Quality Technology & Quantitative Management
Vol. 8, No. 3, pp. 211-236, 2011

QTQM

Performance Analysis of a Two-Server Heterogeneous Retrial Queue with Threshold Policy

Dmitry Efrosinin[1] and Janos Sztrik[2]

[1]Johannes Kepler University of Linz, Linz, Austria
[2]University of Debrecen, Debrecen, Hungary
(*Received* December 2008, *accepted* July 2009)

Abstract: In the paper we deal with a Markovian queueing system with two heterogeneous servers and constant retrial rate. The system operates under a threshold policy which prescribes the activation of the faster server whenever it is idle and a customer tries to occupy it. The slower server can be activated only when the number of waiting customers exceeds a threshold level. The dynamic behaviour of the system is described by a two-dimensional Markov process that can be seen as a quasi-birth-and-death process with infinitesimal matrix depending on the threshold. Using a matrix-geometric approach we perform a stationary analysis of the system and derive expressions for the Laplace transforms of the waiting time as well as arbitrary moments. Illustrative numerical results are presented for the threshold policy that minimizes the mean number of customers in the system and are compared with other heuristic control policies.

Keywords: Controllable queueing system, retrial queue, sojourn time distribution, steady-state probabilities, threshold control policy, waiting time distribution.

---

1. Introduction

Different types of retrial queueing systems for single and multiple server cases have found applications in local area networks and communication protocols. For detailed review of the literature on retrial queues and some concrete examples the reader is referred to [7, 8, 31, 26]. In a retrial queueing system a customer who finds the service area blocked is assigned to an orbit where the customer will repeat the attempt to get a service. Most retrial queues are considered with respect to a classical retrial policy assuming that any particular customer in orbit retries independently of each other, so that the intervals between successive repeated attempts are exponentially distributed with rate depending on the orbit size. Nevertheless, some computer networks do not satisfy this assumption. Sometimes the server must check if the transmission medium available or not, or we may think that there is a delay for search of customers by server. In this case a constant retrial policy, that assume the equal retrial rates, is more preferable. This paper deals with the latter case of retrial policy. The constant retrial policy was introduced in [18], where the $M/M/1$ retrial queue was investigated for a telephone exchange model where the customers in the retrial group form a queue and only the customer at the head of the orbit can request service after an exponentially distributed retrial time. Further in [17] this discipline was called a retrial queue with FCFS orbit. There are a number of retrial queues in the literature, that were investigated under the constant retrial policy, see for instance [3, 5, 23]. For the application of the constant retrial policy in communication system, see e.g. Choi *et al.* [11]. Besides, this kind of retrial policy was used for the stability of the ALOHA protocol. The constant retrial policy also may occur in the

analysis of classical retrial queues. Neuts and Rao [25] use a truncation model exhibiting spatial homogeneity from some orbit level M upwards. In this case retrial queues can be studied by well known matrix-analytic methods for quasi-birth-and-death (QBD) processes.

Multiserver retrial queues have been extensively studied for homogeneous servers (equal service intensities), for recent results confer, for instance [4, 5, 10, 16]. But for heterogeneous servers (different service intensities) we have found only [27, 29]. In fact, for many real applications, where the multiserver retrial queues matches the mathematical models, the assumption of homogeneity is an unjustified limitation that was done only to simplify the model in order to get the nice analytic results. It motivates us to pay major attention in our research to the heterogeneous queues. Queues with heterogeneous servers can be widely used for modeling real systems with heterogeneous environment, e.g. a group of servers with different types of processors as a consequence of system updates, nodes in telecommunication networks with links of different capacities, nodes in wireless systems serving different mobile users. In [22] it was proved that heterogeneous multiserver systems without retrials are superior in performance to the heterogeneous ones. This was confirmed recently in [30] using simulation results. The author made a conclusion that by designing and configuring a multiserver system the heterogeneous structure must be seriously considered as well. In [9] a non-trivial application of the queue with heterogeneous servers to the performance evaluation of a wireless communication system is presented. The systems with heterogeneous servers are mostly investigated with respect to heuristic service policies, e.g. the Fastest Free Server (FFS) or Random Service Selection (RSS) policies.

The motivation for considering the queues with threshold policy comes from the fact that they can be superior in performance to homogeneous ones, as considered e.g. in [5], with the same total service rate, and to the heterogeneous systems operating under heuristic control policies. They may considerably improve a system's performance by reducing the sojourn time (or the number of customers in the system): it is often better to wait until the faster server will be idle than to occupy the slower one. It was shown in [12, 28] that the dynamic programming value function for the heterogeneous system has specific monotonicity properties which imply a threshold and monotonicity structure of the optimal control policy minimizing the mean sojourn time. Similar results were formulated in [13] and [14] for the queues with classical and constant retrial policies, respectively.

In this paper we further consider the problem of determining the stationary waiting and sojourn time distributions. For uncontrolled ordinary queueing systems without retrials the waiting time distribution can be derived by considering the system state at the arrival time of a tagged customer, see e.g. Kleinrock [20]. In the controlled case [15] it was shown that the waiting and sojourn time distributions correspond to a linear combination of Erlang distributions. The waiting time in uncontrolled retrial queues is more difficult to determine. Some methods are described in the monograph by Falin and Templeton [16]. For retrial systems (uncontrolled or controlled) with direct access of primary customers to the service area the waiting time analysis becomes more complicated because the waiting time of a customer in this case depends also on future arrivals. In systems with classical retrial policy it is necessary to take into account that some later arriving customer can be served earlier according to the random order policy for the orbiting customers. A recursive scheme for the computation of the Laplace transforms of the waiting time of a tagged customer is introduced by Artalejo *et al.* [6]. This paper will provide a starting point for our waiting time analysis of controlled queues with constant retrial policy. In systems with constant retrial rate for the orbiting customers future arrivals can influence the waiting time of the tagged customer by influencing the servers that can be active in the future. Therefore in comparison with [19],

where the uncontrollable queue is considered, in this case it is also necessary to consider the arrival process after the arrival of the tagged customer up to the time where service of this customer starts. For the system under consideration the calculation of the waiting time distribution is achieved by analyzing the transient Markov process with absorption at the time the tagged customer starts service. This requires an extension of the state representation since we have to know at each time the position of the tagged customer in the list of orbiting customers.

The analysis of the retrial queue with heterogeneous servers and constant retrial rate includes the following contributions:

(a) We model the system as a quasi-birth-and-death (QBD) process with threshold dependent block-tridiagonal infinitesimal matrix and apply the general theory of matrix-geometric solutions [24] in order to derive the stationary distribution of the system states and ergodic condition.

(b) We calculate the threshold levels that minimize the mean number of customers in the system. Then we analyze the boundary to the areas with different threshold levels in order to derive an approximation formula for the explicit calculation of threshold levels.

(c) We calculate the main performance characteristics for the system with different control policies and analyze the influence of these policies on the quality of service. It is shown that the heterogeneous system may be superior in performance to the homogeneous one with the same total service time.

(d) We obtain the Laplace transforms of the waiting time and sojourn time and develop a recursive algorithm for the computation of any corresponding arbitrary moment.

(e) We derive the discrete distribution function for the number of retrials made by a customer as well as the corresponding moments.

(f) To limit the size of the formulas we only present the case of two servers, but the presented methods can be applied to any number of servers. The methods can also be extended to some other models, e.g. where arrival and service rates are dependent on a modulated Markov process, or with phase type interarrival and service times.

The paper is organized as follows. In Section 2, we describe the mathematical model and develop equations for the computation of steady-state probabilities and mean performance measures. In Section 3 we develop recursive equations for the calculation of the stationary waiting and sojourn time distributions under threshold policy as well as the corresponding moments of an arbitrary order. In Section 4 we derive the (discrete) distribution function for the number of retrials made by a customer. In section 5 we present some illustrative numerical examples for the performance mean measures and the inversion of Laplace transforms and generation functions. The results for threshold and heuristic control policies are compared against each other.

In further sections we will use the notations $\mathbf{e}(n), \mathbf{e}_j(n)$ and I_n, respectively, for the column-vector of dimension n consisting of 1's, the column vector of dimension n with 1 in the j-th (beginning from 0-th) position and 0 elsewhere, and an identity matrix of dimension $n \times n$. The notations will be used without specifying the dimensions if these are clear from the context.

2. Mathematical Model and Steady State Analysis at an Arbitrary Moment

Consider an $M/M/2/N$ queueing system in which primary customers arrive according to a Poisson stream with intensity $\lambda > 0$ and two heterogeneous exponential servers with intensities μ_1, μ_2 where $\mu_1 > \mu_2 > 0$. We assume a threshold policy for the activation of the servers, given by threshold levels q_k for the k-th server, where $1 = q_1 \le q_2 < \infty$. Upon a customer's primary arrival, it receives service immediately if either the faster server is idle or the faster server is busy but the slower one is idle and the number of waiting customers exceeds a threshold level q_2. Otherwise, the customer is compelled to move to the orbit, which has capacity $2 \le N \le \infty$. From there it retries for service after an exponential distributed time with intensity $\gamma > 0$. Customers in orbit form a queue such that only the customer at the head can retry for access to the servers, i.e. we assume a FCFS orbiting discipline. If at the instant of retrial the customer finds both servers occupied or finds the faster server occupied while the slower one is idle but the number of orbiting customers is less then q_2, then he returns to the head of the orbit. Otherwise, if it finds the faster server idle or finds the faster server busy while the slower one is idle and the number of orbiting customers exceeds q_2, then it immediately obtains service. All interarrival times, intervals of successive retrials and service times are assumed to be mutually independent.

Let $Q(t)$ be the number of customers in the orbit at time t, $D_1(t), D_2(t)$ denote the states of the servers via

$$D_k(t) = \begin{cases} 0, & \text{if the } k\text{-th server is idle and} \\ 1, & \text{if the } k\text{-th server is busy.} \end{cases}$$

Then the process

$$\{X(t)\}_{t\ge 0} = \{Q(t), D_1(t), D_2(t)\}_{t\ge 0}, \tag{1}$$

is an irreducible continuous-time Markov chain with state space defined as

$$E = \{x = (q, d_1, d_2);\ 0 \le q \le N,\ d_k = \{0,1\},\ k = 1,2\},$$

where q and d_k denote the number of orbiting customers and states of the servers, respectively. Further the notations $q(x)$ and $d_k(x)$ will be used to define the orbit size and states of servers in specific state x. Let the states $x \in E$ be partitioned as follows:

$$\mathbf{i} = \{(i,0,0), (i,1,0), (i,0,1), (i,1,1)\},\ i \ge 0.$$

Using elementary arguments, the Markov chain $\{X(t)\}_{t>0}$ for $N = \infty$ has an infinitesimal matrix $\Lambda = [a_{xy}]_{x,\,y\in E}$ which has threshold dependent block-tridiagonal structure given by

$$\Lambda = \begin{pmatrix} A_{00} & A_{01} & 0 & 0 & 0 & 0 & 0 & \dots \\ A_{21} & A_{10} & A_{01} & 0 & 0 & 0 & 0 & \dots \\ 0 & A_{21} & A_{10} & A_{01} & 0 & 0 & 0 & \dots \\ \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \dots \\ 0 & \cdots & 0 & A_{21} & A_{11} & A_{02} & 0 & \dots \\ 0 & \cdots & 0 & 0 & A_{22} & A_{12} & A_{02} & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \Bigg\} q_2,$$

where $(A_{00} + A_{01})\mathbf{e} = (A_{21} + A_{10} + A_{01})\mathbf{e} = (A_{21} + A_{11} + A_{02})\mathbf{e} = (A_{22} + A_{12} + A_{02})\mathbf{e} = \mathbf{0}$. It is

obvious that matrix Λ has the format of a quasi-birth-and-death (QBD) process. The submatrices A_{02}, A_{12}, A_{22} describe the homogeneous part of the QBD process and are given by

$$A_{12} = \begin{pmatrix} -(\lambda+\gamma) & \lambda & 0 & 0 \\ \mu_1 & -(\lambda+\mu_1+\gamma) & 0 & \lambda \\ \mu_2 & 0 & -(\lambda+\mu_2+\gamma) & \lambda \\ 0 & \mu_2 & \mu_1 & -(\lambda+\mu_1+\mu_2) \end{pmatrix},$$

$$A_{02} = \lambda \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad A_{22} = \gamma \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

