Sociedad de Estadística e Investigación Operativa
Top (2002) Vol. 10, No. 1, pp. 51–66

Approximation of Multiserver Retrial Queues by Means of Generalized Truncated Models

Vladimir V. Anisimov
GlaxoSmithKline, Research Statistics Unit, New Frontiers Science Park (South) Third Avenue, Harlow, Essex CM19 5AW, United Kingdom
e-mail: Vladimir_V_Anisimov@gsk.com
Jesus R. Artalejo
Department of Statistics and Operations Research
Faculty of Mathematics, University Complutense of Madrid, 28040 Madrid, Spain
e-mail: jesus_artalejo@mat.ucm.es

Abstract

It is well-known that an analytical solution of multiserver retrial queues is difficult and does not lead to numerical implementation. Thus, many papers approximate the original intractable system by the so-called generalized truncated systems which are simpler and converge to the original model. Most papers assume heuristically the convergence but do not provide a rigorous mathematical proof. In this paper, we present a proof based on a synchronization procedure. To this end, we concentrate on the $M/M/c$ retrial queue and the approximation developed by Neuts and Rao (1990). However, the methodology can be employed to establish the convergence of several generalized truncated systems and a variety of Markovian multiserver retrial queues.

Key Words: Retrial queues, stationary distribution, generalized truncated systems, synchronization, stochastic comparability.

AMS subject classification: 60K25, 90B22.

1 Introduction

The main characteristic of a retrial queue is that a primary customer who finds busy the service facility upon arrival immediately leaves the service area, but some time later he repeats his demand. Between trials a customer is said to be in *'orbit'*. The existence of a second flow of repeated attempts arises in many practical applications including telephony, communication protocols, local area networks and other stochastic systems.

J.R. Artalejo thanks the support received from DGES 98-0837.
Manuscript received: March 2001. Final version accepted: January 2002.

A complete review of the main results and the literature can be found in Artalejo (1999a), Artalejo (1999b) and Falin and Templeton (1997).

The existence of closed form solutions for the stationary distribution of multiserver retrial queues is reduced to a few special cases (see Artalejo (1996), Falin and Templeton (1997) and Gomez-Corral and Ramalhoto (1999). Some theoretical contributions present the stationary distribution of the system state in terms of contour integrals (Cohen (1957)) or as limit of continued fractions (Pearce (1989)). Most multiserver retrial queues can be viewed as a level dependent quasi-birth-and-death process. The main feature of its infinitesimal generator is the spatial heterogeneity caused by transitions due to repeated attempts. This lack of homogeneity explains the analytical complexity of retrial queues. More useful in practice is the implementation of a variety of approximations and truncated models. In this sense, Wilkinson (1956) proposes to truncate the capacity of the orbit at some value K. Stepanov (1999) develops more sophisticated methods of truncation based on the exclusion of a set of states with negligible stationary probabilities. In general, finite truncated models imply very demanding computational resources for getting a good accuracy. This drawback can be improved by using generalized truncated models.

The key of a generalized truncation is approximate the initial infinite system by another infinite calculable system. The fact that both (initial and approximate) systems are infinite provides much better accuracy. A variety of generalized truncated models have been considered to approximate the original intrincate retrial queues (see Artalejo and Pozo (2001), Choi et al. (1999), Falin (1983), Li and Yang (1999) and Neuts and Rao (1990)). The intuition indicates that, if the truncation level is large enough, then the generalized truncated model converges to the original one. Until now, this heuristic has been used in many papers but a rigorous mathematical proof is given only by Falin and Templeton (1997) for the model of Falin (1983), and by Anisimov and Artalejo (2001) for a more complicate model with retrials and negative arrivals. As the proof in the latter is technically very cumbersome, our goal in this paper is to provide a new methodology which can be applied to a versatile class of Markovian multiserver retrial queues.

As related works, a number of papers investigate several variants of the main multiserver retrial queue of type $M/M/c$. This literature (see Artalejo (1999a), Artalejo (1999b), Falin and Templeton (1997) and the references therein) includes mixed models with classical waiting line and

repeated attempts, systems with non-persistent customers and feedback, models with negative customers arriving in batches, polling systems with repeated attempts, models with waiting positions, etc. We also mention a number of recent papers devoted to algorithmic methods for retrial queues including the study of models with interarrival and interrepetition times of types $BMAP$, PH, SM, etc. Other papers deal with the investigation of limit theorems for understanding the system behavior under light and heavy traffic, low retrial rate and applications of limit results for switching processes to overloaded retrial queues (see Anisimov (1999)).

The remainder of the paper is organized as follows. In Section 2, we give the mathematical description of the $M/M/c$ retrial queue and describe several generalized truncated models. In Section 3, we illustrate the proposed methodology for the approximation of Neuts and Rao (1990) and demonstrate its convergence to the main $M/M/c$ queue with repeated attempts. Some concluding remarks are given in Section 4. The underlying synchronization procedure which is the key for proving the convergence is described in the Appendix.

