applied sciences

MDPI

Article

Stalling in Queuing Systems with Heterogeneous Channels

Leonidas Sakalauskas [1,*], Liudvikas Kaklauskas [1,2] and Renata Macaitiene [1]

[1] Faculty of Business and Technologies, Šiauliai State Higher Education Institution, Aušros Av. 40, LT-76241 Šiauliai, Lithuania; l.kaklauskas@svako.lt (L.K.), r.macaitiene@svako.lt (R.M.)
[2] Siauliai Academy, Vilnius University, Vytauto Str. 84, LT-76352 Šiauliai, Lithuania
* Correspondence: leonidas.sakalauskas@mif.vu.lt

Abstract: The present paper considers a model of stalling in a queueing system (QS) with any number of different capacities of heterogeneous servers. The state graph and the corresponding linear system for steady-state probabilities are derived using the standard Markov chain technique. The obtained solution of the steady-state probabilities model is numerically stable; the complexity of the corresponding expressions does not depend on the number of QS states; thus, they enable us to analytically study the QS characteristics. Optimization of a stalling buffer is considered as well, and it is shown that stalling helps us to solve the slow server problem under an appropriate choice of stalling buffer size, making the slow servers usable under various values of system load. The asymptotic conditions of optimal query distribution in channels, when the ratio of the capacities of the fast and slow servers is increasing, are also established. Moreover, some applications of the developed model in heterogeneous server clusters and in work productivity modelling for forest harvesting applications are discussed.

Keywords: heterogeneous channels; stalling buffer; queueing system

check for updates

Citation: Sakalauskas, L.; Kaklauskas, L.; Macaitiene, R. Stalling in Queuing Systems with Heterogeneous Channels. *Appl. Sci.* **2024**, *14*, 773. https://doi.org/10.3390/app14020773

Academic Editor: Ernesto Sanchez

Received: 9 December 2023
Revised: 2 January 2024
Accepted: 11 January 2024
Published: 16 January 2024

1. Introduction

Queueing systems (QSs) with different capacity channels are often used in technology and business. Usually, the theory of discrete Markov processes is used for QS modelling, and the probabilities of steady states are calculated from a linear equations system; here, the coefficients are equal to the rates of transition, while the order of the system is equal to the number of states [1]. If all channels of the linear system have the same capacity, this system has an analytical solution with a complexity that does not depend on the number of states. However, in general, this system of equations for the state probabilities of a QS with several heterogeneous channels can depend on the number of states, which might be significant, and it can be unstable.

The queueing strategies in QSs with two or several different channels have been analysed by many authors, see, for example [2–11]. Computer programs are developed to determine the optimal allocation of storage spaces among three heterogeneous servers, by calculating the bounds of sizes of finite sources for different traffic intensities [12]. Particular cases of the steady-state behaviour of a discrete-time queueing problem with S heterogeneous groups with a discipline FCFS and a limited waiting space were analysed in [13,14]; a set of heterogeneous and exponential servers were discussed in [15]. El-Taha and Stidham [16], using deterministic (sample-path) analysis, generalized and extended the fundamental properties of systems with "stationary deterministic flows". They proved the renewal–reward theorem and established a relationship that shows that the "operational analysis" definition of average service times—when considered as the observation period $t \to \infty$—coincides with the standard definition of average service times for all stable queueing systems. Moreover, the equilibrium in a single-server queueing system with retrials and strategic timing of the customers was analysed in [17]. The systems modelled by M/M/n queues with heterogeneous servers and non-informed customers show that there is a value of arrival rate; below this, the slow server should not be used, and above which, it should be used [18]. A multi-server controllable queueing system with heterogeneous servers was considered in [19]; here, several monotonicity properties of optimal policies for such a system were proved. Efrosinin and Rykov studied [7] the queuing system with K heterogeneous servers using a heuristic approach. Efrosinin and Sctrik [20,21] showed that, in an M/M/K queue, the optimal threshold levels may also depend on the states of slower servers; however, this influence is negligible. In [22], a finite-source queueing system serving one class of customers and consisting of heterogeneous servers with unequal service intensities and of one common queue were investigated. The authors of [23] provided an analysis of the dependence of the convergence of experimental results on the type of distribution used by the system parameters in heterogeneous queueing resource system with an unlimited number. In [24], a multi-server infinite buffer queueing system with additional servers (assistants) providing help to the main servers when they encountered problems was analysed through multi-dimensional Markov chains. The authors of [25] analysed heterogeneous queues where servers differ not only in service rates but also in operating costs; they presented a simple heuristic solution.

The proposed that the solutions for systems of heterogeneous services can also be applied in specific tasks of computer networks and multiprocessor systems in addition to solving social tasks and others. Heterogeneous server problems occur in multiprocessor systems, which combine CPU and GPU processors [26–30]. Expressions of the queue length distributions of the GPS and individual traffic flows were presented in [31]. The Quality of Service (QoS) in different service classes with real-time and non-real-time traffic integration is an important issue in WiMAX systems; therefore, the authors of [32,33] proposed a cross-layer QoS support scheduling framework and a corresponding opportunistic scheduling algorithm to provide QoS support to the heterogeneous traffic in a single-carrier WiMAX point-to-multipoint (PMP) system. The presented model is applied to the uplink transmission in the single-carrier WiMAX system as a multi-class priority TDMA queueing system in order to analyse the average packet delays of different service classes. The authors of [34] demonstrated that a processor's fast cores may not be ideal for system workloads and that less can be more in some situations. Furthermore, in [35], a novel fuzzy testing technique, HFuzz, was proposed as a solution enabling efficient testing on real heterogeneous architectures. Research results showed that such a system can perform well, using much more constrained resources than are usually available.

This paper aims to show that stalling is an efficient approach for exploiting and managing heterogeneous systems. The model of stalling in QSs with two heterogeneous servers has been considered in [9] by the first two authors of this paper, where the explicit probabilities of steady states were derived. The results showed that the appropriate choice of stalling buffer size is helpful in solving the slow server problem. The asymptotic approximations of optimal stalling buffer size, when the ratio of the capacities of fast and slow servers is increasing, have also been established. The application of the model developed in computer networks has been discussed as well. Investigation of the developed model enables us to conclude that stalling is a universal solution which can be implemented into any heterogeneous system.

The findings presented in this paper for several-server QSs might be useful for the investigation of systems with any number of heterogeneous channels or servers with different capacities. Since the analytical study of systems with heterogeneous channels is rather complicated, a system with two types of servers, equipped with a stalling buffer and a finite waiting line, is presented and analysed.

2. QS with Heterogeneous Channels and Stalling Buffer

2.1. State Graph of QS

Let us consider a queueing system consisting of $m + n$ heterogeneous channels, where m is the number of fast channels, n is the number of slow channels, K is the length of the stalling buffer, and M is the length of the waiting line and the buffer (see Figure 1). Assume that the interarrival time to the system is distributed exponentially with parameter λ and the service time is distributed also under this law with parameters μ1 and μ2, respectively, for fast and low channels, i.e., $\mu_1 > \mu_2$; the queries are served in the FCFS discipline. If the query—after being released into the system—finds a free fast channel, then it is served immediately; otherwise, it goes to a stalling buffer of *K* length, where it waits until the efficient channels become free. One query is served only by one channel without a break. If all the places in the stalling buffer are occupied, then the arrived query transfers to the slow channel. If the slow channel is occupied as well, then the application waits in the queue at a waiting buffer of length *M*. If all the places in the waiting and stalling buffers are occupied, then the query is rejected and it is deemed lost.

