mathematics

MDPI

Article

Algorithmic Analysis of Finite-Source Multi-Server Heterogeneous Queueing Systems

Dmitry Efrosinin [1,2,3], Natalia Stepanova [3], Janos Sztrik [4,*]

1 Insitute for Stochastics, Johannes Kepler University Linz, Altenbergerstrasse 69, 4040 Linz, Austria; dmitry.efrosinin@jku.at
2 Department of Information Technologies, Faculty of Mathematics and Natural Sciences, Peoples' Friendship University of Russia (RUDN University), Miklukho-Maklaya 6, 117198 Moscow, Russia
3 V.A. Trapeznikov Institute of Control Sciences of RAS, Profsoyuznaya 65, 117997 Moscow, Russia; natalia0410@rambler.ru
4 Department of Informatics and Networks, Faculty of Informatics, University of Debrecen, Egyetem tér 1, 4032 Debrecen, Hungary
* Correspondence: sztrik.janos@inf.unideb.hu

Abstract: The paper deals with a finite-source queueing system serving one class of customers and consisting of heterogeneous servers with unequal service intensities and of one common queue. The main model has a non-preemptive service when the customer can not change the server during its service time. The optimal allocation problem is formulated as a Markov-decision one. We show numerically that the optimal policy which minimizes the long-run average number of customers in the system has a threshold structure. We derive the matrix expressions for performance measures of the system and compare the main model with alternative simplified queuing systems which are analysed for the arbitrary number of servers. We observe that the preemptive heterogeneous model operating under a threshold policy is a good approximation for the main model by calculating the mean number of customers in the system. Moreover,using the preemptive and non-preemptive queueing models with the faster server first policy the lower and upper bounds are calculated for this mean value.

Keywords: finite-source queueing system; preemptive and non-preemptive service; Markov-decision process; policy-iteration algorithm; performance analysis

Citation: Efrosinin, D.; Stepanova, N.; Sztrik, J. Algorithmic Analysis of Finite-Source Multi-Server Heterogeneous Queueing Systems. *Mathematics* **2021**, *9*, 2624. https://doi.org/10.3390/math9202624

Academic Editor: Mark Kelbert

Received: 29 September 2021
Accepted: 15 October 2021
Published: 18 October 2021

1. Introduction

The finite-source or finite-population queueing systems comparing to the ordinary markovian queues have no longer a Poisson arrival stream as in systems with an infinite source of customers, but rather have a finite source capacity N of possible customers. In such systems a customer can be inside the system, consisting in our case of one common queue with capacity N and K heterogeneous servers or outside the system in so-called arriving state. It is assumed that each customer outside arrives to the system in exponentially distributed time. After receiving the service a customer returns to the arriving area. Much attention by the study of finite-source queueing systems has been paid in terms of the machine repairman problem, see e.g., [1,2]. The customers outside the queueing system can be interpreted as unreliable machines with independent exponentially distributed life times. The queueing system represents then the repair facility where the failed machines must be recovered. Such systems are also used in various dispatching problems, they are appropriate queueing models for telephone registration systems, call centers, Ethernet systems, local-area networks, mobile communications, magnetic disk memory systems and so on.

The primary objective of this paper is to analyze optimal control and to evaluate performance metrics under fixed control policy for a non-preemptive finite-source queueing system with one class of customers and heterogeneous servers. In such a system, a customer

that receives service on a slower server cannot change it if a faster server becomes available in the course of service. Unfortunately, performance analysis of this system in analytical form is limited firstly by the need to have a known allocation mechanism between the servers or control policy and secondly by the dimensionality of the corresponding Markov process, which is affected by the number of servers. To calculate the optimal allocation policy with the aim to minimize the long-run average number of customers in the system we formulate the Markov decision process (MDP) and apply the policy-iteration algorithm. This algorithm can be used not only for the optimal allocation policy calculation but also to obtain the mean number of customers in the system operating under that policy. Numerical experiments confirm our expectations that the optimal policy is of threshold type as in the models with an infinite source capacity [3]. According to this policy the fastest server must be activated whenever there is a customer in the system while the slower servers must be used only if the number of customers in the queue reaches some prespecified threshold level. The model of the non-preemptive queue operating under the optimal threshold policy (OTP) will referred to in the paper as the OTP-model.

The task of calculating other system performance characteristics for a given control policy remains unresolved. Furthermore, it should be taken into account that despite of advantages the policy-iteration algorithm has a limitation on the dimensionality of the random process for an arbitrary number of servers. In case of a threshold control policy for a particular states' ordering the corresponding Markov chain is a quasi-birth-and-death (QBD) process with a three-diagonal block infinitesimal matrix, where the blocks depend on the values of thresholds as it was shown in [4] for the infinite population system. In this case, matrix-analytic solution methods can be applied, but for a limited number of servers. This led us to discuss here in addition some simplified variants of the main model. The non-preemptive queueing system operating under a Fastest Server First (FSF) policy which prescribes for service the usage of the fastest idle server in each state and the preemptive queueing system (PS), where the service in a slower server can be interrupted if during the service time the faster server becomes idle. This system will operates according to a threshold policy, when the slower servers are activated or deactivated if the number of customers in the queue respectively exceeds or falls below a certain threshold level. Although these systems are simpler than the main model and have a low dimensions of the state-space, there are very few publications on such specific systems, especially those with analytical results.

Description of standard finite-source models with classical results, motivation examples and literature overview can be found in [5]. In [6] the author obtained the product form solution for the stationary state distribution for the finite population queueing model with a queue-dependent servers. A non-preemptive finite-population queueing system with heterogeneous classes of customers and a single server was studied in [7]. The problem of the throughput maximizing in a finite-source system with parallel queues was analyzed in [8], where some structural properties of the optimal control policy was proved. Heterogeneous multi-server finite-source queues with a FSF-policy and retrial phenomenon have been studied in [9], where numerical results were carried out by the help of the MOSEL tool. The model with machines having non-identical exponential service times was analysed in [10], where the repair policies which minimize the mean processing cost were considered. As for finite-source systems operating under a threshold policy, we know only research papers dedicated to single server systems with a unit threshold N-policy, see e.g., [11]. Here the threshold policy states that the server in a repair facility must be switched on only when the number of failed machines reaches to predefined threshold level N. In [12], the authors generalized the model to the case of two heterogeneous unreliable servers operating under policies N_1 and N_2. Although there is a considerable amount of results on finite-source queues, the controlled systems with heterogeneous servers operating without preemption have not been investigated before. Therefore, the models presented in this paper and the corresponding results of the system analysis differ from those previously discussed.

The main contributions of paper are as follows. The structural or threshold properties of an optimal control policy are numerically established. We derive for the main model the matrix expressions used further by calculating different performance measures such as the mean number of waiting customers, the mean number of busy servers, the mean length of a busy period. The matrix-analytic solution for the stationary state distribution and performance measures are obtained for the FSF-model. Here we used the recursive definition of some blocks in the infinitesimal matrix. For the PS-models we obtain analytical product-form result for an arbitrary number of servers. Moreover, it is shown that performance characteristics of these systems in certain operation modes are the same or very close to those of the main system functioning under the optimal policy. Thus, these simplified models can be used under certain conditions to calculate upper and lower bounds for some performance characteristics and also as approximating models. We develop also the first step analysis to study the mean number of customers served in the system or by the kth server in a busy period and the probability of the maximum queue length observed during this period.

The rest of the paper is organized as follows. In Section 2, we describe the Markov-decision process of the main model and show that the system has a threshold-based optimal allocation policy. In this section we develop also the computational analysis for the mean values of performance measures including those characterizing the behaviour of the system in a busy period. The FSF-Model is presented and analysed in Section 3. Section 4 is devoted to the PS-model. Comparison analysis of the proposed models and illustrative examples are summarized in Section 5.

The following notations will be used throughout this paper: $\mathbf{e}(n)$, $\mathbf{e}_j(n)$ and I_n stand respectively for the unit vector of dimension n, for the basis vector of dimension n in $\mathbb{R}^n$ with $0 \leq j \leq n-1$ or $1 \leq j \leq n$ depending on the context, and for the identity matrix of dimension n. If it is not necessary to specify a vector dimension, we will omit the corresponding argument. For example, $\mathbf{e}$ denotes a column unit vector of an appropriate dimension. The notation $'$ is used for the transpose. The notation $\otimes$ stands for the Kronecker product of two matrices, $1_{\{A\}}$ – for the indicator function, where $1_{\{A\}} = 1$ if the condition A holds, and 0 otherwise. The notation $|A|$ is used for the magnitude of a finite set A.

2. OTP-Model

Here we discuss the main model of the non-preemptive finite-source controlled queueing system of the type $M/M/K/N//N$ illustrated in Figure 1. The system has K heterogeneous servers with different rates $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_K > 0$ and N customers in a source. It operates under the optimal allocation policy which minimizes the mean number of customers in the system. It will be shown that this policy is defined through a sequence of threshold levels $1 = q_1 \leq q_2 \leq \cdots \leq q_K < \infty$ for the queue lengths which prescribe the activation of slower servers. The analysed system can be treated as a model for the machine-repairman problem, where N unreliable machines in a working area with exponential distributed life times and equal rates $\lambda > 0$ must be repaired by K heterogeneous repair stations. The machines fail independently of each other. The stream of failed machines can be treated as an arrival stream of customers to the queueing system. Hereafter, we will refer to the customer as a failed machine which enters the repair system and gets there a repair service. After the repair the machine becomes as good as a new one and it returns to the working area. The aim is to dynamically allocate the customers to the servers in order to minimize the long-run average number of customers in the system and to calculate the corresponding mean performance measures.

2.1. MDP Formulation

We formulate the optimal allocation problem in this machine-repairman system as a Markov Decision Process (MDP) in the following way. The behaviour of the system is described by a multi-dimensional continuous-time Markov-chain

$$\{X(t)\}_{t\geq 0} = \{Q(t), D_1(t), \ldots, D_K(t)\}_{t\geq 0}, \tag{1}$$

where $Q(t)$ stands for the number of customers waiting in the queue at time t and $D_j(t)$ specifies the state of the jth server at time t, where

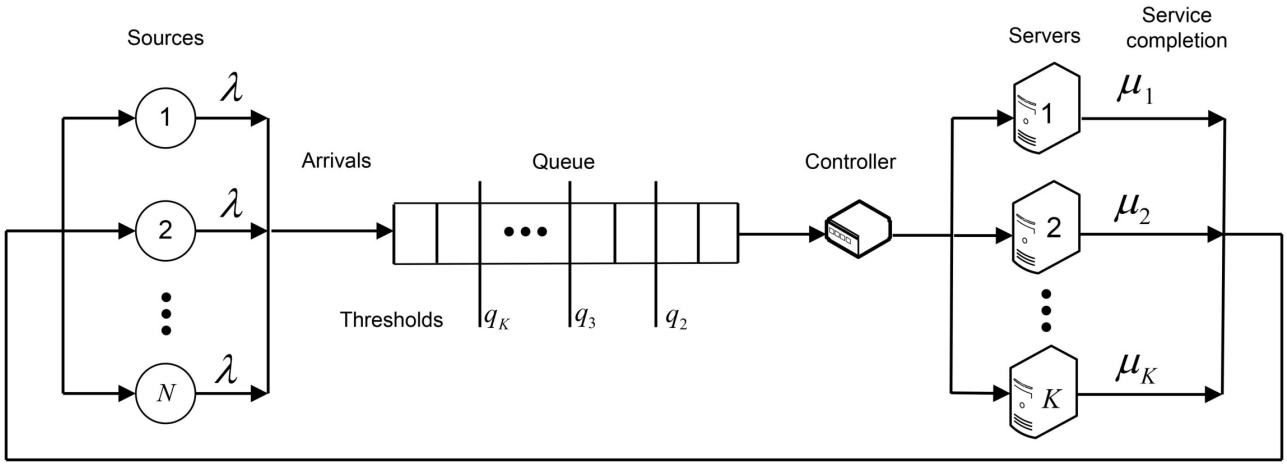


Figure 1. The schema of the finite-source queueing system.