2 The main multiserver retrial queue and its approximation by generalized truncated models

In this section, we focus on the main multiserver model of type $M/M/c$ with retrials. We consider that primary customers arrive according to a Poisson process of rate λ. The service facility consists of c identical servers and customer service times are independent and exponentially distributed with rate ν. An arriving customer finding all servers busy leaves temporary the service area and joins an orbit of blocked customers. We assume that the access from the orbit to the service facility is governed by the classical retrial policy, i.e., each customer in orbit reapplies for service individually after an exponential time of rate μ so the retrial rate given that j customers are in orbit is $j\mu$.

The system state at time t can be described by means of a bivariate process $X = \{(C(t), N(t)); t \geq 0\}$, where $C(t)$ is the number of busy servers and $N(t)$ is the number of customers in orbit at time t. Note that X is an irreducible Markovian process taking values in the lattice semi-strip $S = \{0, \ldots, c\} \times \mathbb{Z}_+$. Its infinitesimal generator, $Q = (q_{ab})$, has the following

elements. For $0 \leq i \leq c-1$, we have:

$$q_{(i,j),(m,n)} = \begin{cases} \lambda, & \text{if } (m,n) = (i+1,j), \\ i\nu, & \text{if } (m,n) = (i-1,j), \\ j\mu, & \text{if } (m,n) = (i+1,j-1), \\ -(\lambda + i\nu + j\mu), & \text{if } (m,n) = (i,j), \\ 0, & \text{otherwise,} \end{cases} \tag{2.1}$$

and for $i = c$:

$$q_{(c,j),(m,n)} = \begin{cases} \lambda, & \text{if } (m,n) = (c,j+1), \\ c\nu, & \text{if } (m,n) = (c-1,j), \\ -(\lambda + c\nu), & \text{if } (m,n) = (c,j), \\ 0, & \text{otherwise.} \end{cases} \tag{2.2}$$

The ergodicity condition of process X is $\rho = \lambda/c\nu < 1$ (see Falin and Templeton (1997)). Then, the stationary probabilities $P_{ij} = \lim_{t\to\infty} P\{C(t) = i, N(t) = j\}$ exist for all $(i,j) \in S$ and are positive. It is well known that neither a closed analytical solution nor a direct algorithmic computation of these limiting probabilities is still available. Thus, we next briefly discuss the use of generalized truncated systems which provide good approximations of the $M/M/c$ retrial queue.

The main feature of a generalized truncated model is to approximate the analysis of the original infinite system by another infinite model which can be successfully solved. Falin (1983) and Falin and Templeton (1997) consider a first simple model and shows numerically the superiority of generalized models over those approximations based on a finite truncation. The intuition tell us, that there should be an orbit level K, such that from the level K up, the process X performs similarly to the standard $M/M/1$ queue with arrival rate λ and service rate $c\nu$. Thus, if we denote by μ_j the retrial rate given that $N(t) = j$, we obtain the generalized truncated system $X^F = \{(C^F(t), N^F(t)); t \geq 0\}$ corresponding to the case

$$\mu_j = \begin{cases} j\mu, & \text{if } 0 \leq j \leq K, \\ \infty, & \text{if } j \geq K+1. \end{cases}$$

It can be proven that the condition $\rho < 1$ is again necessary and sufficient for the ergodicity of X^F. Falin and Templeton (1997) reexpresses X^F as a bidimensional migration process and proves the convergence of the

stationary probabilities $P_{ij}^F(K) = \lim_{t\to\infty} P\left\{C^F(t) = i, N^F(t) = j\right\}$ to P_{ij}, as $K \to \infty$.

A second possibility was introduced in Neuts and Rao (1990) and later used by a variety of authors (see, for instance, Choi et al. (1999) and Li and Yang (1999)). They restrict the number of customers in orbit who are allowed to conduct retrials to a maximum number K, i.e., now the retrial rate is $\mu_j = \min(j, K)\mu$. The approximate process $X^{NR} = \left\{(C^{NR}(t), N^{NR}(t)); t \geq 0\right\}$ can be viewed as a quasi-birth-and-death process with a large number of boundary states. Thus, the mathematical tools for studying the ergodicity of X^{NR} and methods for the recursive computation of the stationary probabilities $P_{ij}^{NR}(K) = \lim_{t\to\infty} P\left\{C^{NR}(t) = i, N^{NR}(t) = j\right\}$ are well investigated in the literature. In particular, the general theory states that the process X^{NR} is ergodic if and only if $\lambda\pi_c < K\mu(1-\pi_c)$, where $\pi_c = \left(\left(\frac{\lambda+K\mu}{\nu}\right)^c \Big/ c!\right)\left(\sum_{k=0}^{c}\left(\frac{\lambda+K\mu}{\nu}\right)^k \Big/ k!\right)^{-1}$. It should be noted that $\pi_c \to 1$ and $K\mu(1-\pi_c) \to c\nu$, as $K \to \infty$. It means that if $\rho < 1$, then at large enough K the process X^{NR} is also ergodic.