The boundary matrices are defined by

$$A_{00} = \begin{pmatrix} -\lambda & \lambda & 0 & 0 \\ \mu_1 & -(\lambda+\mu_1) & 0 & 0 \\ \mu_2 & 0 & -(\lambda+\mu_2) & \lambda \\ 0 & \mu_2 & \mu_1 & -(\lambda+\mu_1+\mu_2) \end{pmatrix},$$

$$A_{10} = \begin{pmatrix} -(\lambda+\gamma) & \lambda & 0 & 0 \\ \mu_1 & -(\lambda+\mu_1) & 0 & 0 \\ \mu_2 & 0 & -(\lambda+\mu_2+\gamma) & \lambda \\ 0 & \mu_2 & \mu_1 & -(\lambda+\mu_1+\mu_2) \end{pmatrix},$$

$$A_{11} = \begin{pmatrix} -(\lambda+\gamma) & \lambda & 0 & 0 \\ \mu_1 & -(\lambda+\mu_1) & 0 & 0 \\ \mu_2 & 0 & -(\lambda+\mu_2+\gamma) & \lambda \\ 0 & \mu_2 & \mu_1 & -(\lambda+\mu_1+\mu_2) \end{pmatrix},$$

$$A_{01} = \lambda \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad A_{21} = \gamma \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

We now derive the condition for the system to reach stationary regime. To accomplish this, we define matrix $A = A_{22} + A_{12} + A_{02}$. According to the results obtained for QBD processes in [24], the necessary and sufficient condition for ergodicity of the process $\{X(t)\}_{t\geq 0}$ is of the form $\mathbf{p}A_{22}\mathbf{e} < \mathbf{p}A_{02}\mathbf{e}$, where the vector $\mathbf{p}$ is given by $\mathbf{p}A = 0$ and $\mathbf{pe} = 1$. After some routine manipulation, the ergodicity condition turns out to be

$$\rho = \frac{\lambda(\lambda+\gamma)^2(\lambda+\mu_2+\gamma)}{M\gamma(\lambda+\gamma)^2 + \gamma\mu_1\mu_2(3(\lambda+\gamma)+\mu_1) + \mu_2^2\gamma(\lambda+\mu_1+\gamma)} < 1.$$

Under the ergodicity condition (2), the stationary probability vector $\boldsymbol{\pi}$ of the infinitesimal matrix Λ exists. The macro-vector $\boldsymbol{\pi}=(\boldsymbol{\pi}_0,\boldsymbol{\pi}_1,\boldsymbol{\pi}_2,...)$, where $\boldsymbol{\pi}_i$ is a sub-vector of probabilities with $x\in\mathbf{i}$, $\pi_x=\lim_{t\to\infty}\mathbf{P}\{X(t)=x\}$, is given by

$$\begin{aligned}&\boldsymbol{\pi}_i=\boldsymbol{\pi}_{q_2}\prod_{j=1}^{q_2-i}M_{q_2-j},\ 0\le i\le q_2-1,\\&\boldsymbol{\pi}_i=\boldsymbol{\pi}_{q_2}R^{i-q_2},\ i\ge q_2,\end{aligned}\tag{3}$$

where the matrices M_i are defined by

$$\begin{aligned}&M_0=-A_{21}A_{00}^{-1},\\&M_i=-A_{21}(M_{i-1}A_{01}+A_{10})^{-1},\ 1\le i\le q_2-2,\\&M_{q_2}=-A_{22}(M_{q_2-2}A_{01}+A_{11})^{-1}.\end{aligned}\tag{4}$$

$\boldsymbol{\pi}_{q_2}$ is the unique solution of the system of equations

$$\boldsymbol{\pi}_{q_2}\left[\sum_{i=0}^{q_2-1}\prod_{j=1}^{q_2-i}M_{q_2-j}+(I-R)^{-1}\right]\mathbf{e}=1,\ \ \boldsymbol{\pi}_{q_2}(M_{q_2-1}A_{02}+A_{12}+RA_{22})=\mathbf{0}.\tag{5}$$

Matrix R is the minimal non-negative solution to the matrix equation

$$R^2A_{22}+RA_{12}+A_{02}=\mathbf{0}.\tag{6}$$

This equation is typically solved numerically using the iteration procedure:

$$R_0=\mathbf{0},\ R_{n+1}=-(A_{02}A_{12}^{-1}+R_n^2A_{22}A_{12}^{-1}).$$

All the inverses in (4) are well defined because the matrices involved are non-singular, a consequence of their probabilistic interpretation as transition matrices of transient Markov chains. The non-singularity can also be proven recursively via explicit calculation, by showing that the involved matrices are row diagonally dominant.

Remark 1 For the same model but assuming a finite orbit size $q_2\le N<\infty$ a similar method can be applied in order to calculate the stationary distribution. Above q_2 the solution is no longer geometric. The blocks $\boldsymbol{\pi}_i$ can be expressed as

$$\boldsymbol{\pi}_i=\boldsymbol{\pi}_N\prod_{j=1}^{N-i}M_{N-j},\ i\ge 0,$$

where the matrices M_i for $i\le q_2-1$ are defined as above and $M_i=-A_{22}(M_{i-1}A_{02}+A_{12})^{-1}$, $i\ge q_2$. Using boundary conditions at $i=N$ and the normalization equation, one obtains a set of linear equations in π_N

$$\boldsymbol{\pi}_N\left[1+\sum_{i=0}^{N-1}\prod_{j=1}^{N-i}M_{N-j}\right]\mathbf{e}=1,$$

$$\boldsymbol{\pi}_N(M_{N-1}+A_{13}A_{02}^{-1})=\mathbf{0},$$

where

$$A_{13} = \begin{pmatrix} -(\lambda+\gamma) & \lambda & 0 & 0 \\ \mu_1 & -(\lambda+\mu_1+\gamma) & 0 & \lambda \\ \mu_2 & 0 & -(\lambda+\mu_2+\gamma) & \lambda \\ 0 & \mu_2 & \mu_1 & -(\mu_1+\mu_2) \end{pmatrix}.$$

The method works efficiently as long as N is not too large. If N is too large, the method fails because it requires both a forward and a backward recursion. These recursions cannot both be made stable at the same time, and round-off errors start accumulating eventually. In this case the matrix geometric solution corresponding to $N = \infty$ is a good approximation.

As soon as the probabilities $\boldsymbol{\pi}_i$, $i \geq 0$ are known, different performance characteristics of the system can be evaluated. Some of them are enumerated below.

Utilization of the system

$$\bar{U} = 1 - \boldsymbol{\pi}_0 \mathbf{e}_0. \tag{7}$$

Utilization of the first and second server

$$\bar{U}_1 = \left[\sum_{i=0}^{q_2-1} \boldsymbol{\pi}_i + \boldsymbol{\pi}_{q_2}(I-R)^{-1}\right](\mathbf{e}_1+\mathbf{e}_3),\ \bar{U}_2 = \left[\sum_{i=0}^{q_2-1} \boldsymbol{\pi}_i + \boldsymbol{\pi}_{q_2}(I-R)^{-1}\right](\mathbf{e}_2+\mathbf{e}_3). \tag{8}$$

Mean number of busy servers

$$\bar{C} = \left[\sum_{i=0}^{q_2-1} \boldsymbol{\pi}_i + \boldsymbol{\pi}_{q_2}(I-R)^{-1}\right](\mathbf{e}_1+\mathbf{e}_2+2\mathbf{e}_3) = \bar{U}_1 + \bar{U}_2. \tag{9}$$

Mean number of customers waiting in the queue

$$\bar{Q} = \left[\sum_{i=0}^{q_2-1} i\boldsymbol{\pi}_i + \boldsymbol{\pi}_{q_2}(R(I-R)^{-1} + q_2 I)(I-R)^{-1}\right]\mathbf{e}. \tag{10}$$

Mean number of customers in the system

$$\begin{aligned} \bar{N} &= \left[\sum_{i=0}^{q_2-1} (i+1)\boldsymbol{\pi}_i + \boldsymbol{\pi}_{q_2}(R(I-R)^{-1} + (q_2+1)I)(I-R)^{-1}\right]\mathbf{e} + \left[\sum_{i=0}^{q_2-1} \boldsymbol{\pi}_i + \boldsymbol{\pi}_{q_2}(I-R)^{-1}\right]\mathbf{e}_3 \\ &= \bar{C} + \bar{Q}. \end{aligned} \tag{11}$$

Mean waiting and sojourn times

$$\bar{W} = \frac{\bar{Q}}{\lambda},\ \bar{T} = \frac{\bar{N}}{\lambda}. \tag{12}$$

Blocking probability

$$P_{blocking} = \sum_{i=0}^{q_2-1} \boldsymbol{\pi}_i(\mathbf{e}_1+\mathbf{e}_3) + \boldsymbol{\pi}_{q_2}(I-R)^{-1}\mathbf{e}_3. \tag{13}$$

Mean busy period

$$\overline{L} = \frac{1}{\lambda}\left(\frac{1}{\boldsymbol{\pi}_0 \mathbf{e}_0} - 1\right). \tag{14}$$

3. Waiting Time Distribution

In this section we shall determine the distribution of the waiting time in a system with threshold level $q_2 \geq 1$ and with orbit capacity $N < \infty$. The restriction on the orbit size is used due to the arguments that will be given in the remark further below. For a queue with heterogeneous servers and a threshold policy the waiting time of a customer can depend on future arrivals as they may change the number of active servers. Moreover, we assumed that an arriving customer may have a direct access to the idle server according to the control policy. Hence the waiting time of a tagged customer in the system under a threshold policy depends not only on its position in the orbit, but also on the orbit length during its waiting time. The calculation of the waiting time distribution is performed by analyzing the auxiliary process just after the arrival of the tagged customer with absorption at the time when it starts the service. Let us introduce the transient Markov chain

$$\hat{X}(t) = (Q(t), D_1(t), D_2(t), J(t)). \tag{15}$$

The state space is

$$\hat{E} = \{x = (q, d_1, d_2, j);\ 0 \leq q \leq N,\ d_k = \{0,1\},\ k = 1,2,\ 0 \leq j \leq q\},$$

where the last component $J(t)$ denotes the position of the tagged customer in the list of orbiting customers at time t. This component can take the values $\{0, 1, 2, \ldots\}$, and decreases for the system under threshold policy

- at a retrial arrival time when the first server is idle,
- at a retrial arrival time when at least one of the servers is idle and the queue length is greater than q_2.

The process is absorbed when the component $J(t)$ become equal to zero. Furthermore, $Q(t) \geq J(t)$ at any time t when the tagged customer has to wait in the orbit. At the point of time of a new arrival t^+ (the initial time for the Markov chain $\{\hat{X}(t)\}_{t \geq 0}$) it is obvious that $J(t^+) = Q(t^+)$ if the tagged customer has to wait in the orbit and $J(t^+) = 0$ if upon arrival the customer can be served immediately.