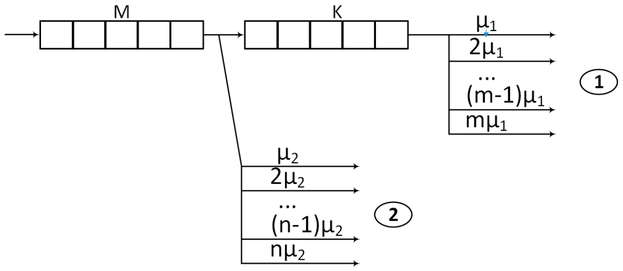

Figure 1. Scheme of QS with stalling buffer.

The main parameter of the QS is the coefficient of utilization $\rho = \frac{\lambda}{m \cdot \mu_1 + n \cdot \mu_2}$. Note, if the system is full ($\rho > 1$), then all the channels, i.e., slow and fast ones, should work because the service is most efficient if the system works at maximal capacity. If some channels work with idle time, then the system will process a smaller number of queries. If more queries come in than the system can serve, then unserved queries will be deemed lost, if the waiting line is finite.

2.2. Calculation of Steady-State Probabilities

Assume that the state of a QS with $m + n$ heterogeneous channels, a stalling buffer of capacity *K*, and waiting line length *M*, be defined by numbers: $\boldsymbol{i},\ \boldsymbol{j},\ \boldsymbol{k}$, where *i* and *j* are the numbers of queries served in fast and slow channels, respectively, and *k* is the number of stalled or waiting queries—$0 \leq \boldsymbol{i} \leq \boldsymbol{m},\ 0 \leq \boldsymbol{j} \leq \boldsymbol{n},\ 0 \leq \boldsymbol{k} \leq \boldsymbol{K} + \boldsymbol{M},\ \boldsymbol{m} \geq 1, \boldsymbol{n} \geq 0,\ \boldsymbol{K} \geq 0,\ \boldsymbol{M} \geq 0$. Denote the corresponding steady-state probabilities by $P_{i,j,k}$. Let us draw the state graph of such a QS inn Figure 2; here, the vertexes denote the states connected by arrows representing transitions with non-null probability from one state to the other [1].

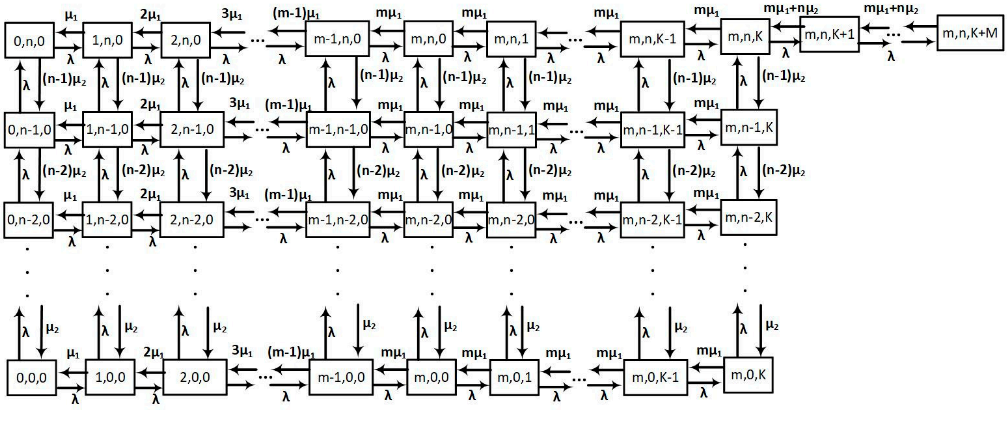

Figure 2. Graph of QS state.

The states on the graph are specified as well:

$P_{0,0,0}$—all channels and stalling buffers are free;

$P_{i,0,0}$—in the fast channels are the *i* queries, the slow channels and stalling buffer are free;

$P_{m,0,k}$—the fast channels are occupied with m queries, the slow channels are free, the k queries are stalled in the stalling buffer;$P_{m,0,K}$—the fast channels are occupied with m queries, the slow channels are free, the stalling buffer is full;

$P_{0,j,0}$—in the slow channels are j queries, the fast channels and the stalling buffer are free;

$P_{i,j,0}$—in the fast channels are i queries, in the slow channels are j queries, the stalling buffer is free;

$P_{m,j,k}$ —the fast channels are occupied with m queries, in the slow channels are j queries, k queries are stalled in the stalling buffer;

$P_{m,j,K}$—the fast channels are occupied with m queries, in the slow channels are j queries, the stalling buffer is full;

$P_{0,n,0}$—only the slow channels are occupied with n queries;

$P_{i,n,0}$—in the fast channels are i queries, in the slow channels are n queries, the stalling buffer is free;

$P_{m,n,k}$—the fast channels are occupied with m queries, the slow channels are occupied with n queries, k queries are stalled in the stalling buffer;

$P_{m,n,k+K}$—the fast channels are occupied with m queries, the slow channels are occupied with n queries, the stalling buffer is full, k queries are waiting in line;

$P_{m,n,K+M}$—the fast channels are occupied with m queries, in the slow channels are n queries, the stalling buffer and the queue are full, and the next arrived query will be lost;

Typically, the probabilities of these states are calculated using a linear equation system whose coefficients are equal to the rates of transition while the order of the system is equal to the number of states [1]. Thus, one can derive steady-state equations according to the steady-state graph in Figure 2. They are given in the following system (1):