$$D_j(t) = \begin{cases} 0 & \text{if the server } j \text{ is idle} \\ 1 & \text{if the server } j \text{ is busy.} \end{cases}$$

State space: The set E_X consists of $K+1$ dimensional row vectors,

$$\begin{aligned} E_X = &\{x = (q(x), d_1(x), \ldots, d_K(x)) : \\ &q(x) \in \{0, 1, 2, \ldots, N - \sum_{j=1}^{K} d_j(x)\},\ d_j(x) \in \{0, 1\},\ j = 1, \ldots, K\}, \end{aligned}$$

where $q(x)$ denotes the number of customers in the queue and $d_j(x)$ – the status of the jth server in state x. The total number of states in the set E_X is equal to $|E_X| = \sum_{j=0}^{K} \binom{K}{j}(N-j+1)$. *Decision epochs:* The arrival and service completion epochs in the system with waiting customers. *Action space:* $A = \{0, 1, \ldots, K\}$. To identify the group of idle and busy servers, the following sets are defined,

$$J_0(x) = \{j : d_j(x) = 0\},\ J_1(x) = \{j : d_j(x) = 1\}.$$

With this notations the set of admissible control actions $A(x) \subseteq A$ in state $x \in E_X$ can be defined as $A(x) = J_0(x) \cup \{0\}$. The action $a \in J_0(x)$ means that in state x a customer must be allocated to an idle server, while $a = 0$ means that the customer must be routed to the queue. At an arrival epoch, which occurs only if the number of customers in the system is less than N, the arrived customer joins the queue and simultaneously another one from the head of the queue must be routed to some idle server or returned back to the queue. At a service completion epoch the same happens, i.e. the customer from the head of the queue is routed either to one of idle servers or to the queue again. By service completion in a system without waiting customers no actions have to be performed.

Immediate cost: The function $l(x)$ specifies the number of customers in a state $x \in E_X$, i.e.,

$$l(x) = q(x) + \sum_{j=1}^{K} d_j(x),$$

which is in fact independent of a control action a.

Transition rates: The policy-dependent infinitesimal matrix $\Lambda^f = [\lambda_{xy}(a)]_{x,y \in E_X}$ of the Markov-chain (1) includes the rates to go from state x to state y given the control action is a defined as

$$\lambda_{xy}(a) = \begin{cases} (N - l(x))\lambda & y = S_a x,\ 0 \le l(x) \le N,\ a \in A(x), \\ \mu_j & y = S_j^{-1} x,\ j \in J_1(x),\ q(x) = 0, \\ \mu_j & y = S_0^{-1} S_j^{-1} S_a x,\ j \in J_1(x),\ q(x) > 0,\ a \in A(S_0^{-1} S_j^{-1} x), \\ -((N - l(x)) & + \sum_{j \in J_1(x)} \mu_j) \qquad y = x, \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

with $\lambda_x(a) = -\lambda_{xx}(a) = -\sum_{y \neq x} \lambda_{xy}(a)$, where S_a and S_j^{-1} stand for the shift operators applied to the vector state x in the following way,

$$S_a x = x + \mathbf{e}_a(K+1),\ a \in J_0(x) \text{ and } S_j^{-1} x = x - \mathbf{e}_j(K+1),\ j \in J_1(x).$$

Due to the finiteness of the state space E_X and boundedness of the immediate cost function $l(x) \le N$, a stationary average-cost optimal policy $f : E \to A$ exists with a finite constant mean value for the number of customers in the system

$$g^f = \limsup_{t \to \infty} \frac{1}{t} \mathbb{E}^f \Big[\int_0^t \Big(Q(t) + \sum_{j=1}^{K} D_j(t) \Big) dt \Big| X(0) = x \Big] = \sum_{x \in E_X} l(x) \pi_x^f < \infty,$$

which is independent of the initial state x. In this case the policy-iteration algorithm introduced in Algorithm 1 converges.

This algorithm consists of two main parts: Policy evaluation and Policy improvement. In the first part, a system of linear equations with immediate costs $l(x)$

$$v^f(x) = -\frac{1}{\lambda_{xx}(a)} \Big(l(x) + \sum_{y \neq x} \lambda_{xy}(a) v^f(y) - g^f \Big) \tag{3}$$

is solved for the unknown real-valued dynamic-programming value function $v^f : E_X \to \mathbb{R}$ and mean value g^f given a control policy is f. The second part of the algorithm is responsible for improving the previous policy, which for a given system consists in determining, for each system state, a control action a that minimizes the value function $v(S_a x)$. The improved control action in state x is defined then as $f^*(x) = \operatorname{argmin}_{a \in A(x)} v(S_a x)$ for $x \in E_X \setminus \{x : l(x) = N\}$. Thus, the algorithm constructs a sequence of improved control policies until it finds one that minimizes the gain g^f.

In Algorithm 1 we perform a conversion of the $K+1$-dimensional state space E_X of the Markov chain (1) to one-dimensional equivalent state space using the function $\Delta : E_X \to \mathbb{N}_0$, where

$$\Delta(x) = q(x) 2^K + \sum_{i=1}^{K} d_i(x) 2^{i-1}. \tag{4}$$

In one-dimensional state space the transitions due to arrivals and service completions can be defined then as

$$\Delta(x \pm \mathbf{e}_0(K+1)) = (q(x) \pm 1)2^K + \sum_{i=1}^{K} d_i(x)2^{i-1} = \Delta(x) \pm 2^K,$$

$$\Delta(x \pm \mathbf{e}_j(K+1)) = q(x)2^K + \sum_{i=1}^{K} d_i(x)2^{i-1} \pm 2^{j-1} = \Delta(x) \pm 2^{j-1},\ 1 \le j \le K.$$

For more details about derivation of the optimality equation for heterogeneous queueing systems the interested reader is referred to relevant publications, e.g., [3].

Algorithm 1 Policy-iteration algorithm

procedure PIA($K, N, \lambda, \mu_j, j = 1, 2, \ldots, K$)

$f^{(0)}(x) = \text{argmax}_{j \in J_0(x)} \left\{ \mu_j \right\}$ ▷ Initial policy

$n \leftarrow 0$

$g^{f^{(n)}} = N\lambda v^{f^{(n)}}(\mathbf{e}_1(K+1))$ ▷ Policy evaluation

for $x = (0, 1, 0, \ldots, 0)$ **to** $(N-K, 1, 1, \ldots, 1)$ **do**

$$v^{f^{(n)}}(x) = \frac{1}{(N-l(x))\lambda + \sum_{j \in J_1(x)} \mu_j} \Big[l(x) - g^{f^{(n)}} + (N-l(x))\lambda v^{f^{(n)}}(S_{f^{(n)}(x)}x) + \sum_{j \in J_1(x)} \mu_j v^{f^{(n)}}(S_j^{-1}x) 1_{\{q(x)=0\}} + \sum_{j \in J_1(x)} \mu_j v^{f^{(n)}}(S_0^{-1}S_j^{-1}S_{f^{(n)}(S_0^{-1}S_j^{-1}x)}x) 1_{\{q(x)>0\}} \Big]$$

end for

▷ Policy improvement

$$f^{(n+1)}(x) = \text{argmin}_{a \in A(x)}\, v^{f^{(n)}}(S_a x)$$

if $f^{(n+1)}(x) = f^{(n)}(x)$, $x \in E^f$ **then return** $f^{(n+1)}(x), v^{f^{(n)}}(x), g^{f^{(n)}}$

else $n \leftarrow n+1$, **go to step 4**

end if

▷ Threshold evaluation

$$q_k : f^{(n+1)}(q, 1, \ldots, 1, 0, d_{k+1}, \ldots, d_K) = \begin{cases} 0 & q \le q_k - 2 \\ k & q > q_k - 2 \end{cases},\ k = 2, \ldots, K$$

end procedure

Numerical analysis confirms our expectation that the optimal control policy in heterogeneous systems for a finite number of customers also belongs to a class of threshold policies, as in infinite population case. Theoretical justification of this statement is still difficult. For this purpose it is necessary to prove that the dynamic-programming operator B defined for our queueing model as

$$\begin{aligned} v(x) &= \frac{1}{(N-l(x))\lambda + \sum\limits_{j \in J_1(x)} \mu_j} \Big[l(x) + (N-l(x))\lambda T_0 v(x) + \sum_{j \in J_1(x)} \mu_j T_j v(x) - g \Big] \\ &= Bv(x), \end{aligned} \tag{5}$$

where T_0 and T_j are the events operators in case of a new arrival and a service completion at server $j \in J_1(x)$,

$$T_0 v(x) = \min_{a \in A(x)} v(S_a x),$$
$$T_j v(x) = v(S_j^{-1} x),\ q(x) = 0,$$
$$T_j v(x) = T_0 v(S_0^{-1} S_j^{-1} x),\ q(x) > 0,$$

preserves the monotonicity properties of the increments of the value function v:

$$v(S_0 x) - v(S_2 x) - v(S_0^2 x) + v(S_0 S_2 x) \leq 0,\ x \in E_X,\ d_1(x) = 1,\ d_2(x) = 0, \quad (6)$$
$$v(S_0 x) - v(x) - v(S_0 S_2 x) + v(S_2 x) \leq 0,\ x \in E_X,\ d_1(x) = 1,\ d_2(x) = 0. \quad (7)$$

In proving the inequality (7) we encounter difficulty. This is due primarily to the form of the operator B in (5). There is a term describing arriving customers whose coefficient $(N - l(x))\lambda$ depends on the system state x. Bringing the terms in inequality (7) to a common denominator by introducing fictitious transitions, we get terms which cannot be proved to be negative. We hope that we will be able to overcome these difficulties in our next paper, but to date we're basing our statement about a threshold structure of the optimal control policy f exclusively on the performed numerical experiments. The following example makes the case vividly.

Example 1. *Consider the system with $K = 5, N = 60$ and $\lambda = 0.3$. The service rates take the following values: $(\mu_1, \mu_2, \mu_3, \mu_4, \mu_5) = (20, 8, 4, 2, 1)$. The Table 1 of optimal control actions $f(x)$ for selected system states x is of the form:*

Table 1. The optimal control actions.

System State x	**Queue Length $q(x)$**													
$d = (d_1, d_2, d_3, d_4, d_5)$	0	1	2	3	4	5	6	7	8	9	10	11	12	...
(0, *, *, *, *)	1	1	1	1	1	1	1	1	1	1	1	1	1	1
(1, 0, *, *, *)	2	2	2	2	2	2	2	2	2	2	2	2	2	2
(1, 1, 0, *, *)	0	3	3	3	3	3	3	3	3	3	3	3	3	3
(1, 1, 1, 0, *)	0	0	0	4	4	4	4	4	4	4	4	4	4	4
(1, 1, 1, 1, 0)	0	0	0	0	0	0	0	0	5	5	5	5	5	5
(1, 1, 1, 1, 1)	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Threshold levels q_k, $2 \leq k \leq K$, are evaluated by comparing the optimal actions $f(x) = 0$ and $f(S_0 x) = k$ for $x = (q(x), 1, \ldots, 1, 0, d_{k+1}(x), \ldots, d_K(x))$, $0 \leq q(x) \leq N - \sum_{j=1}^{K} d_j(x), d_j(x) \in \{0, 1\}$. In this example the optimal policy f is defined here through a sequence of threshold levels $(q_2, q_3, q_4, q_5) = (1, 2, 4, 9)$ and $g^f = 4.91549$.