Further the following notations are used:

W: the waiting time in the system;

W_x: the residual waiting time of a tagged customer given state x;

$\omega_x(t)$: the density function of the residual waiting time;

$\tilde{\omega}_x(s) = \mathbf{E}\left[e^{-sW_x}\right] = \int_0^\infty e^{-st}\omega_x(t)dt,\ Re[s] \geq 0$: the Laplace-Stieltjes transform;

$\tilde{\omega}_x(s)$ and $\tilde{W}(s)$: the unconditional Laplace-Stieltjes transform (LST) and Laplace transform (LT).

Let $\tilde{\mathbf{w}}(s) = (\tilde{\mathbf{w}}_1(s), \ldots, \tilde{\mathbf{w}}_N(s))^t$ denotes the vector of LSTs, partitioned as

$$\tilde{\mathbf{w}}_j(s) = (\tilde{\mathbf{w}}_{j,j}(s), \tilde{\mathbf{w}}_j(s), ..., \tilde{\mathbf{w}}_{N,j}(s))^t,\ 1 \le j \le N,$$

$$\tilde{\mathbf{w}}_{i,j}(s) = (\tilde{\omega}_{(i,0,0,j)}(s), \tilde{\omega}_{(i,1,0,j)}(s), \tilde{\omega}_{(i,0,1,j)}(s), \tilde{\omega}_{(i,1,1,j)}(s))^t,\ i \ge j.$$

Due to the Markov property of the chain $\hat{X}(t)$, the residual waiting time in state x consists of the time the system spends in state x until the next transition with density $a_x e^{-a_x t}, a_x = -a_{xx}$, plus the residual time in a new state y after possible transition from the state x, which takes place with probability a_{xy}/a_x. Thus from the low of total probability for the density $w_x(t)$ we get

$$w_x(t) = \sum_{y \ne x} \frac{a_{xy}}{a_x}\left[\lambda_x e^{-a_x t} * w_y(t)\right],\ x \in E, \tag{16}$$

where $*$ denotes convolution. Applying the LST to the relation (16) we get

$$\tilde{w}_x(s) = \sum_{y \ne x} \frac{a_{xy}}{s + a_x}\tilde{w}_y(s),\ x \in E. \tag{17}$$

Theorem 1 The LST $\tilde{\mathbf{w}}_j(t),\ 1 \le j \le N$ of the conditional waiting times satisfy the following recurrent block system

$$\Lambda_{W,j}(s)\tilde{\mathbf{w}}_j(s) = -\Gamma_j \tilde{\mathbf{w}}_{j-1}(s),\ 1 \le j \le N,$$

$$\tilde{\mathbf{w}}_0(s) = \mathbf{e}(4(N+1)),$$

where square matrices $\Lambda_{W,j}(s) = (\Phi_j - sI_{4(N-j+1)})$ have dimensions $4(N-j+1)$ and the rectangular matrices Γ_j have dimensions $4(N-j+1) \times 4(N-j+2)$. Φ_j is obtained from Λ by removing the first j block-rows and block-columns as well as the lower diagonal

$$\Phi_j = \begin{pmatrix} A_{10} & A_{01} & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & A_{10} & A_{01} & 0 & 0 & 0 & \dots & 0 \\ \ddots & \ddots & \ddots & \ddots & \ddots & & \ddots & \\ 0 & \dots & 0 & A_{11} & A_{02} & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & A_{12} & A_{02} & \dots & 0 \\ \ddots & \ddots & \ddots & \ddots & \ddots & & \ddots & \\ 0 & \dots & 0 & 0 & 0 & 0 & A_{12} & A_{02} \\ 0 & \dots & 0 & 0 & 0 & 0 & 0 & A_{13} \end{pmatrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \} q_2 - 1 \\ \end{matrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \} N - j + 1, \end{matrix}$$

$$\Gamma_j = \Delta(\overbrace{\underbrace{A_{21}, ..., A_{21}}_{q_2 - 1}, A_{22}, ..., A_{22}}^{N-j+1}, \mathbf{0}),$$

where $\Delta(a_1, ..., a_m, \mathbf{0})$ denotes a diagonal block-matrix with blocks $a_1, ..., a_m$ and additional last zero block-column.

Proof: If after a transition $J(t) = 0$, then a tagged customer must immediately be sent to one of idle servers and the remaining waiting time is 0, i.e. $\tilde{w}_x(s) = 1,\ x = (q, d_1, d_2, 0) \in E$. For all other states and positions of the tagged customer we obtain according to equation (17)

$$
\begin{aligned}
&(s+\lambda \mathbf{1}_{\{q(x)<N\}}+\sum_{k=1}^{2} d_k\mu_k+\gamma\sum_{k=1}^{2}\mathbf{1}_{\{A_{qk(x)}\}})\tilde{w}_x(s)\\
&=\lambda\tilde{w}_{x+e_0}(s)\mathbf{1}_{\{\bar{A}_{q_2-1}\vee\sum_{k=1}^{2}d_k(x)=2,\ \ q(x)<N\}}+\lambda\sum_{k=1}^{2}\tilde{w}_{x+e_k}(s)\mathbf{1}_{\{A_{qk-1}(x)\}} \qquad (19)\\
&\quad+\sum_{k=1}^{2} d_k\mu_k\tilde{w}_{x-e_k}(s)+\gamma\sum_{k=1}^{K}\tilde{w}_{x-e_0-e_3+e_k}(s)\mathbf{1}_{\{A_{qk}(x)\}},
\end{aligned}
$$

where $A_{qk}(x)$ and $\bar{A}_{qk}(x)$ denote the following event and its complement

$$A_{qk}(x)=\{q(x)\ge qk,\ d_i(x)=1,\ i=\overline{1,k-1},\ d_k(x)=0\},$$

$$\bar{A}_{qk}(x)=\{q(x)\le qk-1,\ d_i(x)=1,\ i=\overline{1,k-1},\ d_k(x)=0\}.$$

After routine block identification we may express this system of equations as in (18).

The tagged customer must wait in the orbit if upon arrival it finds the system in some state of the subset

$$E_W=\{(q,1,0);\ 0\le q\le q_2-1\}\cup\{(q,1,1);\ 0\le q\le N-1\}.$$

Denote by $\boldsymbol{\pi}_W=(\hat{\pi}_1,\ \hat{\pi}_2,\ldots,\ \hat{\pi}_N)$ the row-vector of dimension $2N(N+1)$ which contains all stationary probabilities in set E_W,

$$\hat{\pi}_i=(0,\pi_{(i-1,1,0)},0,\pi_{(i-1,1,1)},\underbrace{0,\ldots,0}_{4(N-i)}),\ 1\le i\le q_2-1,$$

$$\hat{\pi}_i=(0,0,0,\pi_{(i-1,1,1)},\underbrace{0,\ldots,0}_{4(N-i)}),\ q_2\le i\le N.$$

According to the PASTA property the conditional probability of the state x^- upon arrival coincides with the unconditional one.

Corollary 1 For the unconditional LT of the waiting time with respect to all possible initial states x of the process $X(t)$ and the corresponding states x^- before an arrival we have

$$\tilde{W}(s)=\frac{1}{s}(1-\boldsymbol{\pi}_W e+\boldsymbol{\pi}_W\tilde{\mathbf{w}}(s)), \qquad (20)$$

where the contribution

$$1-\boldsymbol{\pi}_W\mathbf{e}=1-\left[\sum_{q=0}^{q_2-2}\pi_{(q,1,0)}+\sum_{q=0}^{N-1}\pi_{(q,1,1)}\right]=\sum_{q=0}^{N}\left[\pi_{(q,0,1)}+\pi_{(q,0,0)}\right]+\sum_{q=q_2-1}^{N}\pi_{(q,1,0)},$$

represents the stationary probability that a tagged customer does not have to wait for service; the contribution

$$\boldsymbol{\pi}_W\tilde{\mathbf{w}}(s)=\sum_{q=0}^{q_2-2}\pi_{(q,1,0)}\tilde{w}_{(q+1,1,0,q+1)}(s)+\sum_{q=0}^{N-1}\pi_{(q,1,1)}\tilde{w}_{(q+1,1,1,q+1)}(s),$$

represents the LST of the waiting time with density function $\omega_c(t)$ given $W>0$, $\int_0^\infty w_c(u)\mathrm{d}u=\boldsymbol{\pi}_W\mathbf{e}$.

Due to a Tauberian result we apply the limit property of the LST to get the value of the function $w_c(t)$ at the point $t=0$:

$$\lim_{s\to\infty} s\tilde{w}_{(q,d_1,d_2,j)}(s)=\begin{cases}\gamma, & \text{if } q_2\le q\le N,\ d_1=1,\ d_2=0,\ j=1\\ \gamma, & \text{if } 1\le q\le N,\ d_1=0,\ d_2=1,\ j=1\\ \gamma, & \text{if } 1\le q\le N,\ d_1=0,\ d_2=0,\ j=1\\ 0, & \text{otherwise}\end{cases}.$$

Therefore we have

$$\lim_{t\to 0} w_c(t)=\lim_{s\to\infty} s\boldsymbol{\pi}_W\tilde{\mathbf{w}}(s)=0.$$

The inversion of the LST $(1/s)\boldsymbol{\pi}_W\tilde{\mathbf{w}}(s)$ for the distribution function $W_c(t)=\int_0^t w_c(u)\mathrm{d}u$ is used to get the distribution function of the unconditional waiting time

$$W(t)=P[W\le t]=1-\boldsymbol{\pi}_W\mathbf{e}+W_c(t),\ t\ge 0. \tag{21}$$

Remark 2 In principle it is also possible to calculate the LST $\pi_W\tilde{w}(s)$ for $N=\infty$ taking into account that the conditional LST $\tilde{w}_{(j+q_2-1,1,1,j)}(s),\ j\ge 1$ does not depend on the future arrival to the orbit that implies the finite system of equations for each j. However, for an infinite orbit we get an infinite sum of rational functions that must in any case be approximated by a finite sum in order to evaluate the inverse LST. This requires a careful choice of stopping parameters.

Now we obtain the n-th moments of W_x which we denote by $\bar{W}_x(n)=\mathbf{E}[W_x^n]$. Let $\bar{\mathbf{W}}(n)$ be the vector of the moments with subvectors $\bar{\mathbf{W}}_j(n),\ 1\le j\le N$, partitioned in the same way as $\tilde{\mathbf{w}}(s)$. By differentiating the expression (18) with respect to s and taking into account that $\bar{\mathbf{W}}_j(n)=(-1)^n(d^m/ds^n)\tilde{\mathbf{w}}_j(s)\big|_{s=0}$ and $\Lambda_{W,j}(0)=\Phi_j$, we obtain the arbitrary moments of the waiting time.