Recently, Artalejo and Pozo (2001) have studied another way to reduce the initial $M/M/c$ retrial model to a numerically tractable model. To this end, they assume that the retrial rate depends on the system state and, in particular, consider the case

$$\mu_{ij} = \begin{cases} \infty, & \text{if } 0 \leq i \leq c-2,\ j \geq K+1, \\ j\mu, & \text{otherwise.} \end{cases}$$

The corresponding Markovian process $X^{AP} = \left\{(C^{AP}(t), N^{AP}(t)); t \geq 0\right\}$ is a natural extension of the first truncated model X^F but its transitions are closer to the initial $M/M/c$ retrial queue. The ergodicity condition is again $\rho < 1$, as in the initial system. As opposite to the truncated processes X^F and X^{NR}, the new process X^{AP} preserves the non-homogeneity inherent to the existence of a flow of repeated attempts. This explains why the numerical comparison among the truncated models X^F, X^{NR} and X^{AP} gives evidence of the superiority of X^{AP}.

Falin and Templeton (1997) prove that $\lim_{K\to\infty} P_{ij}^F(K) = P_{ij}$. The proof is based on the notion of stochastic comparability and in some results for migration processes. In principle, the method does not hold in the case of the processes X^{NR} and X^{AP}. Therefore we provide another constructive proof

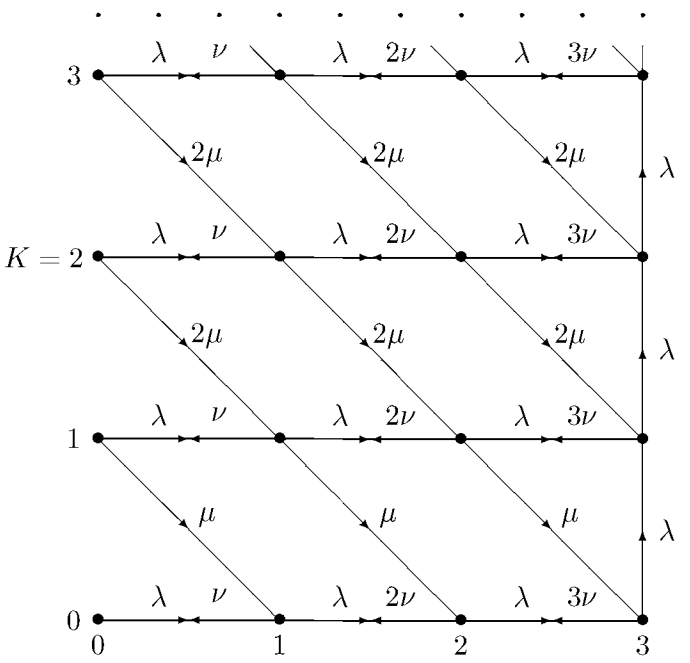

Figure 1: *State space and transitions of process* X^{NR}

based on a synchronization procedure between two equivalent versions of X^{NR} (respectively any other generalized truncated model) and the original process X. The only motivation for the choice of X^{NR} is its wide use in many papers during the last decade. The proof of the convergence of X^{AP} (or X^{F}) to the original model X can be given without significant differences. It follows from a casuistry similar to that given in Proposition B.1 (see Appendix). Hence, in the next section, we illustrate our methodology by proving the convergence of the probabilities $P_{ij}^{NR}(K)$ to the stationary probabilities P_{ij} of the initial system.

3 Convergence of X^{NR} to the initial process X

In this section, we concentrate our attention on the approximation of Neuts and Rao (1990). We need to prove that for any $(i, j) \in S$, we have

$\lim_{K\to\infty} P_{ij}^{NR}(K) = P_{ij}$. The busy period of a multiserver retrial queue is defined as the period that starts when an arriving customer finds an empty system and ends at the next service completion epoch at which the system becomes empty again. Let us denote the busy period of processes X, X^F, X^{NR} and X^{AP} by T, T_F^K, T_{NR}^K and T_{AP}^K, respectively.