$$\begin{gathered}
P_{0,j,0}\cdot(\lambda + j\cdot\mu_2) = P_{1,j,0}\cdot\mu_1 + P_{0,j+1,0}\cdot(j+1)\cdot\mu_2,\quad 0 \le j < n,\\
P_{i,j,0}\cdot(\lambda + i\cdot\mu_1 + j\cdot\mu_2) = P_{i-1,j,0}\cdot\lambda + P_{i+1,j,0}\cdot(i+1)\cdot\mu_1 + P_{i,j+1,0}\cdot(j+1)\cdot\mu_2,\quad 0 < i < m, 0 \le j < n,\\
P_{m,j,0}\cdot(\lambda + m\cdot\mu_1 + j\cdot\mu_2) = P_{m-1,j,0}\cdot\lambda + P_{m,j,1}\cdot m\cdot\mu_1 + P_{m,j+1,0}\cdot(j+1)\cdot\mu_2,\quad 0 \le j < n,\\
P_{m,j,k}\cdot(\lambda + m\cdot\mu_1 + j\cdot\mu_2) = P_{m,j,k-1}\cdot\lambda + P_{m,j,k+1}\cdot m\cdot\mu_1 + P_{m,j+1,k}\cdot(j+1)\cdot\mu_2,\quad 0 \le j < n,\ 0 < k < K,\\
P_{m,0,K}\cdot(\lambda + m\cdot\mu_1) = P_{m,0,K-1}\cdot\lambda + P_{m,1,K}\cdot\mu_2,\\
P_{m,j,K}\cdot(\lambda + m\cdot\mu_1 + j\cdot\mu_2) = P_{m,j,K-1}\cdot\lambda + P_{m,j+1,K}\cdot(j+1)\cdot\mu_2 + P_{m,j-1,K}\cdot\lambda,\quad 0 < j < n,\\
P_{0,n,0}\cdot(\lambda + n\cdot\mu_2) = P_{1,n,0}\cdot\mu_1,\\
P_{i,n,0}\cdot(\lambda + i\cdot\mu_1 + n\cdot\mu_2) = P_{i-1,n,0}\cdot\lambda + P_{i+1,n,0}\cdot(i+1)\cdot\mu_1,\quad 0 < i < m,\\
P_{m,n,0}\cdot(\lambda + m\cdot\mu_1 + n\cdot\mu_2) = P_{m-1,n,0}\cdot\lambda + P_{m,n,1}\cdot m\cdot\mu_1,\\
P_{m,n,k}\cdot(\lambda + m\cdot\mu_1 + n\cdot\mu_2) = P_{m,n,k-1}\cdot\lambda + P_{m,n,k+1}\cdot m\cdot\mu_1,\quad 0 < k < K,\\
P_{m,n,K}\cdot(\lambda + m\cdot\mu_1 + n\cdot\mu_2) = P_{m,n,K-1}\cdot\lambda + P_{m,n,K+1}\cdot(m\cdot\mu_1 + n\cdot\mu_2) + P_{m,n-1,K}\cdot\lambda,\\
P_{m,n,K+k}\cdot(\lambda + m\cdot\mu_1 + n\cdot\mu_2) = P_{m,n,K+k-1}\cdot\lambda\ + P_{m,n,K+k+1}\cdot(m\cdot\mu_1 + n\cdot\mu_2),\quad 0 < k < M,\\
P_{m,n,K+M}\cdot(m\cdot\mu_1 + n\cdot\mu_2) = P_{m,n,K+M-1}\cdot\lambda.
\end{gathered} \tag{1}$$

The obtained linear equation system is homogeneous; thus, the normalization condition is needed to ensure the uniqueness of the following solution:

$$\sum\nolimits_{\boldsymbol{i}=0}^{\boldsymbol{m}}\sum\nolimits_{\boldsymbol{j}=0}^{\boldsymbol{n}}\boldsymbol{P}_{\boldsymbol{i},\boldsymbol{j},0} + \sum\nolimits_{\boldsymbol{k}=1}^{\boldsymbol{K}}\sum\nolimits_{\boldsymbol{j}=0}^{\boldsymbol{n}}\boldsymbol{P}_{\boldsymbol{m},\boldsymbol{j},\boldsymbol{k}} + \sum\nolimits_{\boldsymbol{k}=1}^{\boldsymbol{M}}\boldsymbol{P}_{\boldsymbol{m},\boldsymbol{n},\boldsymbol{K}+\boldsymbol{k}} = 1. \tag{2}$$

The specific case of the steady-state system, when $\boldsymbol{m} = \boldsymbol{n} = 1$, is also given in [9]. Generally, the order of the linear equation system (1)–(2) is equal to the number of QS states $(m + K + 1)\cdot(n + 1) + M$. Note that solving this system using Cramer's rule might become numerically complicated, because the implementation of this rule requires a third-order complexity algorithm from the number of states; in addition, this system is often is ill-posed.

Looking for a more simple, explicit solution led us to apply the following parameters to the characterization of a QS with a stalling buffer:

- Coefficient of utilization of fast servers with switched-out slow channels $\boldsymbol{q} = \frac{\lambda}{m \cdot \mu_1}$;
- Rate of capacity of fast and slow servers $r = \frac{\mu_1}{\mu_2}$;
- Stalling buffer length K;
- Waiting queue length M.

Now, the coefficient of QS utilization is as follows:

$$\rho = \frac{\lambda}{m \cdot \mu_1 + n \cdot \mu_2} = \frac{q}{1 + \frac{n}{m \cdot r}}.$$

Let us denote the polynomial functions and ratios that can be used to explicitly derive the steady-state probabilities from systems (1) and (2):

$$R_{i,j} = \sum\nolimits_{s=j}^{n} \frac{s!}{j! \cdot (s-j)!} P_{\min(i,m),s,\max(i-m,0)} \tag{3}$$

$$A_{i,j} = \frac{R_{i,j}}{R_{0,j}} \tag{4}$$

$$h_j = \frac{R_{m+K,j}}{R_{m+K,n}} \tag{5}$$

The explicit expressions of QS steady-state probabilities are given by Theorem A1.

Theorem A1. *Steady-state probabilities of QS with stalling, defined by system (1)–(2), are as follows, if $q > 0, \rho > 0,\ m \geq 1, n \geq 0,\ K \geq 0, M \geq 0$:*

$$P_{i,j,k} = P_{m,n,K} \cdot \sum\nolimits_{s=j}^{n} h_s \cdot \frac{A_{i+k,s}}{A_{m+K,s}} \frac{(-1)^{s-j} \cdot s!}{(s-j)! \cdot j!},\quad 0 \leq i \leq m, 0 \leq j \leq n,\ 0 \leq k \leq K, \tag{6}$$

$$P_{m,n,k+K} = P_{m,n,K} \cdot \rho^k,\quad 0 \leq k \leq M. \tag{7}$$

$$P_{m,n,K} = \frac{q^K \cdot p_1}{h_0 \cdot (1 + p_1 \cdot C_1) + q^K \cdot p_1 \cdot S_1}, \tag{8}$$

$$\text{where}\quad p_1 = \frac{\frac{(m \cdot q)^m}{m!}}{\sum_0^{m-1} \frac{(m \cdot q)^s}{s!}}, \tag{9}$$

$$C_1 = \sum\nolimits_{s=0}^{K} q^s = \frac{1 - q^{M+1}}{1 - q} \quad if\ q \neq 1,\quad otherwise\ C_1 = K + 1,$$

$$S_1 = \sum\nolimits_{s=1}^{M} \rho^s = \frac{\rho - \rho^{M+1}}{1 - \rho} \quad if\ \rho \neq 1,\quad otherwise\ S_1 = M. \tag{10}$$

Proof of Theorem A1 is given in Appendix A. For the proof of this theorem, we will use two Lemmas given below. Their proofs are also given in Appendix A.