2.2. Evaluation of System Performance Measures

We are concerned in calculation of the system performance measures for a given policy f. The state probabilities and performance characteristics defined here will refer to some particular fixed control policy f, so we will use in notations the corresponding upper index. The states x of the set E_X with $q(x) = 0$ are ordered according to the number of busy servers $|J_1(x)|$ while the states for $q(x) > 0$ are ordered with respect to the queue length, so that the infinitesimal matrix Λ^f has a block three-diagonal structure for the fixed policy f. First we define the performance characteristics:

- The probability that the kth server $1 \leq k \leq K$ is busy, $\bar{U}_k^f = \sum_{x \in E_X} d_k(x) \pi_x^f$;

- The mean number of busy servers, $\bar{C}^f = \sum_{k=1}^{K} \bar{U}_k^f$;
- The mean number of customers in the queue, $\bar{Q}^f = \sum_{x \in E_X} q(x)\pi_x^f$.
- The mean number of customers in the system, $\bar{N}^f = \bar{C}^f + \bar{Q}^f$.

The following vectors of dimension $|E_X| - 1$ comprise the policy-dependent values $a^f(x)$ and policy-independent values $b(x)$,

$$\mathbf{a}^f = (a^f(x) : x \in E_X \setminus \{x_0\}),\ \mathbf{b} = (b(x) : x \in E_X \setminus \{x_0\}),\ x_0 = \mathbf{0}.$$

where the first elements of the vectors are respectively $a^f(\mathbf{e}_1(K+1))$ and $b(\mathbf{e}_1(K+1))$. Denote by $\bar{M}_1^f$ one of the performance characteristics $\bar{U}_k^f, \bar{C}^f, \bar{Q}^f$ and $\bar{N}^f$.

Proposition 1. *The performance measure $\bar{M}_1^f$ satisfies the relation*

$$\bar{M}_1^f = N\lambda \mathbf{e}_1'(|E_X| - 1)\mathbf{a}^f, \tag{8}$$

where the vector $\mathbf{a}^f$ is a solution of the system

$$(\tilde{\Lambda}^f + N\lambda \mathbf{e}'(|E_X| - 1) \otimes \mathbf{e}_1(|E_X| - 1))\mathbf{a}^f = -\mathbf{b}. \tag{9}$$

The matrix $\tilde{\Lambda}^f$ is obtained from Λ^f by removing the first column and the first row, and

$$\bar{M}_1^f = \begin{cases} \bar{U}_k^f & b(x) = d_k(x),\ x \in E_X, \\ \bar{C}^f & b(x) = \sum_{k=1}^{K} d_k(x),\ x \in E_X, \\ \bar{Q}^f & b(x) = q(x),\ x \in E_X, \\ \bar{N}^f & b(x) = l(x),\ x \in E_X. \end{cases} \tag{10}$$

Proof. We multiply the both sides of the equality (9) by the row-vector of the stationary state probabilities $\tilde{\boldsymbol{\pi}}^f = (\pi_x^f : x \in E \setminus \{x_0\})$,

$$\tilde{\boldsymbol{\pi}}^f(\tilde{\Lambda}^f - N\lambda \mathbf{e}'(|E_X| - 1) \otimes \mathbf{e}_1(|E_X| - 1))\mathbf{a}^f = -\tilde{\boldsymbol{\pi}}^f \mathbf{b},$$

where $\tilde{\boldsymbol{\pi}}^f \mathbf{b} = \sum_{x \in E_X \setminus \{x_0\}} b(x)\pi_x^f$ for the corresponding function $b(x)$ is obviously equal to the performance measure $\bar{M}_1^f$. The following sequence of relations

$$\begin{aligned} &\tilde{\boldsymbol{\pi}}(\tilde{\Lambda}^f - N\lambda \mathbf{e}'(|E_X| - 1) \otimes \mathbf{e}_1(|E_X| - 1))\mathbf{a}^f = \\ &\tilde{\boldsymbol{\pi}}^f \tilde{\Lambda}^f \mathbf{a}^f - N\lambda \tilde{\boldsymbol{\pi}}^f \mathbf{e}'(|E_X| - 1) \otimes \mathbf{e}_1(|E_X| - 1)\mathbf{a}^f = \\ &- \pi_{x_0}^f (N\lambda, 0, \ldots, 0)\mathbf{a}^f - N\lambda(1 - \pi_{x_0}^f, 0, \ldots, 0)\mathbf{a}^f = -N\lambda \mathbf{e}_1'(|E_X| - 1)\mathbf{a}^f = -\bar{M}_1^f. \end{aligned}$$

validates the statement. □

The following measures characterize the behaviour of the system in a busy period which we define as a duration starting when the arrived customer enters the empty system in state x_0 and finishes when the system visits x_0 again after a service completion.

- The mean length of a busy period, $\bar{L}^f = \frac{1}{N\lambda}\left(\frac{1}{\pi_{x_0}^f} - 1\right)$;
- The mean number of customers served in a busy period by the kth server, $\bar{N}_{L,k}^f$;
- The total mean number of customers served in a busy period, $\bar{N}_L^f = \sum_{k=1}^{K} \bar{N}_{L,k}^f = \frac{1}{\pi_{x_0}^f}$.

In the following proposition we describe a general way to calculate these characteristics for the fixed control policy f. Denote by $\bar{M}_2^f$ one of the performance characteristics $\bar{L}^f, \bar{N}_{L,k}^f$ and $\bar{N}_L^f$.

Proposition 2. *The performance measure $\bar{M}_2^f$ satisfies the relation*

$$\bar{M}_2^f = \mathbf{e}_1'(|E_X| - 1)\mathbf{a}^f, \tag{11}$$

where the vector $\mathbf{a}^f$ is a solution of the system

$$\tilde{\Lambda}^f \mathbf{a}^f = -\mathbf{b}. \tag{12}$$

The matrix $\tilde{\Lambda}^f$ is obtained from Λ^f by removing the first column and the first row, and

$$\bar{M}_2^f = \begin{cases} \bar{L}^f & b(x) = 1 + \sum_{k=1}^K d_k(x)\mu_k 1_{\{|J_1(x)|=1\}},\ x \in E_X, \\ \bar{N}_{L,k}^f & b(x) = d_k(x)\mu_k,\ x \in E_X, \\ \bar{N}_L^f & b(x) = \sum_{k=1}^K d_k(x)\mu_k,\ x \in E_X. \end{cases}$$

Proof. Denote by $\tilde{\varphi}_x^f(s) = \int_0^\infty \varphi_x^f(t)e^{-st}dt, Re[s] > 0$, the Laplace-Stiltjes transform (LST) of the probability density function (PDF) $\varphi_x^f(t)$ for the first passage time to state x_0 given that the initial state is $x \in E_X$, the control policy is f and by $\bar{L}_x^f = \int_0^\infty t\varphi_x^f(t)dt$ the corresponding first moment. According to the first step analysis we get for the LST the system

$$\begin{aligned} &\tilde{\varphi}_{x_0}^f(s) = 0, \\ &\tilde{\varphi}_x^f = \sum_{y \neq x} \frac{\lambda_{xy}(a)}{s + \lambda_x(a)} \tilde{\varphi}_y^f(s),\ x \in E_X \setminus \{x_0\}. \end{aligned} \tag{13}$$

We take into account that $\bar{L}^f(x) = -\frac{d}{ds}\tilde{\varphi}_x^f(s)\Big|_{s=0}$, we can obtain from (13) the system for the conditional moments

$$\begin{aligned} &\bar{L}^f(x_0) = 1, \\ &\bar{L}^f(x) = \frac{1}{\lambda_x(a)}\Big[1 + \sum_{y \neq x} \lambda_{xy}(a)\bar{L}^f(y)\Big],\ x \in E_X \setminus \{x_0\}. \end{aligned} \tag{14}$$

The system (14) for the transition rates (2) is of the form

$$\begin{aligned} &\Big((N - l(x))\lambda + \sum_{j \in J_1(x)} \mu_j\Big)\bar{L}^f(x) = 1 + \sum_{j \in J_1(x), |J_1(x)|=1} \mu_j 1_{\{q(x)=0\}} + \\ &(N - l(x))\lambda \bar{L}^f(S_{f(x)}x) + \sum_{j \in J_1(x), |J_1(x)|>1} \mu_j \bar{L}(S_j^{-1}x)1_{\{q(x)=0\}} + \\ &\sum_{j \in J_1(x)} \mu_j \bar{L}^f(S_0^{-1}S_j^{-1}S_{f(S_0^{-1}S_j^{-1}x)}x)1_{\{q(x)>0\}},\ x \in E_X \setminus \{x_0\} \end{aligned} \tag{15}$$

By expressing relations (15) in matrix form and taking into account that $\bar{L}^f := \bar{L}^f(\mathbf{e}_1(K+1))$ we obtain the expressions (11) for $a^f(x) = \bar{L}^f(x)$.

Denote now by $\tilde{\psi}_{x,k}^f = \sum_{i=0}^\infty \psi_{x,k}^f(i)z^i, |z| \leq 1$, the probability generating function (PGF) of the PDF $\psi_{x,k}^f(i)$ of the number of service completion at server k up to the end of busy period given that the initial state is $x \in E_X \setminus \{x_0\}$. With respect to the law of the total probability we get the following relations for the function $\psi_{x,k}^f(i)$,

$$\psi_{x,k}^f(i) = \frac{\lambda_{xu}(a)}{\lambda_x(a)}\psi_{u,k}^f(i-1) + \sum_{y \neq x,u} \frac{\lambda_{xy}(a)}{\lambda_x(a)}\psi_{y,k}^f(i),\ i \geq 1. \tag{16}$$

The first term on the right hand side of (16) represents the transition to state u accompanied with an event we count, that is a service completion at server k. The second term

stands for other possible transitions. The system (16) can be rewritten in terms of the PGF in the following form,

$$\tilde{\psi}^f_{x,k}(z) = \frac{z\lambda_{xu}(a)}{\lambda_x(a)}\tilde{\psi}^f_{u,k} + \sum_{y\neq x,u}\frac{\lambda_{xy}(a)}{\lambda_x(a)}\tilde{\psi}^f_{y,k}(z). \tag{17}$$

The expressions (17) can be modified using the property $\bar{N}^f_{L,k}(x) = \frac{d}{dz}\tilde{\psi}_{x,k}(z)\Big|_{z=1}$ in such a way that we get a system for the corresponding first moments,

$$\begin{aligned}&\bar{N}^f_{L,k}(x_0) = 1,\\ &\bar{N}^f_{L,k}(x) = \frac{1}{\lambda_x(a)}\Big[\lambda_{xu}(a) + \sum_{y\neq x}\lambda_{xy}(a)\bar{N}^f_{L,k}(y)\Big],\ x\in E_X\setminus\{x_0\}.\end{aligned} \tag{18}$$