By appealing to the intuition, we may expect the following relationships:

$$E[T_F^K] < E[T_{AP}^K] < E[T],\ K \geq 0, \tag{3.1}$$

$$E[T] < E[T_{NR}^K],\ K \geq 1, \tag{3.2}$$

$$E[T_{NR}^{K+1}] < E[T_{NR}^K],\ K \geq 1, \tag{3.3}$$

$$E[T_F^{K+1}] > E[T_F^K] \text{ and } E[T_{AP}^{K+1}] > E[T_{AP}^K],\ K \geq 0. \tag{3.4}$$

We can verify numerically the validity of inequalities (3.1)-(3.4). For example, taking the system parameters as $(\lambda, \nu, \mu, c) = (1.5, 1.0, 0.1, 5)$ we may compute the value of $E[T_F^K]$, $E[T_{AP}^K]$ and $E[T_{NR}^K]$. A description of suitable algorithms can be found in Artalejo and Pozo (2001).

The numerical example in Table 1 gives support to the correctness of (3.1)-(3.4). However, a rigorous mathematical proof is expected. In fact, similar stochastic relationships among the variables T, T_F^K, T_{NR}^K and T_{AP}^K are needed to establish the convergence of the approximate stationary distributions as $K \to \infty$. In the Appendix, we present a methodology based on the construction of equivalent versions of the original process X and the approximate one (X^{NR}, for example). Then, a synchronization method is used to compare sample paths of the equivalent versions. The construction yields to monotonicity properties which are the key to conclude the convergence of the approximate stationary distribution. This is done in Theorem 3.1.

Theorem 3.1. *Assume that the ergodicity condition $\rho < 1$ is satisfied. Then, for any $(i, j) \in S$, we have*

$$\lim_{K\to\infty} P_{ij}^{NR}(K) = P_{ij}. \tag{3.5}$$

Proof. In Section 2, we see that $\rho < 1$ implies that the process X^{NR}

K	$E[T_F^K]$	$E[T_{AP}^K]$	$E[T_{NR}^K]$
0	2.325892	2.330844	-
1	2.750437	2.752746	3.161210
2	2.855203	2.856112	2.925796
3	2.882898	2.883232	2.899513
4	2.890563	2.890682	2.894897
5	2.892743	2.892785	2.893931
6	2.893375	2.893389	2.893709
7	2.893559	2.893564	2.893656
8	2.893614	2.893616	2.893642

Table 1: *Mean busy period versus* K.
Case $(\lambda, \nu, \mu, c) = (1.5, 1.0, 0.1, 5)$

is also ergodic. Thus, the stationary probabilities $\{P_{ij}; (i,j) \in S\}$ and $\left\{P_{ij}^{NR}(K); (i,j) \in S\right\}$ exist and are positive for $K \geq K^*$.

Now we consider the infinitesimal generator, $Q_{NR}^K = (q_{ab}^K)$, of process X^{NR}. For $j \leq K$, its elements agree with those given in (2.1)-(2.2). If $j > K$, then q_{ab}^K is defined as q_{ab} but replacing $j\mu$ by $K\mu$. Thus, it is clear that the infinitesimal transition rates of Q_{NR}^K converge to the corresponding rates of Q. Then, it follows easily that T_{NR}^K converges weakly to T, as $K \to \infty$. To this end, we now appeal to the Appendix and consider the equivalent processes $\widetilde{X}$ and $\widetilde{X}^{NR}$ with busy periods $\widetilde{T}$ and $\widetilde{T}^{NR}$, respectively. According to the construction, we have that $\widetilde{T}_{NR}^{K+1} \leq \widetilde{T}_{NR}^K$ almost surely for $K \geq K^*$. Hence, it is obvious that $\widetilde{T}_{NR}^{K+1} \leq_d \widetilde{T}_{NR}^K$, where the symbol $\leq_d$ means stochastically smaller in distribution (see Stoyan (1983)). Since the random variables T_{NR}^K and $\widetilde{T}_{NR}^K$ are identically distributed, we have proved in fact that the sequence $\left\{T_{NR}^K; K \geq K^*\right\}$ decreases stochastically to T. Due to this monotone behavior, we have for any $L > 0$ and $K \geq K^*$

$$\int_L^\infty P\left\{T_{NR}^K > x\right\} dx \leq \int_L^\infty P\left\{T_{NR}^{K^*} > x\right\} dx.$$

Thus it follows trivially the uniform integrability of the function $g(x) = x$ with respect to $\left\{F_{T_{NR}^K}, K \geq K^*\right\}$, where $F_{T_{NR}^K}$ denotes the distribution

function of T_{NR}^K. The weak convergence jointly with the uniform integrability guarantee the convergence of expectations (see Laha and Rohatgi (1979)). Thus, we obtain

$$\lim_{K\to\infty} E[T_{NR}^K] = E[T]. \tag{3.6}$$

From expression (3.6) and the theory of regenerative processes, we have

$$\lim_{K\to\infty} P_{00}^{NR}(K) = \lim_{K\to\infty} \frac{1/\lambda}{1/\lambda + E[T_{NR}^K]} = P_{00}.$$