Lemma A1. *The equalities exist as follows, if $\boldsymbol{m} \geq \mathbf{1}, \boldsymbol{n} \geq \mathbf{0}, \boldsymbol{K} \geq \mathbf{0}$:*

$$\boldsymbol{A_{0,j} = 1,\ A_{1,j} = q \cdot m + \frac{j}{r},\ 0 \leq j \leq n,}$$

$$A_{i+1,j} = \frac{\left(\boldsymbol{q} \cdot \boldsymbol{m} + \frac{j}{r} + \mathbf{min}(\boldsymbol{i},\ \boldsymbol{m})\right) \cdot \boldsymbol{A_{i-1,j}}}{\mathbf{min}(\boldsymbol{i} + \mathbf{1}, \boldsymbol{m})} - \frac{\boldsymbol{m} \cdot \boldsymbol{q} \cdot \boldsymbol{A_{i-1,j}}}{\mathbf{min}(\boldsymbol{i} + \mathbf{1}, \boldsymbol{m})}, \mathbf{1} < \boldsymbol{i} \leq \boldsymbol{m} + \mathrm{K},\ 0 \leq \boldsymbol{j} \leq \boldsymbol{n}.$$

Lemma A2. *The ratios in (5) satisfy the recursive relation:*

$$\boldsymbol{h_n = 1,\ h_{j-1} = h_j \cdot \left(\frac{A_{m+K+1,j}}{q \cdot A_{m+K,j}} - 1 \right) + \frac{n!}{(j-1)! \cdot (n-j+1)!},\ 0 < j \le n.}$$

Remark. *Note that the sums* $\boldsymbol{R_{i,0} = \sum_{j=0}^{n} P_{min(i,m),j,max(i-m,0)}}$*,* $\boldsymbol{0 \le i < m + K}$*, according to the definition, are the steady-state probabilities of several queries in fast channels and the stalling buffer occupies both together. Using Lemma A1 and well-known formulas of steady-state probabilities in a multi-channel queueing system with the finite queue* [1]*, we can easily make sure that these probabilities follow to the steady-state probabilities in an M/M/m/K system with a utilization coefficient q and queue length K:*

$$\sum_{j=0}^{n} \boldsymbol{P_{i,j,0}} = \frac{\boldsymbol{(q \cdot m)^i}}{\boldsymbol{i!}} \cdot \sum_{j=0}^{n} \boldsymbol{P_{0,j,0}},\ \ 0 < \boldsymbol{i \le m},$$

$$\sum_{j=0}^{n} \boldsymbol{P_{m,j,k}} = \frac{\boldsymbol{(q \cdot m)^m}}{\boldsymbol{m!}} \cdot \boldsymbol{q^k} \cdot \sum_{j=0}^{n} \boldsymbol{P_{0,j,0}},\ \ 0 < \boldsymbol{k \le K}.$$

3. Queuing Characteristics and Optimization

3.1. Queueing Characteristics

Probabilities (6)–(8) enable us to calculate the steady-state characteristics of multi-channel QS with stalling as probabilities of various QS states and numbers of queries in these states. Hence, one has at first to calculate the following, according to Lemmas A1 and A2: the $(m + K + 1) \times (n + 1)$ matrix of auxiliary functions $A_{i,j}$; $n + 1$ components vector h, used to find the steady-state probabilities according to Theorem A1; and the other characteristics. Note, the obtained explicit expressions are numerically stable; in addition, their complexity does not depend on the number of states, and is only linear with respect to the number of fast channels and stalling buffer length. Some characteristics are given in Table 1, using the following notation for simplicity:

$$p_2 = \frac{p_1}{(1-q) + p_1},$$

$$C_2 = \sum_{s=1}^{K} s \cdot q^s k \frac{q - (K+1) \cdot q^{K+1} + K \cdot q^{K+2}}{(1-q)^2}\ \ if\ q \ne 1,\ otherwise\ C_2 = \frac{K \cdot (K+1)}{2}.$$

$$S_2 = \sum_{s=1}^{M} s \cdot \rho^s = \frac{\rho - (M+1) \cdot \rho^{M+1} + M \cdot \rho^{M+2}}{(1-\rho)^2}\ \ if\ \rho \ne 1,\ \ otherwise\ \ S_2 = \frac{M \cdot (M+1)}{2} \tag{11}$$

A key characteristic of QS is the occupancy probability of all channels and stalling buffer only $P_{m,n,K}$, derived according to Theorem A1. For calculating the characteristics using (6)–(10), the well-known Newton binomial formula is applied:

$$\sum_{i=0}^{u} \frac{u!}{i! \cdot (u-i)!} q^i \cdot (1-q)^{u-i} = 1. \tag{12}$$

The explicitly given characteristics help us to study various effects; for instance, the effect of a query becoming stuck can be studied—the situation when the fast channels and the stalling buffer are free, but the slow channels are busy by service of queries that arrived before, etc.

Table 1. Characteristics of QS with stalling, $q > 0$, $\rho > 0$, $m \geq 1$, $n \geq 0$, $K \geq 0$, $M \geq 0$.

Characteristics	**Denotation**	**Formula**
The occupancy probability of all channels and stalling buffer	$P_{m,n,K}$	$\frac{q^K \cdot p_1}{h_0 \cdot (1+p_1 \cdot C_1)+q^K \cdot p_1 \cdot S_1}$
Downtime probability	$P_{0,0,0}$	$P_{m,n,K} \cdot \sum\limits_{s=0}^{n} \frac{(-1)^s \cdot h_s}{A_{m+K,s}}$
Occupancy probability of only all slow channels	$P_{0,n,0}$	$\frac{P_{m,n,K}}{A_{m+K,n}}$
Stuck probability P_{stuck}	$\sum\limits_{s=1}^{n} P_{0,s,0}$	$P_{m,n,K} \cdot \sum\limits_{s=1}^{n} \frac{(-1)^{s-1} \cdot h_s}{A_{m+K,s}}$
Number of stuck queries $\overline{N}_{stuck}$	$\sum\limits_{s=1}^{n} s \cdot P_{0,s,0}$	$h_1 \cdot \frac{P_{m,n,K}}{A_{m+K,1}}$
Probability of queries in slow channels P_{slow}	$\sum\limits_{i=0}^{m+K} \sum\limits_{s=1}^{n} P_{\min(i,m),s,\max(i-m,0)}$ $+P_{m,n,K} \cdot \sum\limits_{s=1}^{M} \rho^s$	$1 - P_{m,n,K} \cdot \sum\limits_{s=0}^{n} h_s \cdot (-1)^s \cdot \frac{\sum_{i=0}^{m+K} A_{i,s}}{A_{m+K,s}}$
Number of queries in slow channels $\overline{N}_{slow}$	$\sum\limits_{s=1}^{n} \left(s \sum\limits_{i=0}^{m+K} P_{\min(i,m),s,\max(i-m,0)} \right)$ $+P_{m,n,K} \cdot n \cdot \sum\limits_{s=1}^{M} \rho^s$	$P_{m,n,K} \cdot \left(\frac{h_1 \cdot \sum_{i=0}^{m+K} A_{i,1}}{A_{m+K,1}} + n \cdot S_1 \right)$
Probability of queries in fast channels P_{fast}	$\sum\limits_{s=0}^{n} \sum\limits_{i=1}^{m+K} P_{\min(i,m),s,\max(i-m,0)} + P_{m,n,K} \cdot \sum\limits_{s=1}^{M} \rho^s$	$1 - \frac{h_0 \cdot P_{m,n,K}}{q^K \cdot \frac{(q \cdot m)^m}{m!}}$
Number of queries in fast channels $\overline{N}_{fast}$	$\sum\limits_{i=1}^{m-1} i \cdot \sum\limits_{s=0}^{n} P_{i,s,0} + m \cdot \sum\limits_{k=0}^{K} \sum\limits_{s=0}^{n} P_{m,s,k}$ $+P_{m,n,K} \cdot m \cdot \sum\limits_{s=1}^{M} \rho^s$	$m - \frac{m \cdot h_0 \cdot P_{m,n,K}}{q^K \cdot p_2}$
Probability of stalling $P_{stalling}$	$\sum\limits_{k=1}^{K} \sum\limits_{s=0}^{n} P_{m,s,k} + P_{m,n,K} \cdot \sum\limits_{s=1}^{M} \rho^s$ $if\ K > 0\ otherwise\ 0$	$1 - \frac{h_0 \cdot P_{m,n,K} (1+p_1)}{p_1 \cdot q^K}$
Number of queries in stalling buffer $\overline{N}_{stalling}$	$\sum_{s=0}^{n} \sum_{k=1}^{K} k \cdot P_{m,s,k} + P_{m,n,K} \cdot K \cdot \sum_{s=1}^{M} \rho^s,$ $if\ K \geq 0,$ otherwise 0	$\mathrm{K} - \frac{h_0 \cdot P_{m,n,K} \left(\frac{K}{p_1} + C_2 \right)}{q^K}$
Probability of queue P_w	$P_{m,n,K} \cdot \sum\limits_{s=1}^{M} \rho^s$	$P_{m,n,K} \cdot S_1$
Number of queries in waiting line $\overline{N}_w$	$P_{m,n,K} \cdot \sum\limits_{s=1}^{M} (s \cdot \rho^s)$	$P_{n,m,K} S_2$
Probability of queries loss P_{loss}	$P_{m,n,K} \cdot \rho^M$	$P_{m,n,K} \cdot \rho^M$
Average number of queries in QS $\overline{N}$	$\overline{N}_{fast} + \overline{N}_{slow} + \overline{N}_{stalling} + \overline{N}_w =$ $\sum\limits_{i=0}^{m+K} \sum\limits_{s=0}^{n} (i+s) \cdot P_{min(i,m),s,\max(i-m,0)}$ $+P_{m,n,K} \cdot \sum\limits_{s=1}^{M} (s+m+K+n) \cdot \rho^s$	$q \cdot m + \frac{p_2 \cdot q}{(1-q)^2} - P_{m,n,K}$ $\cdot(\left(K + m \cdot (1-q) - \frac{p_m \cdot q}{1-q} \right) \cdot \left(\frac{h_0 \cdot q}{1-q} - S_1 \right) - h_1 \cdot \frac{\sum_{i=0}^{m+K} A_{i,1}}{A_{m+K,1}} +$ $\frac{h_0 \cdot q}{(1-q)^2} - S_1 \cdot n - S_2)$ if $q \neq 1,$ otherwise $m + P_{m,n,K} \cdot \left((K+1-m) \cdot h_0 + h_1 \cdot \frac{\sum_{i=0}^{m+K} A_{i,1}}{A_{m+K,1}} + S_2 + S_1 \cdot (K+n) \right)$