For the model under study the system (18) is of the form

$$\begin{aligned}&\Big((N-l(x))\lambda + \sum_{j\in J_1(x)}\mu_j\Big)\bar{N}^f_{L,k}(x) = d_k(x)\mu_k +\\ &(N-l(x))\lambda\bar{N}^f_{L,k}(S_{f(x)}x) + \sum_{j\in J_1(x)}\mu_j\bar{N}^f_{L,k}(S_j^{-1}x)1_{\{q(x)=0\}} +\\ &\sum_{j\in J_1(x)}\mu_j\bar{N}^f_{L,k}(S_0^{-1}S_j^{-1}S_{f(S_0^{-1}S_j^{-1}x)}x)1_{\{q(x)>0\}},\ x\in E_X\setminus\{x_0\}.\end{aligned} \tag{19}$$

The last system can be also expressed in form (11) for $a^f(x) = \bar{N}^f_{L,k}(x)$ and $\bar{N}^f_{L,k} = \bar{N}_{L,k}(\mathbf{e}_1(K+1))$. For the mean total number of customers served $\bar{N}_L$ the term $d_k(x)\mu_k$ on the right hand side of (19) must be replaced by $\sum_{k=1}^K d_k(x)\mu_k$. □

Finally, one more performance measure in this section is of our interest, namely, the distribution of the maximal queue length in a busy period for the given control policy f. Denote by Q^f_{max} the maximum number of customers waiting in the queue during a busy period. For each fixed value $n\geq 0$ the event $\{Q^f_{max}\leq n\}$ is equivalent to the event that the process $\{X(t)\}_{t\geq 0}$ starting in state $\mathbf{e}_1(K+1)$, where the first server is busy, hits the empty state x_0 before visiting the subset of states

$$\begin{aligned}E_{max,n} =&\{x = (q(x), d_1(x),\dots,d_K(x)):\\ &q(x)\in\{n+1,n+2,\dots,N-\sum_{j=1}^K d_j(x)\},\ d_j(x)\in\{0,1\},\ j=1,\dots,K\}\end{aligned}$$

The probability $\bar{Q}^f_{max,n} = \mathbb{P}[Q^f_{max}\leq n]$ will be calculated by means of absorption probabilities for states in a set of absorbing states $E_{max,n}\cup\{x_0\}$ given that the initial state is $x\in E_{X,n} = E_X\setminus E_{max,n}\cup\{x_0\}$. Denote by

$$\mathbf{a}^f = (a^f(x): x\in E_{X,n}) \text{ and } \mathbf{b} = (b(x): x\in E_{X,n})$$

the column-vectors of dimension $|E_{X,n}| = |E_X| - |E_{max,n}| - 1 = \sum_{j=0}^K\binom{K}{j}(n+1) - 1$, where n is fixed. Denote further by $\bar{M}^f_3$ one of performance characteristics $\bar{Q}^f_{max,n}, n\geq 0$.

Proposition 3. *The performance measure $\bar{M}^f_3$ satisfies the relation*

$$\bar{M}^f_3 = \mathbf{e}_1'(|E_{X,n}|)\mathbf{a}^f, \tag{20}$$

where the vector $\mathbf{a}^f$ *is a solution of the system*

$$\tilde{\Lambda}^f(n)\mathbf{a}^f = -\mathbf{b}. \tag{21}$$

The matrix $\tilde{\Lambda}^f(n)$ *is obtained from* $\tilde{\Lambda}^f$ *by removing all columns and rows starting with the* $n+1$*, and*

$$\bar{M}_3^f = \bar{Q}_{max,n}^f,\ b(x) = \sum_{k=1}^{K} d_k(x)\mu_k 1_{\{|J_1(x)|=1\}},\ x \in E_{X,n}. \tag{22}$$

Proof. Denote by $\bar{Q}_{max,n}^f(x)$ the probability of absorption into empty state x_0 starting in $x \in E_{X,n}$, where $\bar{Q}_{max,n}^f = \bar{Q}_{max,n}^f(\mathbf{e}_1(K+1))$, where $\mathbf{e}_1(K+1)$ as before is the state after an arrival to an empty state x_0. The following system can be obtained by conditioning on the next visited state Using again the first principles,

$$\begin{aligned}
&\bar{Q}_{max,n}^f(x_0) = 1,\\
&\bar{Q}_{max,n}^f(x) = \frac{1}{\lambda_x(a)}\sum_{y\neq x}\lambda_{xy}(a)\bar{Q}_{max,n}^f(y),\ x \in E_{X,n},\\
&\bar{Q}_{max,n}^f(x) = 0,\ x \in E_{max,n}.
\end{aligned} \tag{23}$$

For the queueing system operation under the control policy f the system (23) is of the form,

$$\begin{aligned}
&\Big((N-l(x))\lambda + \sum_{j\in J_1(x)}\mu_j\Big)\bar{Q}_{max,n}^f(x) = (N-l(x))\lambda\bar{Q}_{max,n}^f(S_{f(x)}x) +\\
&\sum_{j\in J_1(x)}\mu_j\bar{Q}_{max,n}^f(S_j^{-1}x)1_{\{q(x)=0\}} +\\
&\sum_{j\in J_1(x)}\mu_j\bar{Q}_{max,n}^f(S_0^{-1}S_j^{-1}S_{f(S_0^{-1}S_j^{-1}x)}x)1_{\{q(x)>0\}},\ x \in E_{X,n}.
\end{aligned} \tag{24}$$

Then after a routine of (block) identification the system (24) can be expressed in form (21), where $a^f(x) = \bar{Q}_{max,n}^f(x), x \in E_{X,n}$. □

As we can see, calculating the performance characteristics requires solving very similar systems of equations. Thus, the same algorithm can be used for this purpose by substituting appropriate values into vectors $\mathbf{a}^f$ and $\mathbf{b}$, This versatility of the proposed approach greatly simplifies the application of algorithmic types of analysis of complex controlled queueing systems. In principle, we assume that for a fixed control threshold policy, the structure of the infinitesimal matrix can be even fully defined for an arbitrary number of servers, as will be proposed in the next section for the special case of the control policy where all thresholds are equal to 1. Thus we believe that matrix expressions can be derived explicitly from the presented matrix systems for performance characteristics. We leave this problem for our research in the near future.

3. FSF-Model

Here we discuss the FSF-Model which is a special case of the OTP-model, where $q_1 = q_2 = \cdots = q_K = 1$. The Markov-chain $\{X(t)\}_{t\geq 0}$ operating under the FSF-policy has a state space

$$E_X = \{x : q(x) = 0, |J_1(x)| < K\} \cup \{x : q(x) \geq 1, |J_1(x)| = K\}.$$

The states in E_X are divided in to levels y in the following way,

$$\mathbf{y} = \{x \in E : q(x) = 0, |J_1(x)| = y\},\ 0 \le y \le K,$$
$$\mathbf{y} = \{x \in E : q(x) = y, |J_1(x)| = K\},\ K+1 \le y \le N.$$

Denote by $s_{i,j} = \binom{K-j+i}{i}$ for $K \ge j$, then $|\mathbf{y}| = s_{y,y}$ for $1 \le y \le K$ and $|\mathbf{y}| = 1$ for $K+1 \le y \le N$. Within each level $y, 1 \le y \le K$, the states are ordered in the lexicographic order, where the rank of x in the level y with $|J_1(x)| = y$ and $|J_0(x)| = K - y$ can be evaluated by

$$\Delta_y(x) = \sum_{i=1, i\in J_0(x)}^{K-1} \frac{n_i(x)(K-i)!}{\left(\sum_{j=i}^{K} d_j(x)\right)!\left(K - \sum_{j=i}^{K} d_j(x)\right)!} + 1, \tag{25}$$

where $n_i(x) = |\{j : d_j(x) = 1, d_i(x) = 0, j > i\}|$ is the number of slower busy servers as the ith idle one. Note that this ordering of states differs from that defined in (4) and used in the policy iteration algorithm. In the lexicographic ordering within each level of states it is possible to obtain explicit matrix expressions for state probabilities in case of an arbitrary number of servers K. Denote further by $L_y, 1 \le y \le K$, matrices whose rows consist of ordered elements of level y.

Proposition 4. *The the system under FSF-policy is described by a QBD process with a block-three diagonal infinitesimal matrix of the form*

$$\Lambda = \begin{pmatrix}
A_{1,0} & A_{0,1} & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\
A_{2,0} & A_{1,1} & A_{0,2} & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\
0 & A_{2,1} & A_{1,2} & A_{0,3} & 0 & 0 & 0 & 0 & \dots & 0 \\
\ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \\
0 & \dots & 0 & A_{2,K-2} & A_{1,K-1} & A_{0,K} & 0 & 0 & \dots & 0 \\
0 & \dots & 0 & 0 & A_{2,K-1} & A_{1,K} & A_{0,K+1} & 0 & \dots & 0 \\
0 & \dots & 0 & 0 & 0 & A_{2,K} & A_{1,K+1} & A_{0,K+2} & \dots & 0 \\
\ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \\
0 & \dots & 0 & 0 & 0 & 0 & 0 & A_{2,N-2} & A_{1,N-1} & A_{0,N} \\
0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 & A_{2,N-1} & A_{1,N}
\end{pmatrix}. \tag{26}$$

The square blocks $A_{1,y}$ of dimension $s_{y,y}$ for $0 \le y \le K-1$ and 1 for $K \le y \le N$ consist of the rates to stay in the yth level, are defined as

$$A_{1,y} = I_{s_{y,y}}(\mathbf{e}'(s_{y,y}) \otimes [L_y B_{0,1} + (N-y)\lambda \mathbf{e}(s_{y,y})]),\ 0 \le y \le K-1,$$
$$A_{1,y} = (N-y)\lambda + m_K,\ K \le y \le N. \tag{27}$$

The blocks $A_{0,y}$ of dimension $s_{y-1,y-1} \times s_{y,y}$ for $1 \le y \le K$ and of dimension 1 for $K+1 \le y \le N$ consist of the rates to move upwards from the level $y-1$ to y due to arrivals and are defined as

$$A_{0,y} = (N-y+1)\lambda \begin{pmatrix}
I_{s_{0,y}} & 0 & 0 & 0 & \dots & 0 \\
 & I_{s_{1,y}} & 0 & 0 & \dots & 0 \\
 & \ddots & \ddots & \ddots & \ddots & \\
 & & & I_{s_{y-1,y}} & 0 & \dots & 0
\end{pmatrix},\ 1 \le y \le K, \tag{28}$$
$$A_{0,y} = (N-y+1)\lambda,\ K+1 \le y \le N.$$

The blocks $A_{2,y}$ of dimension $s_{y,y} \times s_{y-1,y-1}$ for $1 \leq y \leq K$ and of dimension 1 for $K+1 \leq y \leq N$ consist of the rates to move downwards from the level $y+1$ to y due to service completions and are defined as recursive matrices

$$\begin{aligned}
&A_{2,y} = B_{y,y+1},\ 1 \leq y \leq K,\ \textit{where}\\
&\quad B_{0,j} = (\mu_j, \mu_{j+1}, \ldots, \mu_K)',\\
&\quad B_{i,j} = \begin{pmatrix} B_{i-1,j} & & & & & I_{s_{i,j}}\mu_{j-i} \\ 0 & B_{i-1,j+1} & & & & I_{s_{i,j+1}}\mu_{j+1-i} \\ \ddots & \ddots & \ddots & \ddots & \ddots & \\ 0 & 0 & \ldots & 0 & B_{i-1,K} & I_{s_{i,K}}\mu_{K-i} \end{pmatrix},\\
&A_{2,y} = m_K,\ K+1 \leq y \leq N.
\end{aligned} \tag{29}$$