Corollary 2 The conditional moments of the n-th order $\bar{\mathbf{W}}_j(n),\ 1\le j\le N$ of the waiting time satisfy the following recurrent block system

$$\begin{aligned}&\Phi_j\bar{\mathbf{W}}_j(n)=-n\bar{\mathbf{W}}_j(n-1)-\Gamma_j\bar{\mathbf{W}}_{j-1}(n),\ 1\le j\le N, n\ge 1,\\ &\bar{\mathbf{W}}_0(n)=\mathbf{0}(4(N+1)),\ n\ge 1,\\ &\bar{\mathbf{W}}_j(0)=\mathbf{e}(4(N-j+1)),\ 1\le j\le N.\end{aligned} \tag{22}$$

Corollary 3 The unconditional moments $\mathbf{E}[W^n]$ of the waiting time are given by

$$\mathbf{E}[W^n]=\boldsymbol{\pi}_W\bar{\mathbf{W}}(n),\ n\ge 0. \tag{23}$$

To carry out the calculation of the sojourn time distribution it should be pointed out that the service time depends on the customers arriving at the orbit after the tagged customer. Therefore we apply the same approach used above with only one exception that the absorption of the process (15) takes place at the time when a tagged customer completes the service. Define

$\tilde{\mathbf{t}}(s)=(\tilde{\mathbf{t}}_0(s),\tilde{\mathbf{t}}_1(s),\ldots,\tilde{\mathbf{t}}_N(s))^t$-column-vector, where $\tilde{\mathbf{t}}_j(s),\ 1\le j\le N$ contains the

conditional LSTs $\tilde{t}_x(s)$ partitioned as $\tilde{w}_x(s)$ and $\tilde{\mathbf{t}}_0(s) = (\mu_1/(s+\mu_1),\ \mu_2/(s+\mu_2))^t$.

Let $\overline{\mathbf{T}}(n)$ be the vector of he moments with subvectors $\overline{\mathbf{T}}_j(n),\ 0 \le j \le N,$ where $\overline{\mathbf{T}}_0(n) = (n!/\mu_1^n, n!/\mu_2^n)^t$, $\tilde{t}(s)$ and $\tilde{T}(s)$- the unconditional LST and LT.

The vectors $\tilde{\mathbf{t}}(s)$ and $\overline{\mathbf{T}}(n)$ can be calculated using the same arguments adduced in the previous section. We omit the details and summarize the results in the following statements.

Theorem 2 The LST $\tilde{\mathbf{t}}_j(s),\ 1 \le j \le N,$ of the conditional sojourn times satisfy the following recurrent block system

$$\begin{aligned} &\Lambda_{W,1}(s)\tilde{\mathbf{t}}_1(s) = -H_1\mathbf{e}(4(N+1))\frac{\mu_1}{s+\mu_1} - H_2\mathbf{e}(4(N+1))\frac{\mu_2}{s+\mu_2}, \\ &\Lambda_{W,j}(s)\tilde{\mathbf{t}}_j(s) = -\Gamma_j\tilde{\mathbf{t}}_{j-1}(s),\ 2 \le j \le N, \end{aligned} \tag{24}$$

where $H_1 = \Delta(\underbrace{A_{21},\ldots,A_{21},}_{N}\ \mathbf{0}),\ H_2 = \Gamma_1 - H_1.$

Corollary 4 The conditional moments of the n-th order $\overline{\mathbf{T}}_j(n),\ 1 \le j \le N$ of the sojourn time satisfy the following recurrent block system

$$\Phi_1\overline{\mathbf{T}}_1(n) = -n\overline{\mathbf{T}}_1(n-1) - H_1\mathbf{e}(4(N+1))\frac{n!}{\mu_1^n} - H_2\mathbf{e}(4(N+1))\frac{n!}{\mu_2^n},\ n \ge 1,$$

$$\Phi_j\overline{\mathbf{T}}_j(n) = -n\overline{\mathbf{T}}_j(n-1) - \Gamma_j\overline{\mathbf{T}}_{j-1}(n),\ 2 \le j \le N, n \ge 1,$$

$$\overline{\mathbf{T}}_j(0) = \mathbf{e}(4(N-j+1)),\ 1 \le j \le N.$$

A tagged customer joins the system if upon arrival it finds the system state in the set

$$E_T = \{x = (q, d_1, d_2) : d_1 + d_2 = \{0,1\},\ 0 \le q \le N\} \cup \{x = (q,1,1) : 0 \le q \le N-1\}.$$

Denote by $\boldsymbol{\pi}_T = (\hat{\pi}_0,\ \hat{\pi}_1,\ldots,\ \hat{\pi}_N) = (\hat{\pi}_0, \boldsymbol{\pi}_W)$ with $\hat{\pi}_0 = (\sum_{i=0}^{N}[\pi_{(i,0,1)} + \pi_{(i,0,0)}],\ \sum_{i=q_2-1}^{N}\pi_{(i,1,0)})$ a row-vector which contains all stationary probabilities in set E_T.

Corollary 5 The unconditional LT of the sojourn time with respect to all possible initial states x is given by

$$\tilde{T}(s) = \frac{1}{s}\boldsymbol{\pi}_T\tilde{\mathbf{t}}(s), \tag{26}$$

and componentwise

$$\begin{aligned} \boldsymbol{\pi}_T\tilde{\mathbf{t}}(s) = &\sum_{i=0}^{N}\left[\pi_{(i,0,1)} + \pi_{(i,0,0)}\right]\frac{\mu_1}{s+\mu_1} + \sum_{i=q_2-1}^{N}\pi_{(i,1,0)}\frac{\mu_2}{s+\mu_2} \\ &+ \sum_{i=0}^{q_2-2}\pi_{(i,1,0)}\tilde{t}_{(i+1,1,0,i+1)}(s) + \sum_{i=0}^{N-1}\pi_{(i,1,1)}\tilde{t}_{(i+1,1,1,i+1)}(s), \end{aligned}$$

where the first two sums in the right hand side represent the LST of the service time when the customer upon arrival has a direct access to the first or second server, respectively, and the last two sums represent the LST of the sojourn time in case that the customer must wait in orbit.

Corollary 6 The unconditional moments $\mathbf{E}[T^n]$ of the sojourn time are given by

$$\mathbf{E}[T^n] = \boldsymbol{\pi}_T \tilde{\mathbf{T}}(n), \ n \geq 0. \tag{27}$$

Let $t(\tau)$ denotes the unconditional density associated with the Laplace transform $\tilde{t}(s)$. Its value at point τ=0 satisfies

$$\lim_{x\to\infty} \tilde{st}_{(q,\,d_1,\,d_2,\,j)}(s) = \begin{cases} \mu_1, & \text{if } 0 \leq q \leq N,\ d_1 = 0,\ d_2 = 1,\ j = 0 \\ \mu_2, & \text{if } 0 \leq q \leq N,\ d_1 = 1,\ d_2 = 0,\ j = 0. \\ 0, & \text{otherwise} \end{cases}$$

Hence,

$$\lim_{\tau\to 0} t(\tau) = \lim_{s\to\infty} s\boldsymbol{\pi}_T \tilde{\mathbf{t}}(s) = \mu_1 \sum_{i=0}^{N} \boldsymbol{\pi}_i(\mathbf{e}_0 + \mathbf{e}_3) + \mu_2 \sum_{i=q_2-1}^{N} \boldsymbol{\pi}_i \mathbf{e}_1.$$

4. The Number of Retrials Made by a Customer

In this section we consider the random value Ψ of the number of retrials made by a tagged customer until it reaches the service area. This descriptor provides a discrete counterpart of the waiting time W studied in Section 3 and complement the present analysis . Denote by

Ψ_x – the number of retrials made by a tagged customer given that the system is in state x,

$\psi_x(k) = \mathbf{P}[\Psi_x = k]$: the conditional density function of the r.v. Ψ_x,

$\tilde{\psi}_x(z) = E[z^{\Psi_x}] = \sum_{k=0}^{\infty} \psi_x(k) z^k$, $|z| \leq 1$: the corresponding generating function,

$\tilde{\boldsymbol{\psi}}(z)$ and $\bar{\boldsymbol{\Psi}}(n)$: the column-vectors of generating functions $\tilde{\psi}_x(z)$ and n-th factorial moments $\bar{\Psi}_x(n) = \mathbf{E}[\Psi_x(\Psi_x - 1)...(\Psi_x - n + 1)]$, $n \geq 1$, with elements partitioned as before.

For the conditional density $\psi_x(k)$ via the law of total probability the following equality holds

$$\psi_x(k) = \frac{a_{xy'}}{a_x} \psi_{y'}(k-1) + \sum_{y \neq x,\, y'} \frac{a_{xy}}{a_x} \psi_y(k), \tag{28}$$

where the first term in the right hand side stands for the transition due to the retrial, whereas the second term includes other possible transitions. In terms of generation function the latter equality can be expressed as

$$\tilde{\psi}_x(z) = \frac{z a_{xy'}}{a_x} \tilde{\psi}_{y'}(z) + \sum_{y \neq x,\, y'} \frac{a_{xy}}{a_x} \tilde{\psi}_y(z). \tag{29}$$

The following theorem adjusts the relations for the generating functions $\tilde{\psi}_x(z)$.

Theorem 3 The generating functions $\tilde{\psi}_j(z)$, $1 \leq j \leq N$, of the conditional number of retrials are related by the following recurrent block system

$$\Lambda_{\Psi,\,1(z)} \tilde{\psi}_1(z) = -z\Gamma_1 e(4(N+1)), \Phi_j \tilde{\psi}_j(z) = -\Gamma_j \tilde{\psi}_{j-1}(z), \ 2 \leq j \leq N, \tag{30}$$

where $\Lambda_{\Psi,1(z)} = (\Phi_1 + (1-z)V)$ and

$$V = -\frac{\gamma}{\lambda} diag(\overbrace{\underbrace{A_{01},\ldots,A_{01}}_{q_2-1}, A_{02},\ldots,A_{02}}^{N}).$$

Proof: Similarly to the previous section, with respect to (29), we see that

$$\begin{aligned}
&\left(\lambda \mathbf{1}_{\{q(x)<N\}} + \sum_{k=1}^{2} d_k\mu_k + \gamma \sum_{k=1}^{2} \mathbf{1}_{\{A_{qk(x)} \vee \bar{A}_{qk}(x),\, j(x)=1\}} + \gamma \mathbf{1}_{\{\sum_{k=1}^{2} d_k=2,\, j(x)=1\}}\right)\tilde{\psi}_x(z) \\
&= \lambda \tilde{w}_{x+e_0}(s)\mathbf{1}_{\{\bar{A}_{q_2-1} \vee \sum_{k=1}^{2} d_k(x)=2,\, q(x)<N\}} + \lambda \sum_{k=1}^{2} \tilde{w}_{x+e_k}(s)1_{\{A_{qk-1}(x)\}} + \sum_{k=1}^{2} d_k\mu_k \tilde{w}_{x-e_k}(s) \\
&\quad + \gamma \sum_{k=1}^{K} \tilde{w}_{x-e_0-e_3+e_k}(s)\mathbf{1}_{\{A_{qk}(x)\}} + z\gamma \sum_{k=1}^{2} \tilde{\psi}_x(z)\mathbf{1}_{\{\bar{A}_{qk}(x),\, j(x)=1\}} + z\gamma \mathbf{1}_{\{\sum_{k=1}^{2} d_k=2,\, j(x)=1\}}.
\end{aligned} \tag{31}$$

For the derivation of this system we take into account that the retrial in state $x \in \hat{E}$ can be performed by the tagged customer only if it stays at the head of the orbit, i.e. $j(x) = 1$. If according to the threshold policy a service area is ready to accept a customer, a retrial means that the customer gets a service. Otherwise a retrial does not change a system state but nevertheless must be counted.

Corollary 7 The unconditional version of the generating function is of the form

$$\tilde{\psi}(z) = 1 - \pi_W e + \pi_W \tilde{\psi}(z). \tag{32}$$

To get the factorial moments we differentiate the relations (30) taking into account that $\bar{\Psi}_j(n) = (d^n / dz^n)\psi_j(z)\big|_{z=1}$ and $\Lambda_{\Psi,1}(1) = \Phi_1$.