3.2. *Optimization of Stalling Buffer*

An average number of queries in the system $\overline{N}$ is the most important characteristic because it describes the whole system's capacity and the efficiency of the chosen queueing strategy. As obtained in the previous section, expression $\overline{N}$ can be explored analytically, e.g., one can differentiate it in relation to the parameters. Further, following the findings presented in [9], the technique of Chebyshev polynomials for the study of the asymptotic behaviour of a heterogeneous multi-channel OS with stalling is developed. Denote by

$$z_j = \frac{q + 1 + \frac{j}{m \cdot r} + \sqrt{\left(q + 1 + \frac{j}{m \cdot r} \right)^2 - 4 \cdot q}}{2}, t_j = \frac{q + 1 + \frac{j}{m \cdot r} - \sqrt{\left(q + 1 + \frac{j}{m \cdot r} \right)^2 - 4 \cdot q}}{2}, \quad 0 \leq j \leq n.$$

We can easily see that $z_j + t_j = q + 1 + \frac{j}{r}$, $z_j \cdot t_j = q$.

Moreover, it is easily to verify yet that if $0 < q < 1$, $r \geq 1$, then $z > 1$, $t < 1$. Assume for simplicity that $0 < q < 1$.

Note that ratios $A_{i,j}$ are polynomial functions of q and $\frac{1}{r}$.

Lema A3. *Ratios $A_{i,j}$ follow to recurrent presentation:*

$$A_{i,j} = \sum\nolimits_{s=0}^{i} b_{i,s} \cdot \frac{{z_j}^{i+1-s} - {t_j}^{i+1-s}}{z_j - t_j},\ 0 \leq i \leq m + K + 1,\ 0 \leq j \leq \mathrm{n},$$

where $b_{i,\,j} = 0$ if $i \leq 0$ or $s < 0$, except $b_{0,\,0} = 1$, and other coefficients follow to recursive relationship:

$$b_{i,s} = \frac{m \cdot b_{i-1,s}}{min(i,m)} - \frac{(m - min(i-1,m)) \cdot b_{i-1,s-1}}{min(i,m)} + \frac{m \cdot q \cdot (b_{i-1,s-2} - b_{i-2,s-2})}{min(i,m)},\ \text{if } i > 1, s \geq 0.$$

Proof of Lemma A3 is given in Appendix A.

Corollary. *We have*

$$\sum\nolimits_{i=0}^{m+K} \frac{A_{i,1}}{A_{m+K,1}} = \frac{z_1}{z_1 - 1} + o\left(q^K\right),\ \text{and}$$

$$\frac{A_{m+K+1,j}}{A_{m+K,j}} = z_j + O\left(q^K\right),\ 0 \leq j \leq n.$$

Theorem A2. *Assume that $0 < q < 1$, $r \geq 1$, $K \geq 0$, $M \geq 1$. Then, the optimal stalling buffer size, minimizing the average number of queries in QS, follows to approximate expressions:*

$$(i) \qquad K_0 = \frac{\frac{h1}{h0} \cdot \frac{z_1}{z_1 - 1} + \frac{n \cdot S_1 + S_2}{h0} - \frac{q}{(1-q)^2}}{\frac{q}{1-q} - \frac{S_1}{h0}} + \frac{q \cdot p_2}{1-q} - m \cdot (1-q) - \frac{1}{ln(q)} + O(q^K),$$

and

$$(ii) \quad K_0 = r \cdot m \cdot (1-q) - \frac{1+q}{1-q} + \frac{p_m \cdot q}{(1-q)^2} - m \cdot (1-q) - \frac{1}{ln(q)} + O\left(\frac{1}{r}\right).$$

where h h0 and h1 are the following:

$$h0 = \sum\nolimits_{s=1}^{n} \frac{n!}{s! \cdot (n-s)!} \prod\nolimits_{j=1}^{s} \left(\frac{z_j}{q} - 1\right) + 1,$$

$$h1 = \sum\nolimits_{s=2}^{n} \frac{n!}{s! \cdot (n-s)!} \prod\nolimits_{j=2}^{s} \left(\frac{z_j}{q} - 1\right) + n.$$

4. Applications

4.1. Modelling of Heterogeneous Server Cluster with Stalling Buffer

Compatibility problems of heterogeneous networks are solved by offering specialized data maintenance solutions, evaluating their combining cases, and analysing errors, links, etc. Optimization of the network node stalling buffer is important for combining networks of different capacities [31]. Web hosting companies are forced to change old slow servers to new fast servers, to serve an increasing flow of service. The optimal solution to this problem is using both groups—fast and slow servers connected to one heterogeneous server cluster using a stalling buffer (Figure 3).