Proof. Analysing the transitions of the Markov-chain $\{X(t)\}_{t\geq 0}$ we get a system of balance equations in form

$$\begin{aligned}
&((N - \sum_{j=1}^{K} d_j(x) - q(x))\lambda + \sum_{j=1}^{K} d_j(x)\mu_j)\pi_x = (N - \sum_{j=1}^{K} d_j(x) - q(x) + 1)\lambda\\
&\times \sum_{k=1}^{K} 1_{\{\sum_{i=1}^{k} d_i(x) = k\}}\pi_{x-\mathbf{e}_k} + \sum_{j=1}^{K}(1 - d_j(x))\mu_j\pi_{x+\mathbf{e}_j},\ q(x) = 0,\ |J_1(x)| \leq K,\\
&((N - K - q(x))\lambda + m_K)\pi_x = (N - K - q(x) + 1)\lambda\pi_{x-\mathbf{e}_0} + m_K\pi_{x+\mathbf{e}_0},\\
&q(x) > 0,\ |J_1(x)| = K,
\end{aligned} \tag{30}$$

where $\pi_x = \lim_{t\to\infty} \mathbb{P}[X(t) = x], x \in E$. Expressing Equation (30) for the sub-vectors $\boldsymbol{\pi}_y$, $1 \leq y \leq K-1$, and the scalars $\boldsymbol{\pi}_0$ and $\boldsymbol{\pi}_y, K \leq y \leq N$, by means of defined blocks and taking into account the states' ordering (25) we get the system

$$\begin{aligned}
&\boldsymbol{\pi}_0 A_{1,0} = \boldsymbol{\pi}_1 A_{0,1},\\
&\boldsymbol{\pi}_y A_{1,y} = \boldsymbol{\pi}_{y-1} A_{0,y} + \boldsymbol{\pi}_{y+1} A_{2,y},\ 1 \leq y \leq K-2,\\
&\boldsymbol{\pi}_{K-1} A_{1,K-1} = \boldsymbol{\pi}_{K-2} A_{0,K-1} + \pi_K A_{2,K-1},\\
&\boldsymbol{\pi}_K A_{1,K} = \boldsymbol{\pi}_{K-1} A_{0,K} + \pi_{K+1} A_{2,K},\\
&\boldsymbol{\pi}_y A_{1,y} = \pi_{y-1} A_{0,y} + \pi_{y+1} A_{2,y},\ K+1 \leq y \leq N-1,\\
&\boldsymbol{\pi}_N A_{1,N} = \pi_{M-1} A_{0,N}.
\end{aligned} \tag{31}$$

Denote by $\boldsymbol{\pi}$ the macro-vector of the stationary state probabilities, i.e.

$$\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_{K-1}, \boldsymbol{\pi}_K, \ldots, \boldsymbol{\pi}_M).$$

Compiling relations (31) to the the system $\boldsymbol{\pi}\Lambda = \mathbf{0}$ we get then the infinitesimal matrix Λ is the form (26) with blocks defined by (27)–(29). □

Proposition 5. *The elements of the stationary probability macro-vector $\boldsymbol{\pi}$ satisfy the relations*

$$\boldsymbol{\pi}_0 = \prod_{j=1}^{K} M_{K-j}\boldsymbol{\pi}_K, \tag{32}$$

$$\boldsymbol{\pi}_y = \prod_{j=1}^{K-y} M_{K-j}\boldsymbol{\pi}_K,\ 1 \le y \le K-1, \tag{33}$$

$$\boldsymbol{\pi}_y = \frac{(N-K)!}{(N-y)!}\rho_K^{y-K}\boldsymbol{\pi}_K,\ K \le y \le N, \tag{34}$$

$$\boldsymbol{\pi}_K = \Big(\sum_{y=0}^{K-1}\prod_{j=1}^{K-y} M_{K-j} + \sum_{y=K}^{N}\frac{(N-K)!}{(N-y)!}\rho_K^{y-K}\Big)^{-1}, \tag{35}$$

where the matrices M_y satisfies the recursive relations

$$\begin{aligned} M_0 &= A_{0,1}A_{1,0}^{-1}, \\ M_y &= A_{2,y}(A_{1,y} - M_{y-1}A_{0,y})^{-1},\ 1 \le y \le K-1. \end{aligned} \tag{36}$$

Proof. The probability π_0 and sub-vectors $\boldsymbol{\pi}_y, 1 \le y \le K-2$, can be expressed from the balance equations (31) using a block forward elimination-backward substitution as

$$\begin{aligned} &\boldsymbol{\pi}_0 = A_{0,1}A_{1,0}^{-1}\boldsymbol{\pi}_1 = \boldsymbol{\pi}_1 M_0, \\ &\boldsymbol{\pi}_1 A_{1,1} = \boldsymbol{\pi}_1 M_0 A_{0,1} + \boldsymbol{\pi}_2 A_{2,1} \Rightarrow \boldsymbol{\pi}_1 = \boldsymbol{\pi}_2 A_{2,1}(A_{1,1} - M_0 A_{0,1})^{-1} = \boldsymbol{\pi}_2 M_1, \\ &\boldsymbol{\pi}_y A_{1,y} = \boldsymbol{\pi}_y M_{y-1} A_{0,y} + \boldsymbol{\pi}_{y+1} A_{2,y} \Rightarrow \boldsymbol{\pi}_y = \boldsymbol{\pi}_{y+1} A_{2,y}(A_{1,y} - M_{y-1}A_{0,y})^{-1} = \boldsymbol{\pi}_{y+1} M_y. \end{aligned}$$

We similarly obtain an expression for $\boldsymbol{\pi}_{K-1}$,

$$\begin{aligned} &\boldsymbol{\pi}_{K-1}A_{1,K-1} = \boldsymbol{\pi}_{K-1}M_{K-2}A_{0,K-1} + \boldsymbol{\pi}_K A_{2,K-1} \Rightarrow \\ &\boldsymbol{\pi}_{K-1} = A_{2,K-1}(A_{1,K-1} - M_{K-2}A_{0,K-1})^{-1}\boldsymbol{\pi}_K = M_{K-1}\pi_K \end{aligned}$$

The relations (32) and (33) are obtained then through a successive substitution. The relation (34) is obtained by solving (31) for $K \le y \le N$ recursively using

$$\boldsymbol{\pi}_y = \frac{(N-y+1)\lambda}{m_K}\boldsymbol{\pi}_{y-1}$$

starting from the last equation. The relation (35) for the probability π_K is determined using the normalizing condition $\boldsymbol{\pi}\mathbf{e}(N) = 1$. □

To calculate performance characteristics the expressions from the previous section applied to a control policy $f(x) = \text{argmax}_{j \in J_0(x)}\{\mu_j\}$ can be used. As an alternative to the policy-iteration algorithm we can use the proposed matrix-analytic solution (32)–(35) to obtain the matrix expressions for the performance characteristics in an explicit form.

Corollary 1. *The probability that the kth server is busy and the mean number of busy servers,*

$$\begin{aligned} \bar{U}_k = &\Big(\sum_{y=1}^{K-1}\prod_{j=1}^{K-y} M_{K-j}L_y\mathbf{e}_k(s_{y,y}) + \sum_{y=K}^{N}\frac{(N-K)!}{(N-y)!}\rho_K^{y-K}\Big)\times, \\ \times &\Big(\sum_{y=0}^{K-1}\prod_{j=1}^{K-y} M_{K-j} + \sum_{y=K}^{N}\frac{(N-K)!}{(N-y)!}\rho_K^{y-K}\Big)^{-1},\ \bar{C} = \sum_{k=1}^{K}\bar{U}_k. \end{aligned}$$

The mean number of customers in the queue,

$$\bar{Q} = \sum_{y=K}^{N} \frac{(y-K)(N-K)!}{(N-y)!} \rho_K^{y-K} \Big(\sum_{y=0}^{K-1} \prod_{j=1}^{K-y} M_{K-j} + \sum_{y=K}^{N} \frac{(N-K)!}{(N-y)!} \rho_K^{y-K} \Big)^{-1}.$$

The mean number of customers in the system,

$$\bar{N} = \bar{C} + \bar{Q} = \Big(\sum_{y=1}^{K-1} \prod_{j=1}^{K-y} M_{K-j} L_y \mathbf{e}(s_{y,y}) + \sum_{y=K}^{N} \frac{(y-K+1)(N-K)!}{(N-y)!} \rho_K^{y-K} \Big) \times,$$
$$\times \Big(\sum_{y=0}^{K-1} \prod_{j=1}^{K-y} M_{K-j} + \sum_{y=K}^{N} \frac{(N-K)!}{(N-y)!} \rho_K^{y-K} \Big)^{-1}$$

Mean length of a busy period,

$$\bar{L} = \frac{1}{N\lambda} \Big(\Big(\prod_{j=1}^{K} M_{K-j} \Big)^{-1} \Big(\sum_{y=0}^{K-1} \prod_{j=1}^{K-y} M_{K-j} + \sum_{y=K}^{N} \frac{(N-K)!}{(N-y)!} \rho_K^{y-K} \Big) - 1 \Big).$$

Mean number of customers served in a busy period,

$$\bar{N}_L = \Big(\prod_{j=1}^{K} M_{K-j} \Big)^{-1} \Big(\sum_{y=0}^{K-1} \prod_{j=1}^{K-y} M_{K-j} + \sum_{y=K}^{N} \frac{(N-K)!}{(N-y)!} \rho_K^{y-K} \Big).$$

The mean number of customers served by the kth server in a busy period and the distribution of the maximal queue length can be evaluated using the matrix systems (11) and (20) taking into account the structure (26) of the infinitesimal matrix Λ.

Proposition 6. *The mean number $\bar{N}_{L,k}$ of customers served in a busy period by the kth server satisfies the relation*

$$\bar{N}_{L,k} = \mathbf{e}_1'(K) \sum_{y=1}^{N} \Big(\prod_{j=1}^{y-1} T_j \Big) S_y, \; 1 \le k \le K, \tag{37}$$

where

$$\begin{aligned} &S_1 = A_{1,1}^{-1} \mathbf{b}_1, \; T_1 = -A_{1,1}^{-1} A_{0,2}, \\ &S_y = (A_{2,y-1} T_{y-1} + A_{1,y})^{-1} (\mathbf{b}_y - A_{2,y-1} S_{y-1}), \; T_y = -(A_{2,y-1} T_{y-1} + A_{1,y})^{-1} A_{0,y+1} \\ &\quad 2 \le y \le N-1, \\ &S_N = (A_{2,N-1} T_{N-1} + A_{1,N})^{-1} (\mathbf{b}_N - A_{2,N-1} S_{N-1}). \end{aligned} \tag{38}$$

The column-vector $\mathbf{b}_y = L_y \mathbf{e}_k (K+1) \mu_k$ for $1 \le y \le K$ and the scalar $\mathbf{b}_y = \mu_k$ for $K+1 \le y \le N$.