Corollary 8 The moments of the n-th order $\bar{\mathbf{\Psi}}_j(n)$, $1 \le j \le N$, of the conditional number of retrials satisfy the following recurrent block system

$$\begin{aligned}
&\Phi_1 \bar{\mathbf{\Psi}}_1(n) = -\Gamma_1 \mathbf{e}(N+1)1_{\{n=1\}} + nV\bar{\mathbf{\Psi}}_1(n-1),\ n \ge 1, \\
&\Phi_j \bar{\mathbf{\Psi}}_j(n) = -\Gamma_j \bar{\mathbf{\Psi}}_{j-1}(n),\ 2 \le j \le N, n \ge 1, \\
&\bar{\mathbf{\Psi}}_j(0) = \mathbf{e}(4(N-j+1)),\ j \ge 1.
\end{aligned} \tag{33}$$

Corollary 9 The unconditional moments $\mathbf{E}[\Psi(\Psi-1)\ldots(\Psi-n+1)]$ satisfy

$$\mathbf{E}[\Psi(\Psi-1)\ldots(\Psi-n+1)] = \boldsymbol{\pi}_W \bar{\mathbf{\Psi}}(n),\ n \ge 0. \tag{34}$$

Corollary 10 The recurrent substitution in (22) and (23) as well as an equality

$$-\Gamma_1 \mathbf{e}(4(N+1)) + V\mathbf{e}(4N) = -\gamma \mathbf{e}(4N),$$

lead to the relation between the first conditional moments of the waiting time and number of retrials made by a customer,

$$\bar{\mathbf{W}}_j(1) = \frac{1}{\gamma}\bar{\mathbf{\Psi}}_j(1) + \sum_{i=1}^{j-1} (-1)^i \Phi_j^{-1} \prod_{k=1}^{i-1} \Gamma_{j-k+1} \Phi_{j-k}^{-1} e(4(N-j+i)),\ 1 \le j \le N. \tag{35}$$

5. Optimization Problem

The mean performance characteristics introduced in Section 2 depend on the threshold level q_2 and a natural question that may arise in practice is a calculation of an optimal threshold policy (OTP). Thereby we look for the threshold level q_2^* which leads to the minimum of the mean sojourn time of customers in the system $\bar{T}$. By Little's formula this is equivalent to the minimizing the mean number of customers in the system $\bar{N}$. Formally the problem can be written as follows

$$\bar{N} = \bar{N}(\lambda, \mu_1, \mu_2, \gamma, q_2) \to \min_{q_2}. \tag{36}$$

To reduce the number of system parameters we divide all of them by λ. Now solving (36) we find the corresponding optimal threshold levels,

$$q_2^*(\frac{\mu_1}{\lambda}, \frac{\mu_2}{\lambda}, \frac{\gamma}{\lambda}) = \arg\min_{q_2} \bar{N}(\frac{\mu_1}{\lambda}, \frac{\mu_2}{\lambda}, \frac{\gamma}{\lambda}, q_2).$$

To calculate this value a simple exhaustion method is quite appropriate if the state space is not large.

Analytical representation of optimal threshold level is quite complicated but if future arrivals are not taken into account (scheduling problem), i.e. when $\lambda = 0$ and the objective is to minimize the sojourn time for an individual customer, the explicit solution can be evaluated.

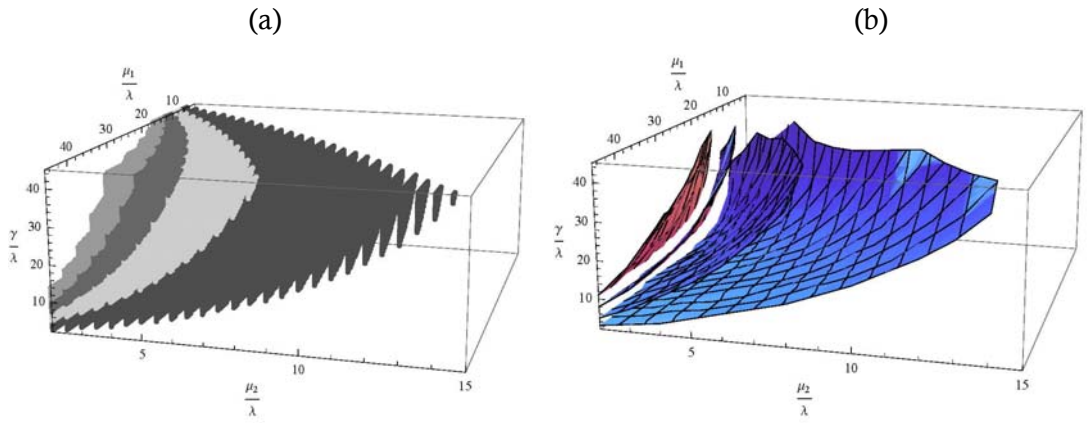

Figure 1. Areas with optimal threshold level $q_2^* = 1, 2, 3, 4$ and corresponding asymptotic surfaces.

Theorem 4 (Scheduling threshold policy) If $\lambda = 0,$ then there exists an optimal threshold level

$$q_2^* = \left\lfloor \frac{\gamma}{\mu_1 + \gamma}\left(\frac{\mu_1}{\mu_2} - 1\right)\right\rfloor, \tag{37}$$

with the following policy: If $q \geq q_2^* - 1$ in state $x = (q, 1, 0),$ then a retrial arrival is dispatched to the slower server. If $q < q_2^* - 1,$ then a retrial is sent back to the orbit.

Proof: Denote by $\bar{V}_x$ the mean sojourn time given an initial state x. Assume that there are already $n > q_2$ customers in the system and consider the problem to make the system idle as soon as possible. The value $\bar{V}_x,\ x \in E$ obviously satisfy the relations

$$\begin{aligned}
&\bar{V}_{(0,1,1)} = \frac{\mu_1+\mu_2}{\mu_1\mu_2},\ \bar{V}_{(0,0,1)} = \frac{1}{\mu_2},\ \bar{V}_{(q-1,1,0)} = \frac{q}{\mu_1} + \bar{V}_{(q-1,0,0)} = \frac{q}{\mu_1} + \frac{(q-1)q}{2}\left(\frac{1}{\gamma}+\frac{1}{\mu_1}\right),\\
&\bar{V}_{(q-1,1,1)} = \frac{1}{\mu_1+\mu_2}(q+1+\mu_1\bar{V}_{(q-1,0,1)} + \mu_2\bar{V}_{(q-1,1,0)}) = \frac{1}{\mu_2} + \bar{V}_{(q-1,1,0)}.
\end{aligned} \tag{38}$$

Since q_2^* is an optimal orbit size to switch on the second server, the following condition must hold: The function $\bar{V}_x$ in the state $x = (q-1,1,1)$ with $q \geq q_2^*$, in which the second server is activated according to the threshold policy, must be less then the function in the state in which the second server is idle, i.e.

$$\bar{V}_{(q-1,1,1)} \leq \bar{V}_{(q,1,0)},\ q \geq q_2^*.$$

Substituting the corresponding expressions (38) into the last inequality we get

$$\frac{1}{\mu_2} + \frac{q}{\mu_1} + \frac{(q-1)q}{2}\left(\frac{1}{\gamma}+\frac{1}{\mu_1}\right) \leq \frac{q+1}{\mu_1} + \frac{q(q+1)}{2}\left(\frac{1}{\gamma}+\frac{1}{\mu_1}\right).$$

Thus the second server must be switched on if

$$q \geq q_2^* = \left\lfloor \frac{\gamma}{\mu_1+\gamma}\left(\frac{\mu_1}{\mu_2}-1\right)\right\rfloor.$$

Remark 3 A numerical analysis of the optimal threshold level q_2 shows that the formula (37) defines the optimal threshold level also for a system with arrival rate $\lambda < \mu_2$.

The problem (36) in general case for $\lambda > 0$ can be solved also in the following way. For the optimal threshold level we identify the regions where a certain level is optimal, e.g. $q_2^* = 1, 2, 3, 4$. This is graphically represented in Figure 1. Using the formula (37) we assume that $q_2^* = \gamma/(\mu_1+\gamma)\hat{q}_2^*$, where $\hat{q}_2^*$ denotes the optimal threshold level for the equivalent ordinary $M/M/2$ queueing system with heterogeneous servers. Then we fit the surfaces between each boundary area by means of the three-dimensional function with parameters $\mu_1/\lambda, \mu_2/\lambda, \gamma/\lambda$, estimating the value $\hat{q}_2^*$ in the same way as it was done in [21]. The following conjecture establishes an approximate relationship between the system parameters $\lambda, \mu_1, \mu_2, \gamma$ and optimal threshold level q_2^* with performance error that never exceeds 1.5 percent.

Conjecture 1 (Approximation to the optimal threshold level) Using the asymptote to the boundary between the area where the optimal threshold is q_2^* and the area with optimal threshold q_2^*+1, an approximation of the optimal threshold level is

$$q_2^* \approx \tilde{q}_2^* = \left\lfloor \frac{\gamma}{\gamma+\mu_1}\left(\frac{\mu_1-\lambda+\sqrt{(\mu_1-\lambda)^2-4\mu_2\lambda}}{2\mu_2}-1\right)\right\rfloor. \tag{39}$$

6. Numerical Examples

Consider the system $M/M/2$ with primary arrival rate λ, retrial rate γ and service rates μ_1 and μ_2. To demonstrate the advantage of an optimal threshold policy we examine additionally three heuristic policies. The Scheduling threshold policy (STP) is defined by the threshold level (37). The Fastest free server selection (FFS) policy prescribes the usage of fastest available server upon an arrival of a primary or a retrial customer. This policy is a threshold policy with level $q_2 = 1$. The Random server selection (RSS) policy with equal probability chooses any free server. Further we consider the system with Homogeneous servers (HS) and the same total service intensity.

By means of *Mathematica* package we have created the procedures

- for the calculation of the stationary probability vector $\boldsymbol{\pi}$, see formulas (3-6)
- for the calculation of the performance measures $\bar{U}, \bar{N}, \bar{W}$ and $\bar{T}$, see formulas (7, 10-12)
- for the derivation of the LTs $\tilde{W}(s)$ and $\tilde{T}(s)$ given by (20) and (26)
- for the derivation of the generating function $\tilde{\psi}(z)$ given by (32).

By inverting the derived LTs (20) and (26) it is possible to evaluate the waiting and sojourn time distribution functions. There are two practical algorithms for the inversion of Laplace transform: The conventional algorithm and the Fourier-series algorithm. Mathematical packages such as *Mathematica*, *Matlab*, *Mathcad*, etc. include standard functions that apply the conventional algorithm. It is based on the partial function expansion method and works efficiently only in case of small order rational functions. Hence we can obtain an accurate representation of the functions $W(t)$ and $T(t)$ only for small t but in symbolic form. The algorithms based on Fourier-series represent numerical inversion methods, e.g. Euler and Post-Widder [2], that can be used for large t as well. The mentioned mathematical software further allows us to invert generating function (32) in order to get the distribution function $\Psi(n)$. However, there are problems inverting functions of higher order. To this aim we have implemented a numerical inversion of generating functions using the Lattice-Poisson algorithm [1].

