TR-I4-02-04, 2004.

[5] Almasi, B., Roszik, J., Sztrik, J., “Homogeneous finite-source retrial queues with server subject to breakdowns and repairs,” *Mathematical and Computer Modelling*, vol. 42, (2005) 673–682.

[6] Artalejo, J.R., “Retrial queues with a finite number of sources,” *J. Korean Math. Soc.*, vol. 35, (1998) 503–525.

[7] Artalejo, J.R., “Accessible bibliography on retrial queues,” *Mathematical and Computer Modelling*, vol. 30, (1999) 1–6.

[8] Artalejo, J.R., “A classified bibliography of research on retrial queues: Progress in 1990-1999,” *TOP*, vol. 7, (1999) 187–211.

[9] Artalejo, J.R., “Stationary analysis of the characteristics of the M/M/2 queue with constant repeated attempts,” *Opsearch*, vol. 33, no. 2, (1996) 83–95.

[10] Artalejo, J.R., Chakravarthy, S.R., “Computational analysis of the maximal queue length in the MAP/M/c retrial queue,” *Applied Mathematics and Computation*, vol. 183, (2006) 1399–1409.

[11] Artalejo, J.R., Chakravarthy, S.R., “Algorithmic analysis of the maximum level length in general-block two-dimensional Markov processes,” *Mathematical Problems in Engineering*, vol. 2006, (2006) 1–15.

[12] Artalejo, J.R., Chakravarthy, S.R., “Algorithmic analysis of the MAP/PH/1 retrial queue,” *TOP*, vol. 14, no. 2, (2006) 293–332.

[13] Artalejo, J.R., Chakravarthy, S.R., Lopez-Herrero, M.J., “The busy period and the waiting time analysis of a MAP/M/c queue with finite retrial group,” *Stochastic Analysis and Applications*, vol. 25, (2007) 445–469.

[14] Artalejo, J.R., Economou, A., Lopez-Herrero, M.J., “Algorithmic analysis of the maximum queue length in a busy period for the M/M/c retrial queue,” *INFORMS J. on Computing*, vol. 19, no. 1, (2007) 121–126.

[15] Artalejo, J.R., Gómez-Corral, A., *Retrial Queueing Systems: A Computational Approach.* Springer Verlag, 2008.

[16] Artalejo, J.R., Joshua, V.C., Krishnamoorthy, A., "An M/G/1 retrial queue with orbital search by the server," in *Advances in Stochastic Modelling*, J. R. Artalejo and A. Krishnamoorthy, Eds. NJ: Notable Publications Inc., (2002) 41–54.

[17] Avrachenkov, K., Yechiali, U., "Retrial networks with finite buffers and their application to internet data traffic," *Probability in the Engineering and Informational Sciences*, vol. 22, (2008) 519–536.

[18] Bolch, G., Greiner, S., Meer, H., Trivedi, K., *Queueing Networks and Markov Chains*, 2nd ed. New York: John Wiley & Sons, 2006.

[19] Chakka, R., Do, T.V., "The $MM\sum_{k=1}^{K} CPP_k/GE/c/L$ G-Queue and Its Application to the Analysis of the Load Balancing in MPLS Networks." in *Proc. of 27th Annual IEEE Conference on Local Computer Networks (LCN 2002), 6-8 November 2002*, Tampa, FL, USA, (2002) 735–736.

[20] Chakka, R., Do, T.V., "The MM $\sum_{k=1}^{K} CPP_k/GE/c/L$ G-queue with heterogeneous servers: Steady state solution and an application to performance evaluation," *Performance Evaluation*, vol. 64, (March 2007) 191–209.

[21] Chakravarthy, S.R., Krishnamoorthy, A., Joshua, V., "Analysis of a multiserver retrial queue with search of customers from the orbit," *Performance Evaluation*, vol. 63, no. 8, (2006) 776–798.

[22] Choi, B.D., Chang, Y., "Single server retrial queues with priority calls," *Mathematical and Computer Modelling*, vol. 30, (1999, invited paper) 7–32.

[23] Cohen, J.W., "Basic problems of telephone traffic theory and the influence of repeated calls," *Philips Telecommun. Rev.*, vol. 18, no. 2, (1957) 49–100.