Now in the same way we can see that as $K \to \infty$ the distribution of the total time $\nu(i,j)_{NR}^K$ spent in any state (i,j) between two sequential returns to $(0,0)$ for the process X^{NR} weakly converges to the corresponding distribution of the time $\nu(i,j)$ for the process X. As $\nu(i,j)_{NR}^K \leq T_{NR}^K$, then it follows also the uniform integrability of the variable $\nu(i,j)_{NR}^K$ and we get

$$\lim_{K\to\infty} E[\nu(i,j)_{NR}^K] = E[\nu(i,j)], \text{ for any } (i,j).$$

The above relation together with ergodic theorem for regenerative processes (see Cinlar (1975)) implies (3.5) for any (i,j). □

Once the convergence of the stationary distribution is established, a variety of performance measures can also be investigated. Some important characteristics are:

1. The marginal distributions

$$P_{i.} = \sum_{j=0}^{\infty} P_{ij}, \ 0 \leq i \leq c,$$

$$P_{.j} = \sum_{i=0}^{c} P_{ij}, \ j \geq 0.$$

In particular, if $i = c$ we obtain the stationary blocking probability $B = P_{c.}$.

2. The mean number of busy servers

$$Y = \sum_{i=0}^{c} iP_{i.}.$$

3. The mean number of customers in orbit

$$N = \sum_{j=0}^{\infty} jP_{.j}.$$

4. The mean waiting time W.

It should be pointed out that Y, N and W can be expressed as follows (see Falin and Templeton (1997)):

$$Y = \frac{\lambda}{\nu}, \tag{3.7}$$

$$N = \frac{\nu + \mu}{\mu(c\nu - \lambda)} \left(\lambda - \nu \left(\sum_{i=0}^{c} i^2 P_{i.} - Y^2 \right) \right), \tag{3.8}$$

$$W = \frac{N}{\lambda}. \tag{3.9}$$

Since $P_{.j}$ only involves a finite sum, the convergence of the approximate marginal distribution $P_{.j}^{NR}(K)$ is trivial. Taking into account the above formulas (3.7)-(3.9) the problem is reduced to establish the convergence of $P_{i.}^{NR}(K)$ to $P_{i.}$. To prove this, we define $\nu(i,.)_{NR}^{K}$ as the total time that the process X^{NR} spent in the set $\{(i,j) \in S \mid j \geq 0\}$ between two successive returns to $(0,0)$. Then the proof follows by repeating the arguments given in the proof of Theorem 3.1.

4 Concluding remarks

We have developed an useful methodology for the investigation of the convergence of the stationary distribution of the so-called generalized truncated systems to the stationary distribution of the main multiserver retrial

queue of the type $M/M/c$. Although we have concentrated on the approximate model of Neuts and Rao (1990), the approach is general and remains valid for the rest of generalized truncated models and other variants of the $M/M/c$ retrial queue. We mention the model with linear retrial rate (see Artalejo and Gomez-Corral (1997)) as an example. However, the method fails in the case of complicate retrial queues where, in addition to regular departures, the customers may leave the system due to the existence of negative arrivals, disasters, impatient behavior, etc. For these cases, the method given in Anisimov and Artalejo (2001) can be used.

Appendix

In this Appendix we present an approach for the investigation of the stochastic comparability of the busy period of the generalized truncated models described in Section 2. As application of this methodology, we may state how the generalized truncated model is related to the main multiserver retrial queue. Concretely we can formalize the convergence of the stationary distribution of the truncated model to the stationary distribution of the $M/M/c$ retrial queue (see Theorem 3.1). Our methodology is valid only for exponentially distributed service times. However, it should be noted that the existing literature recognizes the complexity of multiserver retrial queues even at the Markovian level. Unfortunately, it seems difficult to extend the proof to a more general class of retrial queues.

To illustrate the approach we concentrate on the processes X^{NR} and X. The objective is to prove that $T \leq_d T_{NR}^K$, for any $K \geq 1$. It should be pointed out that the approach provides a general methodology so a minor variant of the arguments yields similar comparisons among T, T_F^K, T_{NR}^K and T_{AP}^K.

We first describe a synchronization mechanism between two equivalent versions of X and X^{NR} which is the key to establish the stochastic comparison for the busy period.

A The synchronization mechanism

Let us assume the initial conditions $(C(0), N(0)) = \left(C^{NR}(0), N^{NR}(0)\right) = (1, 0)$, i.e., we may think that a busy period of the processes X and X^{NR}

starts at time $t = 0$. Suppose that both processes are visiting the same state (i, j) at some time $t^* \geq 0$ and the busy period is still in progress at this time. Let us consider that $(i, j) \in S(K)$, where $S(K) = \{(m, n) \in S \mid n \leq K\} \cup \{(m, n) \in S \mid m = c, n > K\}$.