6.1. Mean Performance Analysis

Next we present some numerical results to show the effect of parameters on the mean performance characteristics. In Tables 1 and 2 the performance measures for the system under different allocation policies and the homogeneous system are given for varying values of primary arrival, retrial and service rates, respectively. We notice the following interesting observation from these tables. As it can be seen, the OTP performs better in terms of the mean number of customers in the system (mean sojourn time) than other heuristic control policies. A comparison of the OTP and the HS system shows that for some values of parameters, e.g. $(\lambda, \mu_1, \mu_2, \gamma) = (0.1, 2.2, 0.3, 8.5)$, the first system is 40% superior to the second one. But this advantage decreases with increasing of λ and decreasing of γ. If $(\lambda, \mu_1, \mu_2, \gamma) = (1.3, 2.2, 0.3, 1.5)$ the HS system is 40% more effective then OTP. The efficiency of the OTP system also decreases with increasing heterogeneity level μ_1 / μ_2 for small λ, e.g. $\lambda = 0.1$, and with decreasing heterogeneity level for large λ, e.g. $\lambda = 1.3$.

Table 1. Utilization and mean number of customers in the system.

$\lambda, \mu_1, \mu_2, \gamma$	$\bar{U}^{(OTP)}$	$\bar{U}^{(STP)}$	$\bar{U}^{(FFS)}$	$\bar{U}^{(RSS)}$	$\bar{U}^{(HS)}$	$\bar{N}^{(OTP)}$	$\bar{N}^{(STP)}$	$\bar{N}^{(FFS)}$	$\bar{N}^{(RSS)}$	$\bar{N}^{(HS)}$
0.1,2.2,0.3,8.5	0.0459	0.0459	0.0552	0.1648	0.0769	0.0482	0.0482	0.0575	0.1717	0.0802
0.1,2.2,0.3,1.5	0.0485	0.0484	0.0554	0.1650	0.0771	0.0510	0.0510	0.0577	0.1722	0.0803
0.1,1.3,1.2,8.5	0.0744	0.0744	0.0744	0.0771	0.0769	0.0775	0.0775	0.0775	0.0803	0.0802
0.1,1.3,1.2,1.5	0.0745	0.0745	0.0745	0.0772	0.0771	0.0777	0.0777	0.0777	0.0805	0.0803
1.3,2.2,0.3,8.5	0.6775	0.6706	0.6706	0.8526	0.7034	1.6820	1.8538	1.8520	2.0310	1.5904
1.3,2.2,0.3,1.5	0.8692	0.8799	0.8814	0.9364	0.8043	4.4750	4.6931	4.4960	5.5240	2.8376
1.3,1.3,1.2,8.5	0.6992	0.6992	0.6992	0.7037	0.7034	1.5799	1.5799	1.5800	1.5900	1.5904
1.3,1.3,1.2,185	0.7997	0.7997	0.7997	0.8040	0.8043	2.7940	2.7940	2.7940	2.8250	2.8376

Table 2. Mean waiting time and mean sojourn time of a customer.

$\lambda, \mu_1, \mu_2, \gamma$	$\bar{W}^{(OTP)}$	$\bar{W}^{(STP)}$	$\bar{W}^{(FFS)}$	$\bar{W}^{(RSS)}$	$\bar{W}^{(HS)}$	$\bar{T}^{(OTP)}$	$\bar{T}^{(STP)}$	$\bar{T}^{(FFS)}$	$\bar{T}^{(RSS)}$	$\bar{T}^{(HS)}$
0.1,2.2,0.3,8.5	0.0275	0.0275	0.0012	0.0036	0.0017	0.4821	0.4821	0.5754	1.7173	0.8017
0.1,2.2,0.3,1.5	0.0546	0.0546	0.0025	0.0074	0.0034	0.5096	0.5096	0.5765	1.7218	0.8034
0.1,1.3,1.2,8.5	0.0016	0.0016	0.0016	0.0017	0.0017	0.7753	0.7753	0.7753	0.8028	0.8017
0.1,1.3,1.2,1.5	0.0033	0.0033	0.0033	0.0034	0.0034	0.7770	0.7770	0.7770	0.8046	0.8034
1.3,2.2,0.3,8.5	0.6437	0.9152	0.5253	0.5779	0.4234	1.2939	1.4260	1.4246	1.5623	1.2234
1.3,2.2,0.3,1.5	2.5848	2.8059	2.5317	3.2481	1.3828	3.4421	3.6101	3.4586	4.2493	2.1827
1.3,1.3,1.2,8.5	0.4202	0.4202	0.4202	0.4231	0.4234	1.2154	1.2154	1.2154	1.2233	1.2234
1.3,1.3,1.2,185	1.3545	1.3545	1.3545	1.3728	1.3828	2.1492	2.1492	2.1492	2.1727	2.1828

The following figures represent the system utilization (Figure 2(a, b)), mean number of customers in the system (Figure 3(a, b)), mean waiting time (Figure 4(a, b)) and mean sojourn time of customers (Figure 5(a, b)) for service intensities $\mu_1 = 2.2, \mu_2 = 0.3$. The retrial intensity $\gamma = 2.5$ for figures labeled by "a" and $\gamma = 18.5$ for figures labeled by "b", with varying primary arrival intensity $0.05 \le \lambda \le 1.7$.

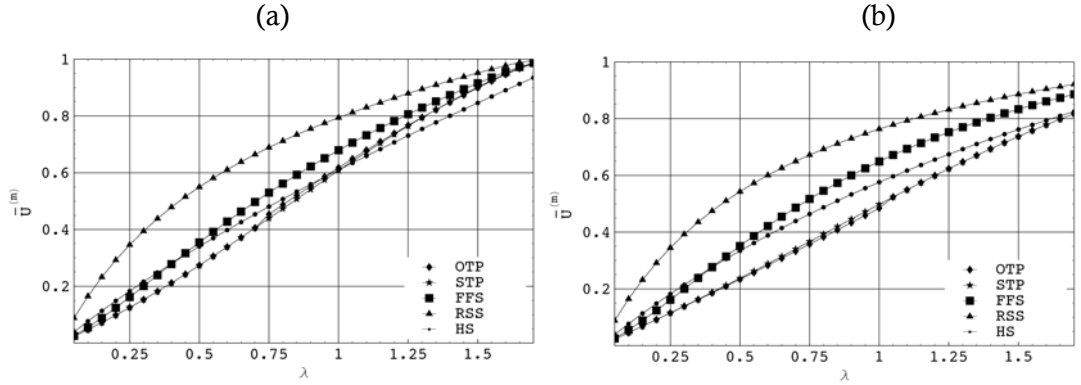

Figure 2. System utilization versus primary arrival and retrial rate (a) $\gamma = 2.5$ (b) $\gamma = 18.5$ Analyzing the figures we have noticed the following.

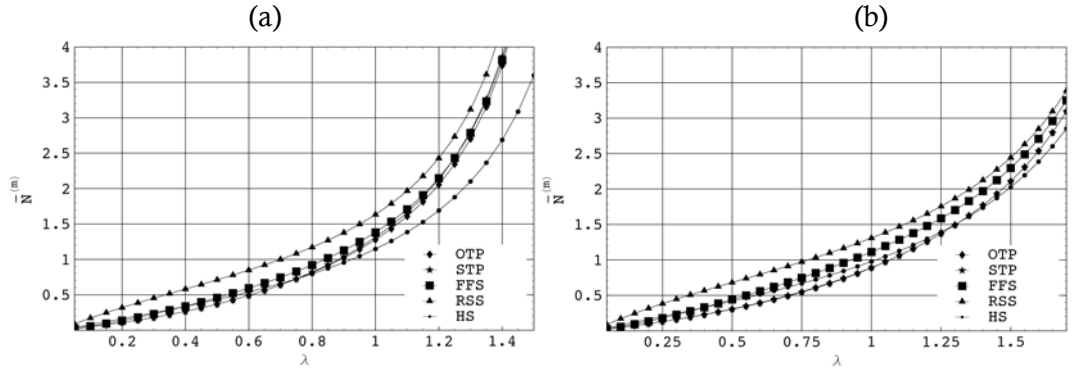

Figure 3. Mean number of customers versus primary arrival and retrial rate (a) $\gamma = 2.5$ (b) $\gamma = 18.5$.

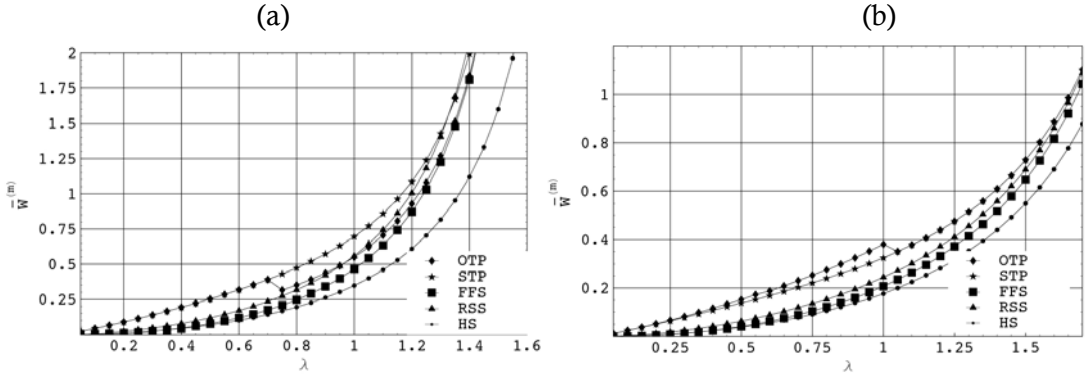

Figure 4. Mean waiting time versus primary arrival and retrial rate (a) $\gamma = 2.5$ (b) $\gamma = 18.5$.

1. When the primary arrival rate λ makes a relative small contribution to the load factor ρ of the system, then the OPT and STP coincide. This leads to equal values of the performance measures. Otherwise the policies as well as the performance measures are different.
2. When the primary arrival rate is quite large and the load factor tends to 1 ("heavy traffic") then difference between the policies can be neglected.
3. The curve of the mean waiting time for the FFS lies below other graphs, i.e. this policy minimizes the mean number of jobs in the orbit. This does not contradict with the optimality of the OTP since this policy optimizes the mean sojourn time. This is confirmed by the corresponding figure.
4. All curves are monotone except for the mean waiting time in case of the OTP and mean sojourn time in case of the RSS policy. In the first case, the threshold levels of the OTP decrease as λ increases. This leads to a reduction of the waiting time in the queue. In the second case, the convexity of the curve is connected with the mean service time that contributes to the mean sojourn time. If the rate is very small a new arrival most likely sees two servers idle and will occupy them with equal probability under the RSS policy. Upon increasing the arrival rate the slower server will be made more busy and fastest server is used more intensively, resulting in less service time.
5. In comparison to classical queues [15], where the system with heterogeneous servers under OTP is always superior to the homogeneous one with the same total service

time, as we have noticed, retrial queues with homogeneous servers are better than heterogeneous ones if $\mu_2 < \lambda$ for large μ_1 / λ and small γ / λ or vice versa.