[24] Do, T.V., "An efficient solution to a retrial queue for the performability evaluation of DHCP," *Computers and Operations Research*, vol. In Press, Corrected Proof, (2009) [Online]. Available: `http://www.sciencedirect.com/science/article/B6VC5-4WGVPXN-1/2/48b8e2e8550847fd28abec83ff41f72d`

[25] Dudin, A.N., Krishnamoorthy, A., Joshua, V., Tsarenkov, G.V., "Analysis of the BMAP/G/1 retrial system with search of customers from the orbit." *Eur. J. Operational Research*, vol. 157, no. 1, (2004) 169–179.

[26] Falin, G., "A survey of retrial queues," *Queueing Systems*, vol. 7, no. 2, (1990) 127–167.

[27] Falin, G., Artalejo, J., "A finite source retrial queue," *Eur. J. Operational Research*, vol. 108, (1998) 409–424.

[28] Falin, G., Templeton, J., *Retrial Queues.* Chapman & Hall, 1997.

[29] Gaver, D., Jacobs, P., Latouche, G., "Finite birth-and-death models in randomly changing environments," *Adv. Appl. Prob.*, vol. 16, (1984) 715–731.

[30] Hanschke, T., "Explicit formulas for the characteristics of the M/M/2/2 queue with repeated attempts," *Appl. Prob.*, vol. 24, (1987) 486–494.

[31] Kulkarni, V.G., Liang, H.M., *Frontiers in Queueing: Models and Applications in Science and Engineering.* CRC Press, 1997, ch. Retrial queues revisited, 19–34.

[32] Marsan, M.A., Carolis, G., Leonardi, E., Lo Cigno, R., Meo, M., "Efficient estimation of call blocking probabilities in cellular mobile telephony networks with customer retrials," *IEEE Journal on Selected Areas in Communications*, vol. 19, no. 2, (February 2001) 332–346.

[33] Neuts, M.F., Ramalhoto, M.F., "A service model in which the server is required to search for customers," *Appl. Prob.*, vol. 21, no. 1, (March 1984) 157–166.

[34] Roszik, J., Kim, C., Sztrik, J., "Retrial queues in the performance modeling of cellular mobile networks using MOSEL," *International Journal of Simulation: Systems, Science and Technology*, vol. 6, (2005) 38–47.

[35] Roszik, J., Sztrik, J., Virtamo, J., "Performance analysis of finite-source retrial queues operating in random environments," *Int. J. Operational Research*, vol. 2, no. 3, (2007) 254–268.

[36] Sherman, N.P., Kharoufeh, J.P., "An M/M/1 retrial queue with unreliable server," *Operations Research Letters*, vol. 34, (2006) 697–705.

[37] Stewart, G.W., "On the adjugate matrix," *Linear Algebra and its Applications*, vol. 283, (1998) 151–164.

[38] Sztrik, J., Almasi, B., Roszik, J., "Heterogeneous finite-source retrial queues with server subject to breakdowns and repairs," *Math. Sciences*, vol. 132, (2006) 677–685.

[39] Wüchner, P., Meer, H., Barner, J., Bolch, G., "A brief introduction to MOSEL-2," in *Proc. of MMB 2006 Conference*, R. German and A. Heindl, Eds., GI/ITG/MMB, University of Erlangen. VDE Verlag, 2006.

[40] Wüchner, P., Sztrik, J., Meer, H., "Homogeneous finite-source retrial queues with search of customers from the orbit," in *Proc. of 14th GI/ITG Conference on Measurement, Modelling and Evaluation of Computer and Communication Systems (MMB 2008)*, Dortmund, Germany, March 2008.

[41] Wüchner, P., Sztrik, J., Meer, H., "Finite-source M/M/S retrial queue with search for balking and impatient customers from the orbit," *Computer Networks*, vol. 53, (2009) 1264–1273.

[42] Wüchner, P., Sztrik, J., Meer, H., "The impact of retrials on the performance of self-organizing systems," *Praxis der Informationsverarbeitung und Kommunikation (PIK)*, vol. 31, no. 1, (March 2008) 29–33.

[43] Yang, T., Templeton, J.G.C., "A survey on retrial queues," *Queueing Systems*, vol. 2, (1987) 201–233.

Patrick Wüchner, Hermann de Meer
Faculty of Informatics and Mathematics, University of Passau
Innstraße 43, 94032 Passau, Germany
e-mail: `{patrick.wuechner,hermann.demeer}@uni-passau.de`

János Sztrik
Faculty of Informatics, University of Debrecen
Egyetem tér 1, P.O. Box 12
4010 Debrecen, Hungary
e-mail: `jsztrik@inf.unideb.hu`