Note that the next transition of X is determined by the competition among the components of the random vector $\mathbf{v}_{ij} = (\xi, \eta_1, \ldots, \eta_i, \gamma_1, \ldots, \gamma_j)$, where ξ is exponentially distributed with rate λ, η_k are exponentially distributed with rate ν, for $1 \leq k \leq i$, and γ_k are exponentially distributed with rate μ, for $1 \leq k \leq j$. It should be noted that the components of vector $\mathbf{v}_{ij}$ are mutually independent. Furthermore, the subvector $(\eta_1, \ldots, \eta_i)$ has no components when $i = 0$. Analogously, the subvector $(\gamma_1, \ldots, \gamma_j)$ is not considered if $j = 0$ and/or $i = c$. A second vector $\mathbf{v}_{ij}^K = \left(\xi^K, \eta_1^K, \ldots, \eta_i^K, \gamma_1^K, \ldots, \gamma_j^K\right)$ identically distributed to $\mathbf{v}_{ij}$ determines the transition of the process X^{NR}.

We now construct two synchronized processes, $\widetilde{X}$ and $\widetilde{X}^{NR}$, in such a way that the first transition after t^* is determined respectively by the random vectors $\widetilde{\mathbf{v}}_{ij} = \left(\widetilde{\xi}, ..., \widetilde{\gamma}_j\right)$ and $\widetilde{\mathbf{v}}_{ij}^K = \left(\widetilde{\xi}^K, \ldots, \widetilde{\gamma}_j^K\right)$. These vectors are distributed as the original ones $\mathbf{v}_{ij}$ and $\mathbf{v}_{ij}^K$ but their components are chosen to be identical one by one, i.e., $\widetilde{\xi} = \widetilde{\xi}^K$, $\ldots$, $\widetilde{\gamma}_j = \widetilde{\gamma}_j^K$.

Let t_0 be the first epoch at which the processes $\widetilde{X}$ and $\widetilde{X}^{NR}$ leave the subset $S(K)$. It is clear from the construction that $\widetilde{X}$ and $\widetilde{X}^{NR}$ have identical sample paths until the time t_0. Due to the construction it follows that $\{(C(t), N(t)); 0 \leq t \leq t_0\}$ and $\left\{(\widetilde{C}(t), \widetilde{N}(t)); 0 \leq t \leq t_0\right\}$ (respectively $\left\{(C^{NR}(t), N^{NR}(t)); 0 \leq t \leq t_0\right\}$ and $\left\{(\widetilde{C}^{NR}(t), \widetilde{N}^{NR}(t)); 0 \leq t \leq t_0\right\}$) have the same finite-dimensional distributions. Hence, we have constructed equivalent versions of the initial processes until the time t_0.

Note that the system state at time t_0 is $\left(\widetilde{C}(t_0), \widetilde{N}(t_0)\right) = \left(\widetilde{C}^{NR}(t_0), \widetilde{N}^{NR}(t_0)\right) = (c - 1, j)$, for any $j > K$. Now the next transitions of $\widetilde{X}$ and $\widetilde{X}^{NR}$ are determined respectively by the competition among the random components of the vectors $\widetilde{\mathbf{v}}_{c-1,j} = \left(\widetilde{\xi}, \widetilde{\eta}_1, \ldots, \widetilde{\eta}_{c-1}, \widetilde{\gamma}_1, \ldots, \widetilde{\gamma}_j\right)$ and $\widetilde{\mathbf{v}}_{c-1,j}^K = \left(\widetilde{\xi}^K, \widetilde{\eta}_1^K, \ldots, \widetilde{\eta}_{c-1}^K, \widetilde{\gamma}_1^K, \ldots, \widetilde{\gamma}_K^K\right)$, where $\widetilde{\xi} = \widetilde{\xi}^K$, $\widetilde{\eta}_k = \widetilde{\eta}_k^K$, for $1 \leq k \leq c - 1$, $\widetilde{\gamma}_k = \widetilde{\gamma}_k^K$, for $1 \leq k \leq K$, i.e., all the components of $\widetilde{\mathbf{v}}_{c-1,j}^K$ are identical to

the corresponding component of $\widetilde{\mathbf{v}}_{c-1,j}$, but the vector $\widetilde{\mathbf{v}}_{c-1,j}$ has an extra subvector $(\widetilde{\gamma}_{K+1}, \ldots, \widetilde{\gamma}_j)$ with exponentially distributed components of rate μ. In other words, $\widetilde{\mathbf{v}}_{c-1,j} = \left(\widetilde{\mathbf{v}}^K_{c-1,j}, \widetilde{\gamma}_{K+1}, \ldots, \widetilde{\gamma}_j\right)$ and $\widetilde{\mathbf{v}}^K_{c-1,j} = \widetilde{\mathbf{v}}_{c-1,K}$, for $j \geq K+1$. Let us imagine that the next transition of $\widetilde{X}$ is caused by one of these specific variables $\widetilde{\gamma}_k$, for $K+1 \leq k \leq j$. Then, the process $\widetilde{X}$ moves to $(c, j-1)$ whereas $\widetilde{X}^{NR}$ remains at the state $(c-1, j)$. At this point, we say that the process $\widetilde{X}$ obtains a potential advantage over $\widetilde{X}^{NR}$.