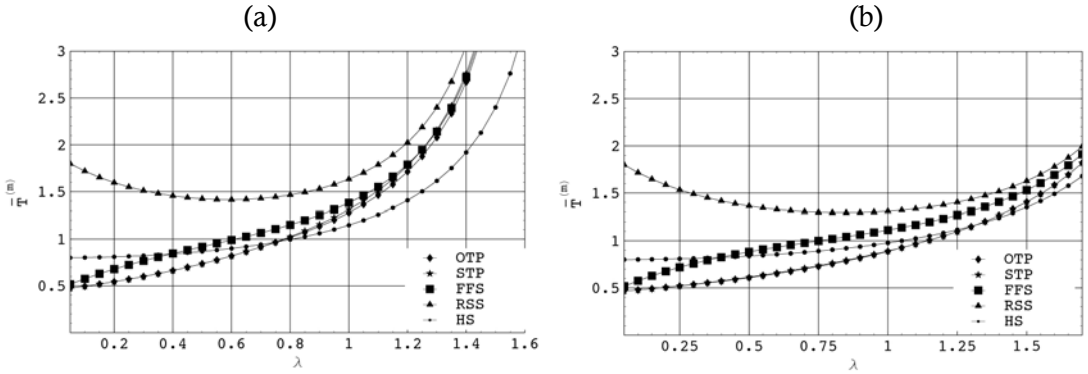

Figure 5. Mean sojourn time versus primary arrival and retrial rate (a) $\gamma = 2.5$ (b) $\gamma = 18.5$.

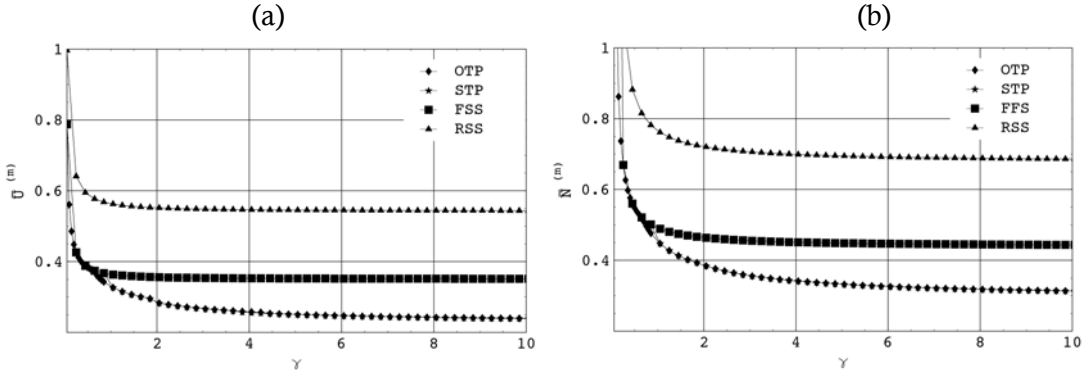

Figure 6. (a) System utilization and (b) mean number of customers in the system versus retrial rate.

The next figures represent the system utilization (Figure 6(a)), mean number of customers in the system (Figure 6(b)), mean waiting time (Figure 7(a)) and mean sojourn time of a customer (Figure 7(b)) for service intensities $\mu_1 = 2.2, \mu_2 = 0.3,$ primary arrival rate $\lambda = 0.5$ with varying retrial rate $0.05 \le \gamma \le 10.$

With respect to the given performance measures we observe the following.

1. For a small retrial rate the threshold control policies prescribe to use the slower server when $q_2^* = 1,$ that coincides with the heuristic policies (FFS, RSS). As the retrial rate increases the advantage of the threshold policies with respect to the mean number of customers in the system or mean sojourn time becomes more evident.
2. For large retrial rates the illustrated curves converge to their asymptotic values that resemble very close to the corresponding performance measures of the classical $M / M / 2$ queue.
3. All curves monotone decrease, except for the mean waiting time under threshold policies (OPT, STP). This can be explained by the fact that as γ increases the threshold levels also increase, which in turn leads to a significant increase of the mean waiting time.

6.2 Waiting and sojourn time analysis

In Figures 8-11 we have indicated the waiting time (the figures labeled by letter "a") and the sojourn time (the figures labeled by letter "b") distribution functions for different values of the primary customer arrival rate λ and retrial customer rate γ. In our examples we fix the service rates $\mu_1 = 2.2, \mu_2 = 0.3$. The following observation can be noticed from these figures:

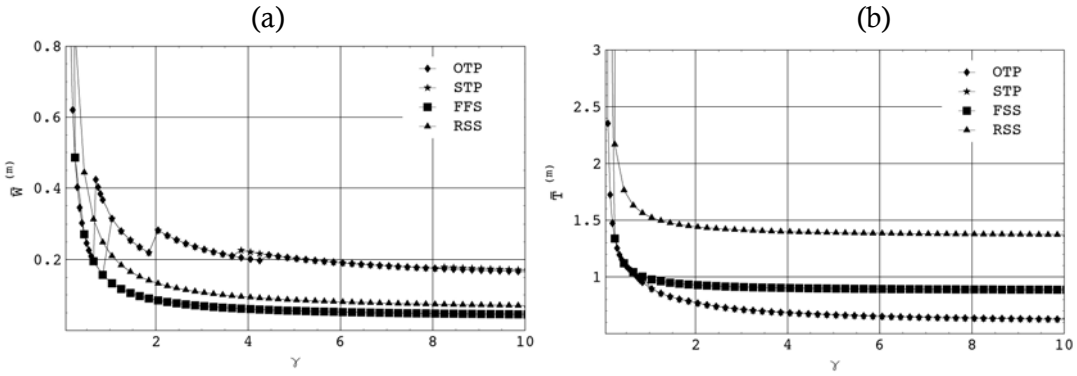

Figure 7. (a) Mean waiting time and (b) mean sojourn time of a customer versus retrial rate.

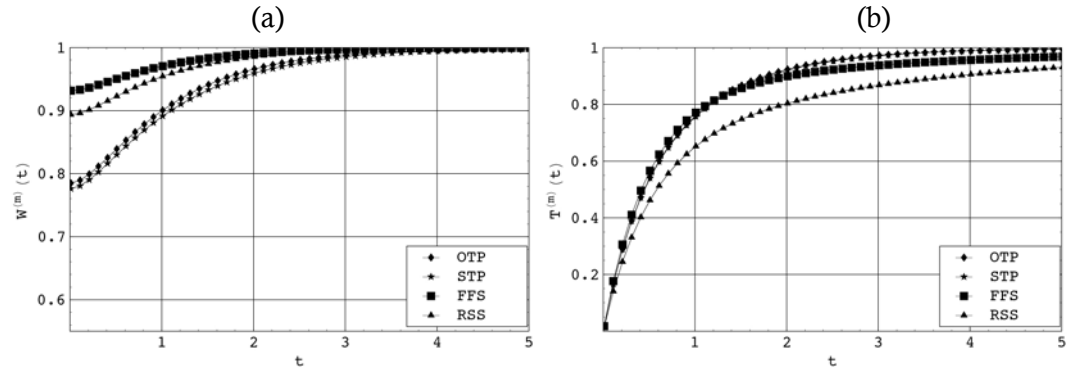

Figure 8. Distribution functions (a) $W(t)$ (b) $T(t)$ for $\lambda = 0.5, \mu_1 = 2.2, \mu_2 = 0.3, \gamma = 2.5$.

1. The curves of the waiting time distributions $W(t)$ for the system under threshold control policies (OTP, STP) lie below the other curves (FFS, RSS). This means that the waiting time of a customer in the orbit is larger for the threshold systems. This does not contradict the optimality of the threshold policy since it minimizes the sojourn time. Regarding the sojourn time distributions $T(t)$ one can notice that for some small values of argument t the curves for the OTP can lie below other graphs but starting from some point of time t they are above the others. Nevertheless the mean sojourn time for the optimal policy turns out to be the smallest. The curves of the waiting time distribution for the system under FFS control policy lie above the other curves, which illustrates that this policy minimizes the waiting time of a customer in the orbit. At the same time the largest sojourn time belongs to the system under RSS

policy. It can be explained by the fact that this policy assigns a customer to the faster or slower server with equal probability, which significantly increases the sojourn time.

2. In Figures 8, 9 and 10, 11 the primary arrival rate λ and retrial rate γ are varied, respectively. As λ or γ increases, the load factor ρ also increases, which leads to distributions with heavier tails. While in Figure 9 and 11 the curves for the threshold systems (OTP, STP) look very similar, in Figure 10 and 12 a large difference can be noticed. Thus we can assume that if the load factor is sufficiently small, i.e. the system has a so-called "light traffic", then the scheduling threshold policy may be a good approximation for the optimal one.

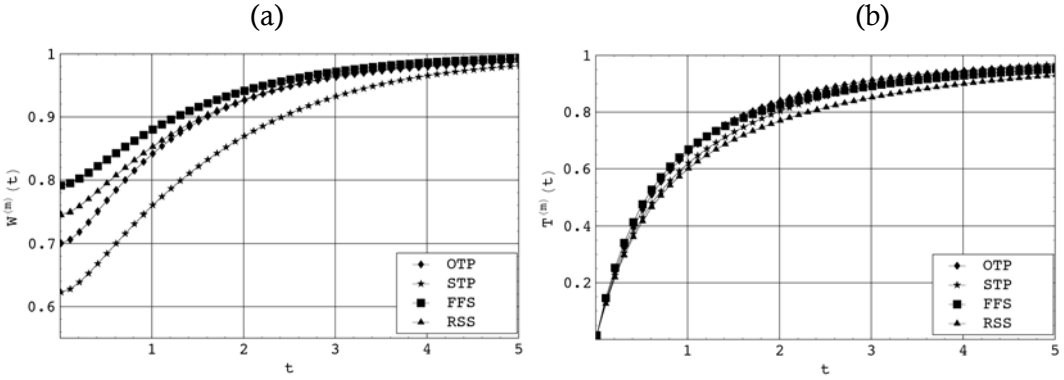

Figure 9. Distribution functions (a) $W(t)$ (b) $T(t)$ or $\lambda = 0.9, \mu_1 = 2.2, \mu_2 = 0.3, \gamma = 2.5$.

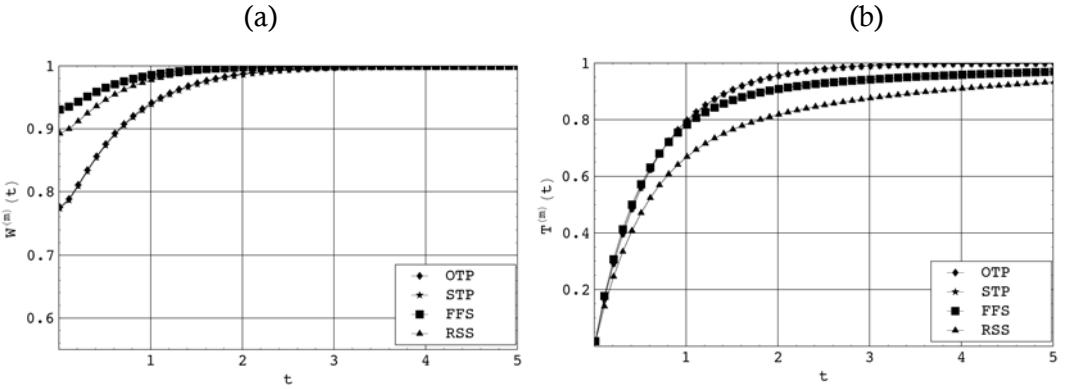

Figure 10. Distribution functions (a) $W(t)$ (b) $T(t)$ for $\lambda = 0.5, \mu_1 = 2.2, \mu_2 = 0.3, \gamma = 8.5$.

The next Figures 12 ($\lambda = 0.5$ and $\lambda = 0.9$ in the figures labeled by letter "a" and "b", respectively) and 13 ($\gamma = 2.5$ and $\gamma = 8.5$ in the figures labeled by letter "a" and "b", respectively) illustrate the discrete distribution function $\Psi(n)$ for the number of retrials made by an orbiting customer until it reaches the service area. The following conclusions cam be drawn from the graphs.