Proof. The system (11) can be rewritten for appropriate blocks in form

$$\begin{aligned} &A_{1,1} \mathbf{a}_1 + A_{0,2} \mathbf{a}_2 = \mathbf{b}_1, \\ &A_{2,y-1} \mathbf{a}_{y-1} + A_{1,y} \mathbf{a}_y + A_{0,y+1} \mathbf{a}_{y+1} = \mathbf{b}_y, \; 2 \le y \le N-1, \\ &A_{2,N-1} \mathbf{a}_{N-1} + A_{1,N} \mathbf{a}_N = \mathbf{b}_N \end{aligned}$$

The elements of $\mathbf{b}_y$ are equal to μ_k if for some state x of the level $\mathbf{y}$ $d_k(x) = 1$. This implies the relations for $\mathbf{b}_y$. Using a forward elimination - backward substitution we get the recursive relations

$$\mathbf{a}_y = S_y + T_y \mathbf{a}_{y+1},\ 1 \le y \le N-1,\ \mathbf{a}_N = S_N,$$

where S_y and T_y are defined by (38). This statement follows through recurrence substitution taking into account that $\bar{N}_{L,k} = \mathbf{e}_1'(K)\mathbf{a}_1$, since the level 1 consists of K states. □

The following statement for the matrix equation (20) can be proved in a similar way taking into account the structure (26) of the infinitesimal matrix Λ.

Proposition 7. *The probability of the maximum queue length in a busy period satisfies the relation*

$$\bar{Q}_{max,n} = \mathbf{e}_1'(K) \sum_{y=1}^{n} \Big(\prod_{j=1}^{y-1} T_j \Big) S_y, \tag{39}$$

where

$$\begin{aligned} &S_1 = A_{1,1}^{-1} A_{2,0},\ T_1 = -A_{1,1}^{-1} A_{0,2}, \\ &S_y = -(A_{2,y-1} T_{y-1} + A_{1,y})^{-1} A_{2,y-1} S_{y-1},\ T_y = -(A_{2,y-1} T_{y-1} + A_{1,y})^{-1} A_{0,y+1} \\ &\qquad 2 \le y \le n-1, \\ &S_n = -(A_{2,n-1} T_{n-1} + A_{1,n})^{-1} A_{2,n-1} S_{n-1}. \end{aligned} \tag{40}$$

4. PS-Model

In this section we discuss a queueing system with a preemption operating under a general threshold policy f defined as a sequence of threshold levels $(q_2, \dots, q_K)$. The first server in this system is permanently available for service while the jth slower server must be used as soon as the number of customers in the system increases up to the value $q_{j-1} + j - 2$. This server must be removed from the system when the number of customers becomes again less as $q_{j-1} + j - 2$. Denote by $\{Y(t)\}_{t \ge 0}$ the continuous-time Markov-chain with a state space $E_Y = \{y : y \in \mathbb{N}_0\}$. All the rates are the same as in the model without preemption. The infinitesimal matrix $\Lambda_Y^f = \lambda_{xy}(q_2, \dots, q_K)$ is then of the form:

$$\lambda_{xy}(q_2, \dots, q_K) = \begin{cases} \lambda & y = x+1, \\ m_j & y = x-1, \\ & q_{j-1} + j - 2 \le y \le q_j + j - 2,\ j = 2, \dots, K, \end{cases} \tag{41}$$

where $m_j = \sum_{i=1}^{j} \mu_i$ and $q_1 = 1$. The state transition diagram of the process $\{Y(t)\}_{t \ge 0}$ is illustrated in Figure 2.

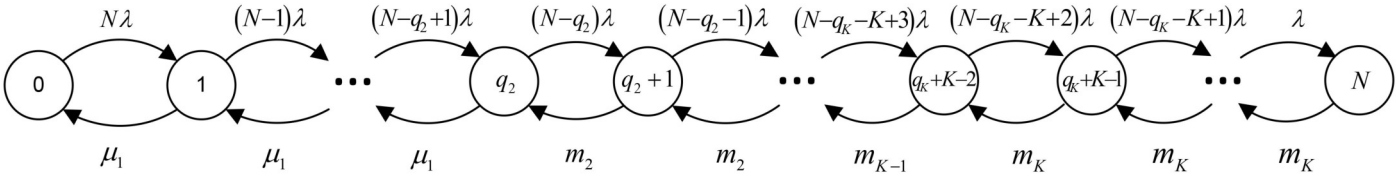


Figure 2. The state transition diagram for the queueing system S_2.

Proposition 8. *The steady-state probabilities* $\pi_y = \lim_{t\to\infty} \mathbb{P}[Y(t) = y]$ *of the PS-Model satisfy the relations*

$$\begin{aligned}
\pi_y &= \frac{N!}{(N-y)!}\prod_{i=1}^{j-1}\rho_i^{q_i-q_{i-1}+1}\rho_j^{y-q_j-j+2}\pi_0,\ q_j+j-1 \le y \le q_{j+1}+j-1,\\
&\qquad\qquad j = 1,\dots,K-1,\\
\pi_y &= \frac{N!}{(N-y)!}\prod_{i=1}^{K-1}\rho_i^{q_i-q_{i-1}+1}\rho_K^{y-q_K-K+2}\pi_0,\ q_K+K-1 \le y \le N,\\
\pi_0 &= \Big(1+\sum_{j=1}^{K}\sum_{y=q_j+j-1}^{q_{j+1}+j-1}\frac{N!}{(N-y)!}\prod_{i=1}^{j-1}\rho_i^{q_i-q_{i-1}+1}\rho_j^{y-q_j-j+2}+\\
&\quad \sum_{y=q_K+K-1}^{N}\frac{N!}{(N-y)!}\prod_{i=1}^{K-1}\rho_i^{q_i-q_{i-1}+1}\rho_K^{y-q_K-K+2}\Big)^{-1},
\end{aligned} \tag{42}$$

where $\rho_j = \frac{\lambda}{m_j}$, $j = 1,\dots,K$ *and* $\prod_{i=1}^{0}\ldots = 1$.

Proof. The proposition follows by solving the following equations

$$\begin{aligned}
N\lambda\pi_0 &= \mu_1\pi_1,\\
((N-q_{j+1}-j+1)\lambda + m_j)\pi_{q_{j+1}+j-1} &= (N-q_{j+1}-j+2)\lambda\pi_{q_{j+1}+j-2}\\
&\quad + m_{j+1}\pi_{q_{j+1}+j},\\
((N-y)\lambda + m_j)\pi_y &= (N-y+1)\lambda\pi_{y-1} + m_j\pi_{y+1},\\
&\quad q_j+j-1 \le y \le q_{j+1}+j-2,\\
((N-y)\lambda + m_K)\pi_y &= (N-y+1)\lambda\pi_{y-1} + m_K\pi_{y+1},\\
&\quad q_K+K-1 \le y \le N-1,\\
m_K\pi_N &= \lambda\pi_{N-1}
\end{aligned}$$

recursively for $j = 1,\dots,K-1$, where π_0 is calculated by means of the normalizing condition $\sum_{y=0}^{M}\pi_y = 1$. □

Corollary 2. *The kth server is busy with a probability,*

$$\bar{U}_k^f = \Bigg[\sum_{j=k}^{K}\sum_{y=q_j+j-1}^{q_{j+1}+j-1}\frac{M!}{(M-y)!}\prod_{i=1}^{j-1}\rho_i^{q_i-q_{i-1}+1}\rho_j^{y-q_j-j+2}+ \sum_{y=q_K+K-1}^{N}\frac{M!}{(M-y)!}\prod_{i=1}^{K-1}\rho_i^{q_i-q_{i-1}+1}\rho_K^{y-q_K-K+2}\Bigg]\pi_0.$$

The mean number of busy servers,

$$\bar{C}^f = \sum_{k=1}^{K}\bar{U}_k^f.$$

The mean number of customers in the queue,

$$\bar{Q}^f = \Bigg[\sum_{j=1}^{K-1}\sum_{y=q_j+j-1}^{q_{j+1}+j-1}\frac{(y-j)N!}{(N-y)!}\prod_{i=1}^{j-1}\rho_i^{q_i-q_{i-1}+1}\rho_j^{y-q_j-j+2} + \sum_{y=q_K+K-1}^{N}\frac{(y-K)N!}{(N-y)!}\prod_{i=1}^{K-1}\rho_i^{q_i-q_{i-1}+1}\rho_K^{y-q_K-K+2}\Bigg]\pi_0.$$

The mean number of customers in the system $\bar{N}^f = \bar{C}^f + \bar{Q}^f$.

The mean length of busy period,

$$\bar{L}^f = \frac{1}{N\lambda}\left[\sum_{j=1}^{K}\sum_{y=q_j+j-1}^{q_{j+1}+j-1}\frac{N!}{(N-y)!}\prod_{i=1}^{j-1}\rho_i^{q_i-q_{i-1}+1}\rho_j^{y-q_j-j+2}+\sum_{y=q_K+K-1}^{N}\frac{N!}{(N-y)!}\prod_{i=1}^{K-1}\rho_i^{q_i-q_{i-1}+1}\rho_K^{y-q_K-K+2}\right].$$

The mean number of customers served in a busy period,

$$\bar{N}_L^f = 1+\sum_{j=1}^{K}\sum_{y=q_j+j-1}^{q_{j+1}+j-1}\frac{N!}{(N-y)!}\prod_{i=1}^{j-1}\rho_i^{q_i-q_{i-1}+1}\rho_j^{y-q_j-j+2}+\sum_{y=q_K+K-1}^{N}\frac{N!}{(N-y)!}\prod_{i=1}^{K-1}\rho_i^{q_i-q_{i-1}+1}\rho_K^{y-q_K-K+2}.$$

Further we use a similar methodology that has been employed in previous section to derive expressions for $\bar{N}_{L,k}^f$ and $\bar{Q}_{max,n}^f$ with the knowledge that all levels **y** consist now of only one state, and hence in the sequel we omit some details.

Proposition 9. *The mean number of customers served by the kth server in a busy period satisfies the relation*

$$\bar{N}_{L,k}^f = \sum_{y=1}^{N}\Big(\prod_{j=1}^{y-1}T_j\Big)S_y, \tag{43}$$

where

$$\begin{aligned}
S_1 &= \frac{m_1+\mu_1 1_{\{k=1\}}}{(N-1)\lambda+m_1},\ T_1 = \frac{(N-1)\lambda}{(N-1)\lambda+m_1}\\
S_y &= \frac{m_j S_{y-1}+\mu_k 1_{\{j\geq k\}}}{(N-y)\lambda+m_j-m_jT_{y-1}},\ T_y = \frac{(N-y)\lambda}{(N-y)\lambda+m_j-m_jT_{y-1}},\\
&\quad q_j+j-1\leq y\leq q_{j+1}+j-1,\ 1\leq j\leq K-1,\\
S_y &= \frac{m_j S_{y-1}+\mu_K}{(N-y)\lambda+m_j-m_jT_{y-1}},\ T_y = \frac{(N-y)\lambda}{(N-y)\lambda+m_K-m_KT_{y-1}},\\
&\quad q_K+K-1\leq y\leq N-1,\\
S_N &= \frac{S_{N-1}}{1-T_{N-1}}.
\end{aligned} \tag{44}$$

Proof. The proof follows directly from (11) by forward elimination - backward substitution taking into account the structure (41) of the infinitesimal matrix Λ^f. □

Proposition 10. *The distribution function of the maximum queue length observed in a busy period satisfies the relation*

$$\bar{Q}_{max,n}^f = \sum_{y=1}^{n}\Big(\prod_{j=1}^{y-1}T_j\Big)S_y, \tag{45}$$

where

$$\begin{aligned}
S_1 =& \frac{m_1}{(N-1)\lambda + m_1},\ T_1 = \frac{(N-1)\lambda}{(N-1)\lambda + m_1} \\
S_y =& \frac{m_j S_{y-1} + \mu_k 1_{\{j \geq k\}}}{(N-y)\lambda + m_j - m_j T_{y-1}},\ T_y = \frac{(N-y)\lambda}{(N-y)\lambda + m_j - m_j T_{y-1}}, \\
& q_j + j - 1 \leq y \leq \min\{n-1, q_{j+1} + j - 1\},\ 1 \leq j \leq K-1, \\
S_y =& \frac{m_j S_{y-1} + \mu_K}{(N-y)\lambda + m_j - m_j T_{y-1}},\ T_y = \frac{(N-y)\lambda}{(N-y)\lambda + m_K - m_K T_{y-1}}, \\
& q_K + K - 1 \leq y \leq \min\{n-1, N-1\}, \\
S_n =& \frac{m_j S_{n-1} + \mu_K}{(N-n)\lambda + m_j - m_j T_{n-1}},\ n < N,\ S_n = \frac{S_{n-1}}{1 - T_{n-1}},\ n = N.
\end{aligned} \tag{46}$$

The last result can be rewritten in explicit form as well.