Performance of servers' W is calculated by using disks and disk arrays IOPS (input/output operations per second) :

$$IOPS_{disk} = \frac{MeX_{rpm}}{\frac{MeX_{seek}}{1000} + \frac{MeX_{latency}}{1000}},$$

where MeX_{rpm}—disk rotational speed per minute; MeX_{seek}—disc seek time in milliseconds; $MeX_{latency}$—disc latency in milliseconds. Disk array (RAID) IOPS is calculated as follows:

$$IOPS_{raid} = IOPS_{disk} \cdot \left(N_{disks} - D_{rp}\right) \cdot D_{R\%} + \frac{IOPS_{disk} \cdot \left(N_{disks} - D_{wp}\right) \cdot D_{w\%}}{C},$$

where $\mathrm{N_{disks}}$—the number of disks in an array; $\mathrm{D_{rp}}$—reading parity for disks array; $\mathrm{D_{wp}}$—writing parity for disks array; $\mathrm{D_{R\%}}$—total percent of reading; $\mathrm{D_{w\%}}$—total percent of writing; C—disks array overhead. Disk array RAID5 in slow and fast servers is used. For RAID5: $\mathrm{D_{rp}}$ = 1, $\mathrm{D_{wp}}$ = 0, $\mathrm{D_{R\%}}$ = 70, $\mathrm{D_{w\%}}$ = 30, C = 4.

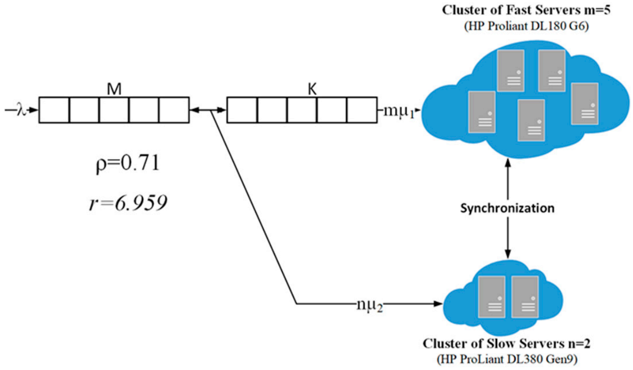

Figure 3. Heterogeneous servers' clusters with a stalling buffer.

Slow servers are HP Proliant DL180 G6 [36] width two Intel Xeon L5640@ 2.27 GHz processors (CPU Benchmarks 6358), with 128 GB DDR3 operating memory (read speed—13,887 MB/s; write speed—10,076 MB/s; $\mathrm{MiN_{r/wspeed}}$ = 11,982, $\mathrm{MiN_{type}}$ = 8); we used five disks ($\mathrm{N_{disks}}$ = 5) HP 3 TB 3.5″ LFF 6 G Dual Port SAS 7.2 K RPM ($\mathrm{MeX_{r/wdata}}$ = 97 Mb/s, data transferring rate $\mathrm{MeX_{rate}}$ = 6.0 Gb/sec, $\mathrm{IOPS_{disk}}$ = 57) connected to RAID5 ($\mathrm{IOPS_{raid}}$ = 18,098).

Fast servers are HP ProLiant DL380 Gen9 [37] width two Intel Xeon E5-2699 v4 @ 2.20 GHz processors (CPU Benchmarks 23200), with 128 GB DDR4 operating memory (read speed—22,075 MB/s; write speed—16,842 MB/s; $\mathrm{MiN_{r/wspeed}}$ = 19,458.5, $\mathrm{MiN_{type}}$ = 16); we used seven disks ($\mathrm{N_{disks}}$ = 7) HPE 2.4 TB SAS 12 G Enterprise 10 K LFF ($\mathrm{MeX_{r/wdata}}$ = 195 Mb/s, data transferring rate $\mathrm{MeX_{rate}}$ = 12 Gb/s, $\mathrm{IOPS_{disk}}$ = 133) connected to RAID5 ($\mathrm{IOPS_{raid}}$ = 62843). Servers were divided into two synchronized clusters. The interarrival traffic time was distributed under the exponential law with parameter λ and the length of service was also distributed under this law with parameters μ_1 and μ_2. This was the case in systems using M length waiting buffer and K length stalling buffer (see Figure 3).

Performance of servers in clusters are calculated using the following formula (Figure 4):

$$W = \frac{U \cdot MeX_{rate} \cdot MeX_{r/wdata} \cdot IOPS_{raid}}{MiN_{r/wspeed} \cdot MiN_{type}},$$

where U—summarized performance processors of server controller (CPU Benchmarks); $\mathrm{MeX_{rate}}$—number of input output transferring's to array of disks; $\mathrm{MeX_{r/wdata}}$—amount of data (bytes) transmitted per input/output operation; $\mathrm{MiN_{r/wspeed}}$—data transferring rate; $\mathrm{MiN_{type}}$—coefficient taking into account the type of memory [38]. Calculated coefficient of performance for fast servers $\mathrm{W_{fast}}$ = 1190 and for slow server $\mathrm{W_{slow}}$ = 171.

The characteristics of the QS cluster with five fast and two slow servers are computed, and the results are displayed in Table 2.

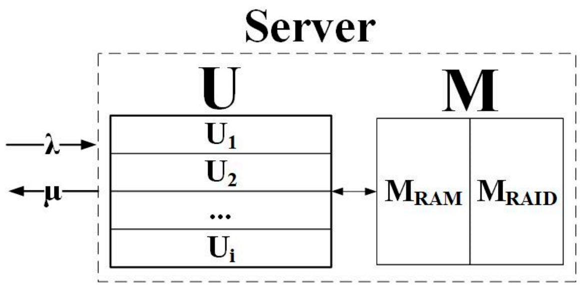

Figure 4. Server components by queuing model.

Table 2. Characteristics of the server's cluster with stalling.

Parameters	**Denotation**	**Formula**	**Value**
	Basic characteristics of clusters with stalling		
Intensity of interarrival	Λ	-	0.58 (Gbps)
Capacity of fast server	μ_1	-	0.146 (Gbps)
Capacity of slow server	μ_2	-	0.045 (Gbps)
Number of fast servers	m	-	5
Number of slow servers	n	-	2
Coefficient of QS utilization	ρ	$\frac{\lambda}{m \cdot \mu_1 + n \cdot \mu_2}$	0.71
Coefficient of utilization of fast server	Q	$\frac{\lambda}{m \cdot \mu_1} = \rho \cdot \left(1 + \frac{n}{m \cdot r}\right)$	0.751
Rate of capacity of fast and slow servers	R	$r = \frac{\mu_1}{\mu_2}$	6.959
	Calculated characteristics of clusters with stalling		
Characteristics	Denotation		Value
Occupancy probability of fast channel in QS without slow channel and stalling	P_0		0.464
The occupancy probability of the fast channels and stalling buffer and free slow channels	$P_{m,n,K}$		0.981
Downtime probability	$P_{0,0,0}$		0.0185
Stuck probability	P_{stuck}		0
Number of stuck queries	$\overline{N}_{stuck}$		0
Occupancy probability of all slow channel with free fast channel with stalling buffer	$P_{0,n,0}$		0
Probability of queries in slow channels	P_{slow}		0
Number of queries in slow channels	$\overline{N}_{slow}$		0
Probability of queries in fast channels	$\sum_{i=1}^{m+K} \sum_{s=0}^{n} p_{i,s} + p_{m+K,n} \cdot \frac{\rho - \rho^{M+1}}{1-\rho}$		0.9814
Number of queries in fast channels	P_{fast}		3.7561
Probability of stalling	$P_{stalling}$		0.3486
Number of queries in stalling buffer	$\overline{N}_{stalling}$		1.4014
Probability of queue	P_w		0
Number of queries in waiting line	$\overline{N}_w$		0
Probability of loss queries	$P_{m+K,n} \cdot \rho^M$		0
Average number of queries	$\overline{N}$		5.1575