1. At the point $t = 0$ the jump of the functions corresponds to the case when no retrials will be made by a customer and equals the probability that the customer will be served directly, i.e. $\Psi(0) = W(0)$.
2. As λ and γ increase, the distributions reveal heavier tails. Under the FFS and RSS control policies the customer makes less retrials than under threshold policies (OTP, STP) because the first two disciplines imply a shorter waiting time.
3. The number of retrials strongly depends on the retrial rate. In case when γ is small, see Figure 12(a, b), the number of retrials will, with a large probability, not be very high, in this example $\Psi(6) > 0.99$ and $\Psi(7) > 0.99$. Otherwise, when γ is large, see Figure 13(a, b), then the number of retrial significantly increases, e.g. $\Psi(10) > 0.96$ and $\Psi(10) > 0.94$.

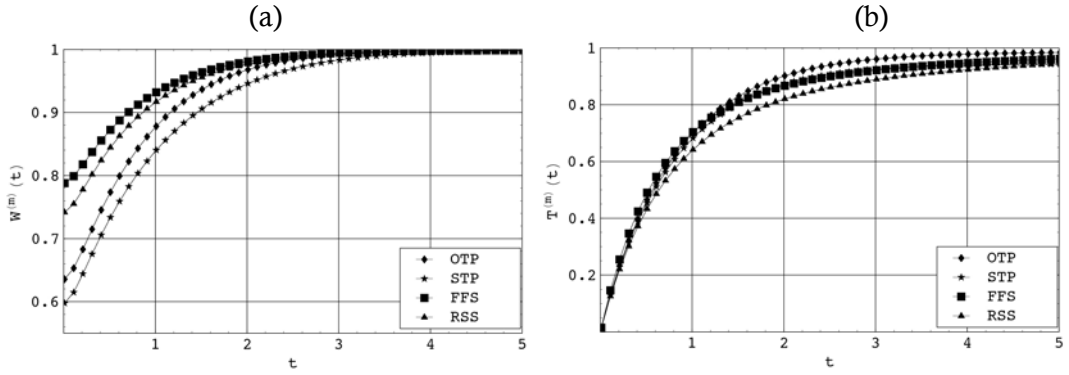

Figure 11. Distribution functions (a) $W(t)$ (b) $T(t)$ for $\lambda = 0.9, \mu_1 = 2.2, \mu_2 = 0.3, \gamma = 8.5$.

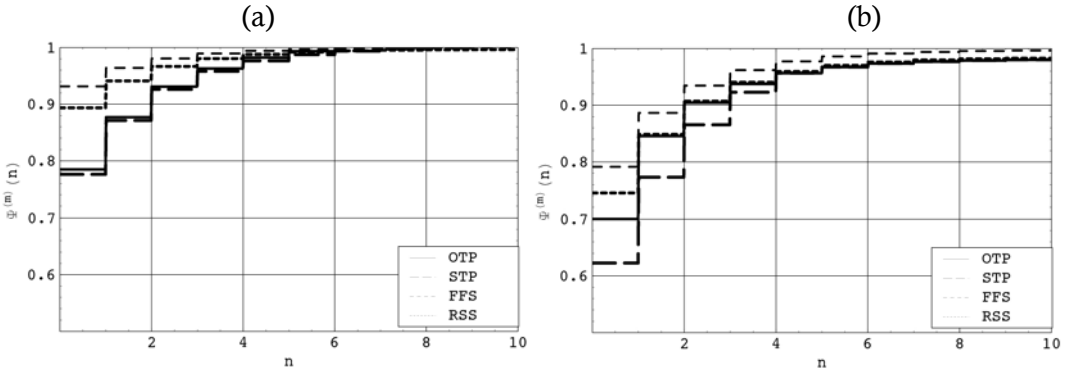

Figure 12. The distribution function $\Psi(n)$ (a) $\lambda = 0.5$ (b) $\lambda = 0.9, \mu_1 = 2.2, \mu_2 = 0.3, \gamma = 2.5$.

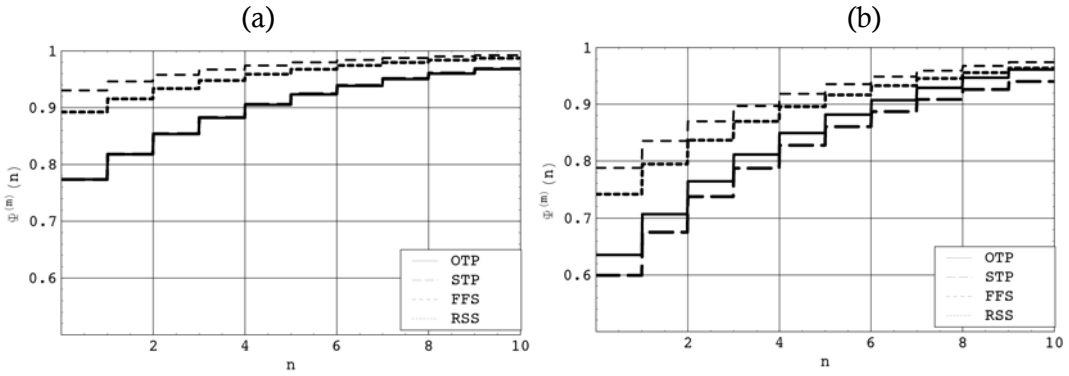

Figure 13. The distribution function $\Psi(n)$ (a) $\lambda = 0.5$ (b) $\lambda = 0.9, \mu_1 = 2.2, \mu_2 = 0.3, \gamma = 8.5$.

Acknowledgements

The authors thank an anonymous referee for reading this paper and for his/her detailed comments that helped to improve the results' presentation. This work was supported by the Austro-Hungarian Cooperation Grant No 72öu6 between Johannes Kepler University Linz and Debrecen University.

References

1. Abate, J. and Whitt, W. (1992). Numerical inversion of probability generating functions. *Operation Research Letters*, 12, 245-251.
2. Abate, J. and Whitt, W. (1995). Numerical inversion of Laplace transforms of probability distributions. *ORSA Journal on Computing*, 7, 36-43.
3. Aissani, A. (2000). An $M^X / G / 1$ retrial queue with exhaustive vacations. Journal of Statistics and Managements, 3, 269-286.
4. Artalejo, J. R. (1996). Stationary analysis of the characteristics of the $M / M / 2$ queue with constant repeated attempts. *Opsearch*, 33(2), 83-95.
5. Artalejo, J. R., Gomez-Corral, A. and Neuts, M.F. (2001). Analysis of multiserver queues with constant retrial rate. *European Journal of Operations Research*, 135, 569-581.
6. Artalejo, J. R.,Chakravarthy, S. R. and Lopez-Herrero, M. J. (2007). The busy period and the waiting time analysis of $MAP / M / c$ queue with finite retrial group. *Stochastic Analysis and Applications*, 25, 445-469.
7. Artalejo, J. R. and Gomez-Corral, A. (2008). Retrial queueing systems. Springer-Verlag, Berlin.
8. Breuer, L., Klimenok, V., Birukov, A., Dudin, A. and Krieger, U. R. (2005). Modeling the access to a wireless network at hot spots. *European Transactions on Telecommunication*, 16, 309-316.
9. Chakka, R. and Do, T. V. (2007). The $MM\sum_{k=1}^{K} CPP_k / GE / c / L$ G-queue with heterogeneous servers: Steady state solution and application to performance evaluation. *Performance Evaluation*, 64, 191-209.
10. Chakravarthy, S. R., Krishnamoorty, A. and Joshua, V. C. (2006). Analysis of a multi-server retrial queue with search of customers from the orbit. *Performance Evaluation*, 63, 776-798.

11. Choi, B. D., Shin, Y. W. and Ahn, W. C. (1992). Retrial queues with collision arising from unslotted CSMA/CD protocol. *Queueing Systems*, 11, 335-356.

12. Efrosinin, D. and Rykov, V. (2004). Optimal control of queueing systems with heterogeneous servers. *Queueing Systems*, 46, 389-407.

13. Efrosinin, D. and Breuer, B. (2006). Threshold policies for controlled retrial queues with heterogeneous servers. *Annals of Operation Research*, 141, 139-162.

14. Efrosinin, D. and Sztrik, J. (2007). Stochastic analysis of controlled retrial queues with heterogeneous servers and constant retrial rate Berichte der mathematischen Institute. *Johannes Kepler University*, 563, 1-21.

15. Efrosinin, D. and Rykov, V. (2008). On performance characteristics for queueing systems with heterogeneous servers. *Automation and Remote Control*, 1, 64-82.

16. Falin, G. I. and Templeton, J. G. C. (1997). Retrial queues. Chapman and Hall, London.

17. Farahmand, K. (1990). Single line queue with repeated demands. *Queueing Systems*, 6, 223-228.

18. Fayolle, G. (1986). A simple telephone exchange with delayed feedbacks. In *Teletraffic Analysis and Computer Performance Evaluation* (Edited by O. J. Boxma, J. W. Cohen and H. C. Tijms), 245-253. Elseiver Science.

19. Gomez-Corral, A. and Ramalhoto, M. F. (2000). On the waiting time distribution and the busy period of a retrial queue with constant retrial rate. *Stochastic Modeling and Applications*, 3(2), 37-47.

20. Kleinrock, L. (1975). *Queueing Systems*, Vol. I, Wiley, New York.

21. Larsen, R. L. and Agrawala, K. (1983). Control of heterogeneous two-server exponential queueing system. *IEEE Transactions on Software Engineering*, 4, 522-526.

22. Lehtonen, T. (1983). Stochastic comparisons for many server queues with non-homogeneous exponential servers. *Opsearch*, 20(1), 1-15.

23. Li, W. and Zhao, Y. Q. (2005). A retrial queue with a constant retrial rate, server downs and impatient customers. *Stochastic Models*, 21, 531-550.

24. Neuts, M. F.. (1981). Matrix-geometric solutions in stochastic models. The John Hopkins University Press, Baltimore.

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

26. Ohmura, H. and Takahashi, Y. (1985). An analysis of repeated call model with a finite number of sources. *Electronic and Communications in Japan*, 68(6), 112-121.

27. Pourbabai, B. (1987). Markovian queueing systems with retrials and heterogeneous servers. *Computers and Mathematics with Applications*, 13, 917-923.

28. Rykov, V. (2001). Monotone Control of Queueing Systems with Heterogeneous. Servers, *Queueing systems*, 37, 391-403.

29. Sztrik, J. and Roszik, J. (2007). Performance analysis of finite-source retrial queueing systems with nonreliable heterogeneous servers. *Journal of Mathematical Sciences*, 146(4), 6033-6038.

30. Trancoso, P. (2005). One size does not fit all: A case for heterogeneous multiprocessor systems. *Proceedings of the IADIS International Conference Applied Computing*, Algarve, Portugal.

31. Tran-Gia, P. and Mandjes, M. (1997). Modeling of customer retrial phenomenon in cellular mobile networks. *IEEE Journal of Selected Areas in Communication*s, 15, 1406-1414.

Authors' Biographies:

Dmitry Efrosinin is University-Assistant for Research and Teaching in the Institute for Stochastics, Johannes Kepler University of Linz, Austria. He is a Research consultant at the Department of Probability theory and Math. Statistics, Peoples Friendship University, Russia. His main scientific interests are in queueing and reliability systems, optimization problems and structural properties of optimal control policies.

Janos Sztrik is Professor in the Department of Informatics Systems and Networks, University of Debrecen, Hungary. He is a Member of J. Bolyai Mathematical Society, Budapest and London Mathematical Society. His research interests are in mathematical statistics, queueing theory and computer performance evaluation.