Proposition 11. *The distribution function of the maximum queue length observed in a busy period is given by*

$$\bar{Q}_{max,n} = \frac{\sum_{y=1}^{n} F(y)}{1 + \sum_{y=1}^{n} F(y)}, \tag{47}$$

where the function $F(n)$ has the following product form,

$$\begin{aligned}
F(y) =& \frac{m_j^{y-q_j-j+2}}{\prod_{i=q_j+j-1}^{y}((N-i)\lambda + m_j)} \prod_{i=1}^{j-1} \frac{m_i^{q_{i+1}-q_i+1}}{\prod_{s=q_i+i-1}^{q_{i+1}+i-1}((N-s)\lambda + m_i)}, \\
& q_j + j - 1 \leq y \leq q_{j+1} + j - 1,\ 1 \leq j \leq K-1, \\
F(y) =& \frac{m_K^{y-q_K-K+2}}{\prod_{i=q_K+K-1}^{y}((N-i)\lambda + m_K)} \prod_{i=1}^{K-1} \frac{m_i^{q_{i+1}-q_i+1}}{\prod_{s=q_i+i-1}^{q_{i+1}+i-1}((N-s)\lambda + m_i)}, \\
& q_K + K - 1 \leq y \leq N.
\end{aligned} \tag{48}$$

Proof. The function $\bar{Q}^f_{max,n}(x), x \in E_Y$ for the given policy f satisfy the following system,

$$\begin{aligned}
& \bar{Q}^f_{max,n}(0) = 1, \\
& ((N-y)\lambda + m_j)\bar{Q}^f_{max,n}(y) = (N-y)\lambda \bar{Q}^f_{max,n}(y+1) + m_j \bar{Q}^f_{max,n}(y+1), \\
& \qquad q_j + j - 1 \leq y \leq \min\{n-1, q_{j+1} + j - 1\},\ 1 \leq j \leq K-1, \\
& ((N-y)\lambda + m_K)\bar{Q}^f_{max,n}(y) = (N-y)\lambda + \bar{Q}^f_{max,n}(y+1) + m_K \bar{Q}^f_{max,n}(y-1), \\
& \qquad q_K + K - 1 \leq y \leq \min\{n-1, N\}, \\
& ((N-y)\lambda + m_K)\bar{Q}^f_{max,n}(n) = m_K \bar{Q}^f_{max,n}(n-1),\ n < N.
\end{aligned} \tag{49}$$

These difference equations can be rewritten as recurrent relation for $1 \leq y \leq n$,

$$\bar{Q}^f_{max,n}(y+1) - \bar{Q}^f_{max,n}(y) = \frac{m_j}{(N-y)\lambda + m_j}(\bar{Q}^f_{max,n}(y) - \bar{Q}^f_{max,n}(y-1)). \tag{50}$$

By iterating (50), taking into account the structure of difference equations for the threshold policy we obtain

$$\bar{Q}^f_{max,n}(y+1) - \bar{Q}^f_{max,n}(y) = F(y)(\bar{Q}^f_{max,n}(1) - \bar{Q}^f_{max,n}(0)), \tag{51}$$

where the function $F(y)$ has a product form (48). Summing (51) for $y = 1, \ldots, n$ yields

$$\bar{Q}^f_{max,n}(n+1) - \bar{Q}^f_{max,n}(1) = \sum_{y=1}^{n} F(n)(\bar{Q}^f_{max,n}(1) - \bar{Q}^f_{max,n}(0)), \tag{52}$$

where $\bar{Q}^f_{max,n}(n+1) = 0$ and $\bar{Q}^f_{max,n}(0) = 1$. Expressing $\bar{Q}^f_{max,n}(1)$ we obtain the explicit formula (47). □

5. Comparison Analysis

In this section we discuss the results after having computed the performance metrics for the following finite-source heterogeneous queueing models: Non-preemptive queueing system operating under the optimal threshold policy (OTP-Model), non-preemptive queueing system with a fastest server first policy (FSF-Model), preemptive queueing system (PS1-Model), where the kth server is used when at least k customers present in the system, and preemptive queueing system (PS2-Model) operating according to a given threshold policy. This policy we calculate using a similar heuristic formula obtained in [13], which can be rewritten in form

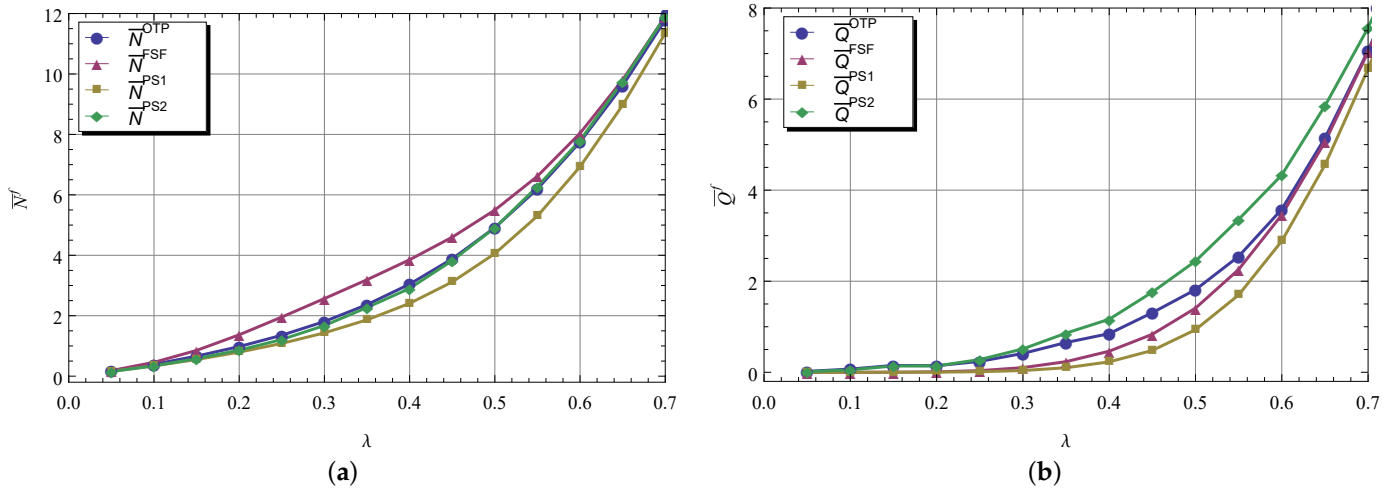
$$q_k = \max\left\{q_{k-1}, \left\lfloor \Big(\sum_{j=1}^{k-1} \mu_j - (N - \bar{N}^{PS1})\lambda\Big)\Big(\frac{1}{\mu_k} - \frac{k-1}{\sum_{j=1}^{k-1}\mu_j}\Big)\right\rfloor + k\right\}, \; 2 \le k \le K,$$

where $q_1 = 1$ and $(N - \bar{N}^{PS1})\lambda$ is an average arrival rate in the PS1-Model which is derived in explicit form.

In our experiments we fix the number of servers $K = 5$, the source capacity $N = 60$ and service intensities $(\mu_1, \mu_2, \mu_3, \mu_4, \mu_5) = (20, 8, 4, 2, 1)$. The rate λ will be varied in the interval $[0.01, 0.7]$. The choice of this interval is not random. At higher values of λ, the analysed functions become indistinguishable, since the corresponding queueing systems will have similar stochastic behaviour in a so-called heavy-traffic mode.

In Figure 3, we display the functions $\bar{N}^f$ (figure labeled by "a") and $\bar{Q}^f$ (figure labeled by "b") for different models as λ varies. We observe that the functions $\bar{N}^{FSF}$ and $\bar{Q}^{PS1}$ models are the natural upper and low bounds for $\bar{N}^{OTP}$. It is clear that the FSF-model is a particular case of the OTP and the queue with a preemption is always superior in performance comparing to the non-preemptive case.

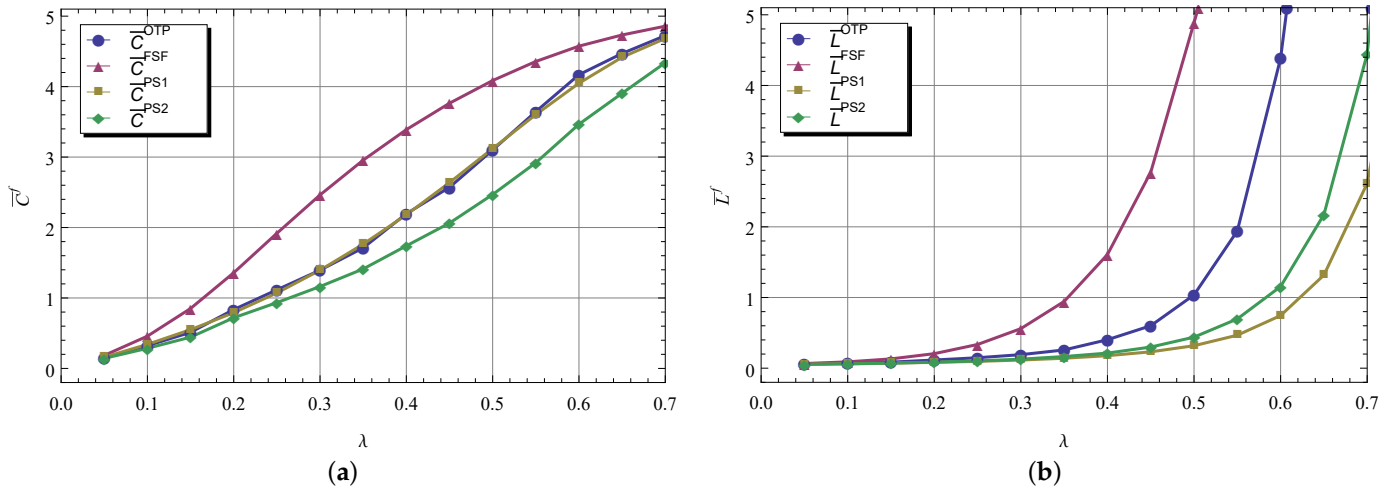
Differences between the functions $\bar{N}^{OTP}$ and $\bar{N}^{PS2}$ are almost not visible. This effect is also observed for other values of system parameters. It allows the preemptive system under a threshold policy to be used as an approximation of the original OTP-model. In contrary, the PS2-model exhibits the higher values of queue lengths while the PS1-model—the shortest. The OTP-model also has in average more waiting customers as in FSF-model which is not surprising, since the optimal policy minimizes in our case the mean number of customers in the system but not in the queue. It should also be noted that the higher the degree of heterogeneity of the servers, the greater the differences in performance functions for different models become.