Let $\{t_n; n \geq 1\}$ be the times of sequential jumps of $\widetilde{X}$ and $\widetilde{X}^{NR}$ from the epoch t_0 up, i.e., we join the transition epochs of processes $\widetilde{X}$ and $\widetilde{X}^{NR}$. We now define some auxiliary concepts.

Definition A.1 (compensation, potential and real advantages). Let us denote the state of $\widetilde{X}$ and $\widetilde{X}^{NR}$ at time t_n by (i, j) and (i', j'), respectively.

(i) We say that $\widetilde{X}$ and $\widetilde{X}^{NR}$ are compensated at time t_n if $(i, j) = (i', j')$.

(ii) We say that $\widetilde{X}$ has a potential advantage of order $k \geq 1$ over $\widetilde{X}^{NR}$ $(PA(k))$ at time t_n if $i + j = i' + j'$ and $j = j' - k$.

(iii) We say that $\widetilde{X}$ has a real advantage of order (l, k), $l \geq 1$, $k \geq 0$, over $\widetilde{X}^{NR}$ $(RA(l, k))$ at time t_n if $i + j = i' + j' - l$ and $j = j' - k$.

B Analysis of the transition $t_n \to t_{n+1}$

The next objective is to analyze the transition $t_n \to t_{n+1}$ in terms of the concepts introduced in Definition A.1. First, we observe that the construction of $\widetilde{X}$ and $\widetilde{X}^{NR}$ at time t_0 can be extended to any epoch t_n, $n \geq 1$. To this end, we consider $\min(i, i')$ variables exponentially distributed with rate ν and $\min(j, \min(j', K))$ variables exponentially distributed with rate μ. These random variables are common to both processes $\widetilde{X}$ and $\widetilde{X}^{NR}$. One more exponential variable with rate λ is also common. The rest of possible variables are specific for one of the processes. This completes the synchronization mechanism and provides equivalent versions of the initial processes $\widetilde{X}$ and $\widetilde{X}^{NR}$.

Proposition B.1. *Let us assume that at time t_n the process $\widetilde{X}$ has any type of advantage over $\widetilde{X}^{NR}$. Then, at time t_{n+1}, the advantage is either*

compensated or the process $\widetilde{X}$ *still has some advantage over* $\widetilde{X}^{NR}$*, but* $\widetilde{X}^{NR}$ *cannot get advantage over* $\widetilde{X}$ *during the transition* $t_n \rightarrow t_{n+1}$.

Proof. We first consider that $\widetilde{X}$ has a $PA(k)$ over $\widetilde{X}^{NR}$. Then, the occurrence of one of the following events implies a modification of the current situation:

- A common arrival occurs and $i = c$, then $\widetilde{X}$ has a $PA(k-1)$ over $\widetilde{X}^{NR}$ at time t_{n+1}. If $k = 1$, it means a compensation of both processes.
- A specific departure of $\widetilde{X}$ occurs, then $\widetilde{X}$ gets a $RA(1,k)$ over $\widetilde{X}^{NR}$.
- A specific retrial of $\widetilde{X}$ occurs, then the potential advantage increases in one unit, i.e., at time t_{n+1}, $\widetilde{X}$ has a $PA(k+1)$ over $\widetilde{X}^{NR}$.
- A specific retrial of $\widetilde{X}^{NR}$ occurs, then $\widetilde{X}$ has a $PA(k-1)$ over $\widetilde{X}^{NR}$ at time t_{n+1}. If $k = 1$ a compensation occurs.

Figure 1 is useful to understand the above casuistics. The rest of possible transitions preserve the relationship that processes $\widetilde{X}$ and $\widetilde{X}^{NR}$ had at time t_n.

The analysis of the case in which $\widetilde{X}$ has a $RA(l,k)$ over $\widetilde{X}^{NR}$ at time t_n is similar. The following events yield a modification of the previous relationship between $\widetilde{X}$ and $\widetilde{X}^{NR}$.