4.2. Modelling of Harvesters and Chainsaws Work Productivity

Coordination of technologies, processes, or devices with different capacities arises very often. Let us consider the problem of coordinated usage of harvesters (Timberjack 1270D) and chainsaws in spruce stands using data of comparative work productivity analysis [39].

Harvester (μ_1) and chainsaw (μ_2) work productivity have been compared in Table 3, taking into account the volume of the tree trunk.

Table 3. Harvester and chainsaw work productivity comparison by volume of tree trunk [39].

Volume of Tree Trunk (m^3)	Harvester		Chainsaw	
	Service (Lumbering) Time (h/m^3)	Work Intensity (μ_1, m^3/h)	Service (Lumbering) time (h/m^3)	Work Intensity (μ_2, m^3/h)
0.1	0.111	9	0.872	1.2
0.2	0.047	21.2	0.71	1.4
0.3	0.035	28.4	0.641	1.6
0.4	0.03	33.4	0.599	1.7
0.5	0.027	37.4	0.571	1.8
0.6	0.025	40.6	0.549	1.8
0.7	0.023	43.3	0.532	1.9
0.8	0.022	45.6	0.518	1.9
0.9	0.021	47.7	0.507	2
1	0.02	49.6	0.497	2

Figure 5 shows how many harvesters and chainsaws will be needed for optimal systems work when the scope of forest harvesting is changed. We can see that, when using one harvester (m = 1) and a system load of 50 m^3 for optimal systems work, 10 chainsaws are needed (Figure 5, graph on the left—see the yellow point on the graph and on the chainsaw axis is showing a 10, and on load axis we can see a 50). In contrast, when using two harvesters (m = 2) and a system load of 100 m^3 for optimal systems work, 20 chainsaws will be needed (Figure 5, graph on the right—see the yellow point on the graph and on chainsaw axis is showing a 20, and on the load axis we can see a 100).

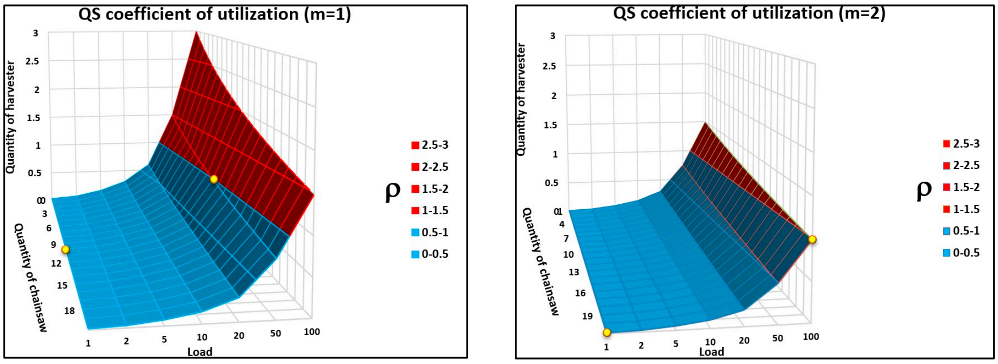

Figure 5. The harvesters and chainsaws quantity dependence from forest harvesting scope.

5. Conclusions

In the paper, the model of stalling in a QS with heterogeneous channels and stalling buffer is presented, deriving the explicit probabilities of steady states using Chebyshev polynomials of the second order. The obtained expressions are numerically stable; their complexity does not depend on the number of states, and they provide a way of analytically studying QS characteristics. Moreover, the existence of a finite optimal size of the stalling buffer is proved and a numerical approach for the optimization of buffer size is developed.

The results showed that the appropriate choice of stalling buffer size provides assistance in solving the slow channel or server problem. The asymptotic conditions of optimal query distribution in channels when the ratio of the capacities of the fast and slow servers increases are also established. An application of the developed model in heterogeneous server clusters for work productivity modelling in the context of forest harvesting is also discussed herein.

This investigation of the developed model enables us to conclude that stalling is a universal solution that can be implemented into any heterogeneous system with any number of heterogeneous channels or servers.

Author Contributions: Conceptualization, L.S. and R.M.; methodology, L.S.; software, L.K.; validation, L.K. and R.M.; resources, L.K.; writing—original draft preparation, L.S. and L.K.; writing—review and editing, L.S., L.K. and R.M.; visualization, L.K.; supervision, L.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in article.

Acknowledgments: We take an opportunity to kindly thank the anonymous reviewers for their valuable remarks.

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

Appendix A

Proof of Lemma A1. Let for $0 \le j \le n$, denote $A_{-1,j} = 1$. At first, note the easily verifiable identity:

$$\sum\nolimits_{s=j}^{n} \frac{(s-j)\cdot s!}{j!\cdot(s-j)!} P_{min(i-1,m),s,max(i-m,0)} = \sum\nolimits_{s=j}^{n-1} \frac{\cdot(s+1)\cdot s!}{j!\cdot(s-j)!} P_{min(i-1,m),s+1,max(i-m,0)}\cdot \tag{13}$$

We can derive the relation between coefficients $R_{i,j}$, whenever $0 \le j \le n$, $1 \le i \le m+K$. Using the Equality (A1), the respective equations of steady-state probabilities in (1) and identities (3), we have