(**a**) (**b**)

Figure 3. $\bar{N}^f$ (**a**) and $\bar{Q}^f$ (**b**) versus λ.

Figure 4 illustrates the influence of λ and model types on the functions $\bar{C}^f$ (figure labelled by "a") and $\bar{L}^f$ (figure labelled by "b"). The functions of the mean number of busy servers for the OTP- and PS1-models are very close to each other. Thus, by subtracting the mean number of busy servers in PS1-model from the mean number of customers in PS2-model, an approximation can be obtained for the mean queue length of the OTP-model. The functions $\bar{C}^{FSF}$ and $\bar{C}^{PS2}$ represent the upper and low bound for $\bar{C}^{OTP}$. The longest busy period appears in FSF-model. In this case the slower servers can be occupied with higher probability and then these servers remain busy for a very long time. As expected, the shortest busy period exhibits the preemptive PS1-model.



(**a**) (**b**)

Figure 4. $\bar{C}^f$ (**a**) and $\bar{L}^f$ (**b**) versus λ.

Figure 5 shows the effect of the service speed of kth server, $1 \leq k \leq K$, to the mean number of customers $\bar{N}_{L,k}$ served in a busy period (figures are labeled respectively by "a"–"f"). We observe that the slow servers begin to contribute to the number of customers served as the intensity of λ increases. The functions $\bar{N}^f_{L,k}$ are proportional to the rate λ, they are simply shifted to the right as λ is getting higher without changing their form. It can be observed also that the FSF-policy maximizes the number of customers served in a busy period at any server. This observation coincides with a statement in [14] that the fastest available server stochastically maximizes the number of service completions.

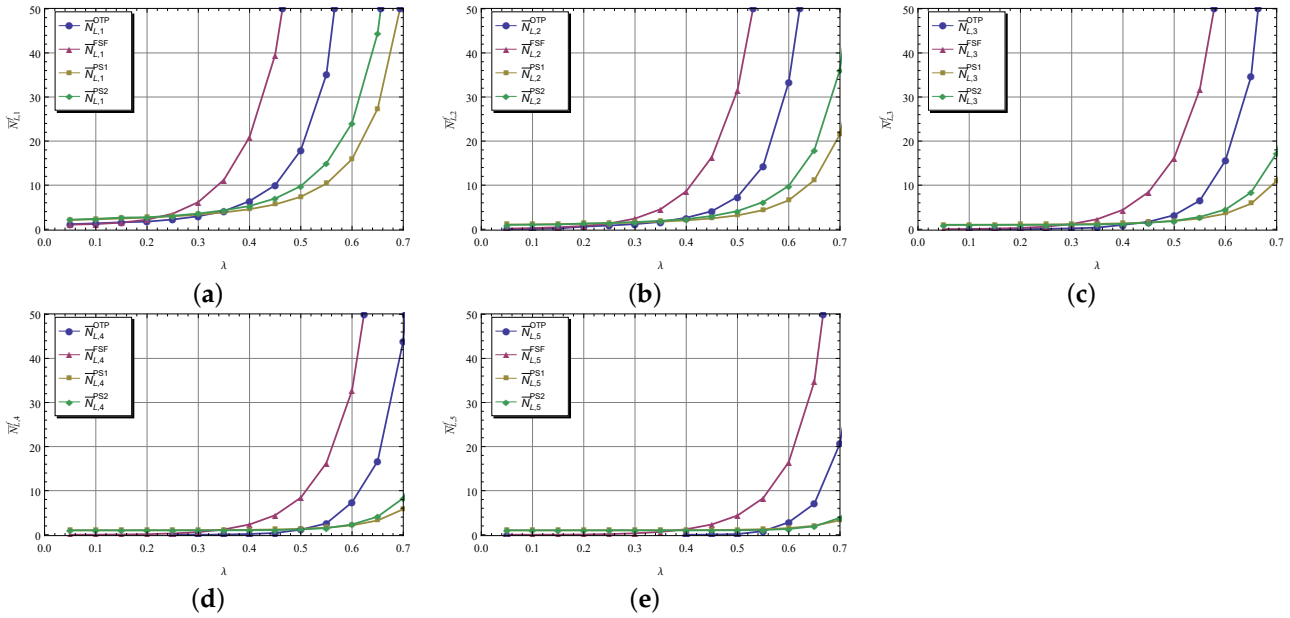


Figure 5. The mean number of customers $\bar{N}_{L,k}$ for $k = 1, \ldots, 5$ respectively in (**a**–**e**) versus λ.

We now focus on the results obtained for the maximum queue length observed during a busy period. To study the influence of system parameters and model type we summarized the results in Table 2 for OTP- and FSF-models and in Table 3—for PS1- and PS2-models. In tables we vary the rate λ keeping as before other parameters constant. The results compiled and presented in tables correlate with the graphs for the mean length of the busy period. The longer the busy period is, the more likely there will be fewer waiting customers in the queue. In the FSF-model it is more likely that there is an empty waiting line. As λ increases, the queues grow and hence we observe that for all models that the 99th percentile increases.

Table 2. The distribution function of the maximum queue length $\bar{Q}^{f}_{max,n}$ as λ varies for OTP and FSF.

n	$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 0.7$	n	$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 0.7$
0	0.79903	0.58580	0.47423	0.40562	0	0.99944	0.87367	0.60804	0.44909
1	0.94864	0.75918	0.61516	0.52072	1	0.99992	0.91039	0.61822	0.45087
2	0.98649	0.84411	0.68819	0.58156	2	0.99998	0.94535	0.63089	0.45254
3	0.99963	0.94722	0.77455	0.64604	3	0.99999	0.97086	0.64667	0.45413
4	0.99995	0.96327	0.80985	0.67923	4	1	0.98591	0.64667	0.45568
5	1	0.97991	0.84369	0.70681	5	1	0.99361	0.69029	0.45723
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
10	1	0.99956	0.96445	0.84265	10	1	0.999993	0.86795	0.46603
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
20	1	0.99987	0.99772	0.91243	20	1	1	0.99910	0.53571
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
40	1	1	0.98523	0.93216	40	1	1	1	0.99998

Table 3. The distribution function of the maximum queue length $\bar{Q}^{f}_{max,n}$ as λ varies for PS1 and PS2.

n	$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 0.7$	n	$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 0.7$
0	0.77220	0.53050	0.40404	0.32626	0	0.77220	0.53050	0.40404	0.32626
1	0.95182	0.74672	0.57128	0.45002	1	0.93781	0.74672	0.57128	0.45002
2	0.99112	0.86396	0.67401	0.52063	2	0.98639	0.85561	0.66394	0.51281
3	0.99851	0.92976	0.74748	0.56866	3	0.99723	0.91588	0.72366	0.54955
4	0.99976	0.96533	0.80376	0.60468	4	0.99945	0.95076	0.77102	0.57612
5	0.99996	0.98344	0.84775	0.63305	5	0.99989	0.97293	0.80967	0.59627
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
10	1	0.99971	0.96288	0.72316	10	1	0.99923	0.93623	0.66395
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
20	1	1	0.99951	0.86029	20	1	1	0.99854	0.79730
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
40	1	1	1	0.99999	40	1	1	1	0.99998

We have also conducted various experiments where we analyzed the effect of the number of servers, the source capacity, the level of heterogeneity and so on to performance metrics of non-preemptive heterogeneous systems and possible approximations through their preemptive equivalents. Due to the space limitations of the paper, we omit these results. As a generalisation, we can state that the main observations we made in the presented examples remain valid also for other values of system parameters.

6. Conclusions

Finite-source multi-server heterogeneous systems without priority service interruption are described using a multivariate Markov-chains. For such a systems we have found the optimal threshold policy and calculated the corresponding performance measure. Both analytical and numerical studies of such a system face constraints on the dimensionality of the problem, i.e., on the number of servers. In this paper we have also tried to understand, whether there are simplified variations of the main model which are appropriate for boundary values calculation or even for approximation of the main model but without constraint on the number of servers. We have analyzed non-preemptive and preemptive queues and provided comparison analysis of the performance characteristics.

Author Contributions: Conceptualization, D.E. and N.S.; formal analysis, investigation, methodology, software and writing, D.E., N.S. and J.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by the Austro-Hungarian Cooperation (OMAA) Grant No. 106öu4, 2021. (D. Efrosinin and J. Sztrik: Mathematical model development, investigation, methodology, N. Stepanova: Numerical analysis and validation). This paper has been supported by the RUDN University Strategic Academic Leadership Program (D. Efrosinin: Formal analysis).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors are very grateful to the reviewers of the paper for their comments which improved its quality.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ke, J.; Wang, K. Cost analysis of the $M/M/R$ machine repair problem with balking, reneging and server breakdown. *J. Oper. Res. Soc.* **1999**, *50*, 275–282.
2. Stecke, K.E. Machine interference: Assignment of machines to operators. In *Handbook of Industrial Engineering*, 2nd ed.; John Wiley & Sons Inc.: New York, NY, USA, 1992; pp. 1460–1494.
3. Efrosinin, D. *Controlled Queueing Systems with Heterogeneous Servers: Dynamic Optimization and Monotonicity Properties of Optimal Control Policies in Multiserver Heterogeneous Servers*; VDM Verlag: Saarbrücken, Germany, 2008.
4. Efrosinin, D.V.; Rykov, V.V. On performance characteristics for queueing systems with heterogeneous servers. *Autom. Remote Control* **2008**, *69*, 61–75.
5. Sztrik, J. Finite-source queueing systems and their applications. In *Formal Methods in Computing*; Ferenczi, M., Pataricza, A., Rnyai, L., Eds.; Akadémia Kiadó: Budapest, Hungary, 2005; pp. 311–356.
6. Jain, M. Finite source $M/M/r$ queueing system with queue-dependent servers. *Comput. Math. Appl.* **2005**, *50*, 187–199.
7. Iravani, S.M.R.; Krishnamurthy, V.; Chao, G.H. Optimal server scheduling in nonpreemptive finite-population queueing systems. *Queueing Syst.* **2007**, *55*, 95–105.
8. Deslay, M.; Kolfal, B.; Ingolfsson, A. Maximizing throughput in finite-Source parallel queue systems. *Eur. J. Oper. Res.* **2012**, *217*, 554–559.
9. Sztrik, J.; Roszik, J. *Performance Analysis of Finite-Source Retrial Queueing Systems with Heterogeneous Non-Reliable Servers and Different Service Policies*; Research Report; University Debrecen: Debrecen, Hungary, 2001.
10. Chakka, R.; Mitrani I. Heterogeneous multiprocessor systems with breakdowns. *Theor. Comput. Sci.* **1994**, *125*, 91–109.
11. Jain, M.; Rakhee, M.S. N-policy for a machine repair system with spares and reneging. *Appl. Math. Model.* **2004**, *28*, 513–531.
12. Jain, M.; Meena, R.K. Vacation model for Markov machine repair problem with two heterogeneous unreliable servers and threshold recovery. *J. Ind. Eng. Int.* **2018** *14*, 143–152.
13. Efrosinin, D.; Stepanova, N.; Sztrik, J.; Plank, A. Approximations in performance analysis of a controllable queueing System with heterogeneous servers. *Mathematics* **2020**, *8*, 1803.
14. Righter, R. Optimal policies for scheduling repairs and allocating heterogeneous servers. *J. Appl. Probab.* **1996**, *33*, 536–547.