- A common arrival occurs, $i = c$ and $i' < c$, then $\widetilde{X}$ has a $RA(l,k-1)$ over $\widetilde{X}^{NR}$ at time t_{n+1}.
- A common arrival occurs, $i < c$ and $i' = c$, we now obtain a $RA(l,k+1)$ of $\widetilde{X}$ over $\widetilde{X}^{NR}$.
- A specific departure of $\widetilde{X}$ occurs, then $\widetilde{X}$ gets a $RA(l+1,k)$ over $\widetilde{X}^{NR}$. Note that in this case $l < k$.
- A specific departure of $\widetilde{X}^{NR}$ occurs, then $\widetilde{X}$ has a $RA(l-1,k)$ over $\widetilde{X}^{NR}$ at time t_{n+1}. In this case $0 \leq k < l$ and $l > 1$.
- A specific retrial of $\widetilde{X}$ occurs, then $\widetilde{X}$ has a $RA(l,k+1)$ over $\widetilde{X}^{NR}$ at time t_{n+1}.
- A specific retrial of $\widetilde{X}^{NR}$ occurs, then $\widetilde{X}$ has a $RA(l,k-1)$ over $\widetilde{X}^{NR}$.

From the above discussion, it is clear that the dynamic between two successive transition times does not allow a chance to $\widetilde{X}^{NR}$ for getting any type of advantage over $\widetilde{X}$, given that at the initial time t_n the advantage corresponds to $\widetilde{X}$. This proves the result. □

Now if the processes $\widetilde{X}$ and $\widetilde{X}^{NR}$ start from the same state then, as it was proved, the process $\widetilde{X}$ can obtain an initial advantage over $\widetilde{X}^{NR}$and due to Proposition B.1 this situation cannot be inverted. To this end we mention that the set of pairs (i, j) that have some type of advantage over the state $(0, 0)$ is empty. This means that the event $\widetilde{X}^{NR} = (0, 0)$ implies the event $\widetilde{X} = (0, 0)$, and finally $\widetilde{T}(\omega) \leq \widetilde{T}^{K}_{NR}(\omega)$ for all sample paths ω, and for some paths the inequality is strict. Thus, we have proved that $\widetilde{T} \leq_d \widetilde{T}^{K}_{NR}$ and consequently $T \leq_d T^{K}_{NR}$.

References

Anisimov V.V. (1999). Switching stochastic models and applications in retrial queues. *Top* 7, 169-186.

Anisimov V.V. and Artalejo J.R. (2001). Analysis of Markov multiserver retrial queues with negative arrivals. *Queueing Systems* 39, 157-182.

Artalejo J.R. (1996). Stationary analysis of the characteristics of the $M/M/2$ queue with constant repeated attempts. *Opsearch* 33, 83-95.

Artalejo J.R. (1999a). Accessible bibliography on retrial queues. *Mathematical and Computer Modelling* 30, 1-6.

Artalejo J.R. (1999b). A classified bibliography of research on retrial queues: Progress in 1990-1999. *Top* 7, 187-211.

Artalejo J.R. and Gomez-Corral A. (1997). Steady state solution of a single-server queue with linear repeated requests. *Journal of Applied Probability* 34, 223-233.

Artalejo J.R. and Pozo M. (2001). Numerical calculation of the stationary distribution of the main multiserver retrial queue. *Annals of Operations Research* 111 (to appear).

Choi B.D., Chang Y. and Kim B. (1999). $MAP_1, MAP_2/M/c$ retrial queue with guard channels and its application to cellular networks. *Top* 7, 231-248.

Cinlar E. (1975). *Introduction to Stochastic Processes.* Pretince-Hall.

Cohen J.W. (1957). Basic problems of telephone traffic theory and the influence

of repeated calls. *Phillips Telecommunication Review* 18, 49-100.

Falin G.I. (1983). Calculation of probability characteristics of a multiline system with repeat calls. *Moscow University Computational Mathematics and Cybernetics* 1, 43-49.

Falin G.I. and Templeton J.G.C. (1997). *Retrial Queues.* Chapman and Hall.

Gomez-Corral A. and Ramalhoto M.F. (1999). The stationary distribution of a Markovian process arising in the theory of multiserver retrial queueing systems. *Mathematical and Computer Modelling* 30, 141-158.

Laha R.G. and Rohatgi V.K. (1979). *Probability Theory.* John Wiley.

Li H. and Yang T. (1999). Steady-state queue size distribution of discrete-time $PH/Geo/1$ retrial queues. *Mathematical and Computer Modelling* 30, 51-63.

Neuts M.F. and Rao B.M. (1990). Numerical investigation of a multiserver retrial model. *Queueing Systems* 7, 169-190.

Pearce C.E.M. (1989). Extended continued fraction, recurrence relations and two-dimensional Markov processes. *Advances in Applied Probability* 21, 357-375.

Stepanov S.N. (1999). Markov models with retrials: the calculation of stationary performance measures based on the concept of truncation. *Mathematical and Computer Modelling* 30, 207-228.

Stoyan D. (1983). *Comparison Methods for Queues and other Stochastic Models.* John Wiley.

Wilkinson, R.I. (1956). Theories for toll traffic engineering in the U.S.A. *The Bell System Technical Journal* 35, 421-514.