$$\begin{aligned}
(\lambda + i\cdot\mu_1 + j\cdot\mu_2)\cdot R_{i,j} &= \sum_{s=j}^{n} \tfrac{s!}{j!\cdot(s-j)!}\cdot(\lambda + i\cdot\mu_1 + j\cdot\mu_2)\cdot P_{min(i,m),s,max(i-m,0)} \\
&= (\lambda + i\cdot\mu_1 + n\cdot\mu_2)\cdot \tfrac{n!}{j!\cdot(n-j)!}\cdot P_{min(i,m),n,max(i-m,0)} \\
&\quad + \sum_{s=j}^{n-1} \tfrac{s!}{j!\cdot(s-j)!}(\lambda + i\cdot\mu_1 + s\cdot\mu_2)\cdot P_{min(i,m),s,max(i-m,0)} \\
&\quad - \sum_{s=j}^{n} \tfrac{(s-j)\cdot s!}{j!\cdot(s-j)!}\cdot\mu_2\cdot P_{min(i,m),s,max(i-m,0)} \\
&= \Big(P_{min(i+1,m),n,max(i+1-m,0)}\cdot max(i+1,m)\cdot\mu_1 \\
&\quad + P_{min(i-1,m),n,max(i-m-1,0)}\cdot\lambda\Big)\cdot\tfrac{n!}{j!\cdot(n-j)!} \\
&\quad + \sum_{s=j}^{n-1} \tfrac{s!}{j!\cdot(s-j)!}\cdot\Big(P_{min(i+1,m),s,max(i+1-m,0)}\cdot(i+1)\cdot\mu_1 \\
&\quad + P_{min(i+1,m),s,max(i-1-m,0)}\cdot(s+1)\cdot\mu_2 + P_{min(i-1,m),s,max(i-1-m,0)}\cdot\lambda\Big) \\
&\quad - \sum_{s=j}^{n} \tfrac{(s-j)\cdot s!}{j!\cdot(s-j)!}\cdot\mu_2\cdot P_{min(i,m),s,max(i-m,0)} \\
&= \sum_{s=j}^{n} \tfrac{s!}{j!\cdot(s-j)!}\cdot P_{min(i+1,m),s,max(i+1-m,0)}\cdot max(i+1,m)\cdot\mu_1 \\
&\quad + \sum_{s=j}^{n} \tfrac{s!}{j!\cdot(s-j)!}\cdot P_{min(i+1,m),s,max(i+1-m,0)}\cdot max(i+1,m)\cdot\mu_1 \\
&\quad + \sum_{s=j}^{n} \tfrac{s!}{j!\cdot(s-j)!}\cdot P_{min(i-1,m),s,max(i-1-m,0)}\cdot\lambda \\
&= max(i+1,\ m)\cdot\mu_1\cdot R_{i+1,j} + \lambda\cdot R_{i-1,j}
\end{aligned}$$

Established relationship and definition (3) imply the Lemma.□

Proof of Lemma A2. Indeed, $h_n = 1$. Hence, according to (5),

$$h_j = \frac{R_{m+K,j}}{R_{m+K,n}} = \frac{\sum_{s=j}^{n} \frac{s!}{j!\cdot(s-j)!} P_{m,s,K}}{P_{m,n,K}},\ 0 \le j \le n.$$

After elementary steps by virtue of latter relationship and Lemma 1, one can make sure that

$$\begin{gathered}
h_j \cdot \left(\frac{A_{m+K+1,j}}{A_{m+K,j}} - 1\right) + \frac{n!}{(j-1)!\cdot(n-j+1)!} \\
= \left(\frac{j}{r\cdot m\cdot q} + \frac{1}{q}\right) \cdot \sum_{s=j}^{n} \frac{s!}{j!\cdot(s-j)!} \cdot \frac{P_{m,s,K}}{P_{m,n,K}} - \sum_{s=j}^{n-1} \frac{s!}{j!\cdot(s-j)!} \cdot \frac{P_{m,s,K-1}}{P_{m,n,K}} \\
- \frac{n!}{j!\cdot(n-j)!} \cdot \frac{P_{m,n,K-1}}{P_{m,n,K}} + \frac{n!}{(j-1)!\cdot(n-j+1)!} \\
= \left(\frac{j}{r\cdot m\cdot q} + \frac{1}{q}\right) \cdot \sum_{s=j}^{n} \frac{s!}{j!\cdot(s-j)!} \cdot \frac{P_{m,s,K}}{P_{m,n,K}} - \sum_{s=j}^{n-1} \frac{s!}{j!\cdot(s-j)!} \cdot \frac{P_{m,s,K-1}}{P_{m,n,K}} \\
- \frac{n!}{j!\cdot(n-j)!} \cdot \frac{P_{m,n,K-1}}{P_{m,n,K}} + \frac{n!}{(j-1)!\cdot(n-j+1)!} \\
= \left(\frac{j}{r\cdot m\cdot q} + \frac{1}{q}\right) \cdot \sum_{s=j}^{n} \frac{s!}{j!\cdot(s-j)!} \cdot \frac{P_{m,s,K}}{P_{m,n,K}} \\
- \sum_{s=j}^{n-1} \frac{s!}{j!\cdot(s-j)!} \\
\cdot \frac{P_{m,s,K}\cdot\left(1+\frac{1}{q}+\frac{s}{r\cdot m\cdot q}\right) - P_{m,s-1,K} - P_{m,s+1,K}\cdot\frac{s+1}{r\cdot m\cdot q}}{P_{m,n,K}} \\
- \frac{n!}{j!\cdot(n-j)!} \cdot \frac{P_{m,n,K}\cdot\left(\frac{1}{q}+\frac{n}{r\cdot m\cdot q}\right) - P_{m,n-1,K}}{P_{m,n,K}} \\
+ \frac{n!}{(j-1)!\cdot(n-j+1)!} \\
= \sum_{s=j}^{n-1} \frac{s!}{j!\cdot(s-j)!} \cdot \frac{P_{m,s-1,K} - P_{m,s,K}}{P_{m,n,K}} - \frac{n!}{j!\cdot(n-j)!} \cdot \frac{P_{m,n,K-1}}{P_{m,n,K}} \\
+ \frac{n!}{(j-1)!\cdot(n-j+1)!} \\
= \sum_{s=j}^{n-1} \frac{s!}{(j-1)!\cdot(s-j+1)!} \cdot \frac{P_{m,s,K}}{P_{m,n,K}} + \frac{n!}{(j-1)!\cdot(n-j+1)!} = h_{J-1}\square
\end{gathered}$$

Proof of Theorem A1. Let us consider Equation (1), for $0 \le i < m + \mathrm{K}, 1 \le j \le n$, or for $i = m + \mathrm{K}, 1 \le \mathrm{j} < n$, and assume that the waiting line is empty. Respectively, define, $P_{-1,j,0} = 0$, and $P_{i,-1,k} = 0$, $P_{i,n+1,k} = 0$. Then, according to (6), we have that

$$\begin{gathered}
\left(q + min\left(\frac{i}{m}, 1\right) + \frac{j}{r\cdot m}\right) \cdot P_{min(i,m),j,max(i-m,0)} \\
= \left(q + min\left(\frac{i}{m}, 1\right) + \frac{j}{r\cdot m}\right) \cdot P_{m,n,K} \cdot \sum_{s=j}^{n} h_s \cdot \frac{A_{i,s}}{A_{m+K,s}} \frac{(-1)^{s-j}\cdot s!}{j!\cdot(s-j)!} \cdot \\
= P_{m,n,K} \cdot \sum_{s=j}^{n} h_s \cdot \frac{A_{i,s}}{A_{m+K,s}} \frac{(j-s)}{r\cdot m} \frac{(-1)^{s-j}\cdot s!}{j!\cdot(s-j)!} + P_{m,n,K} \\
\cdot \sum_{s=j}^{n} h_s \cdot \left(q + min\left(\frac{i}{m}, 1\right) + \frac{s}{r\cdot m}\right) \cdot \frac{A_{i,s}}{A_{m+K,s}} \frac{(-1)^{s-j}\cdot s!}{j!\cdot(s-j)!}
\end{gathered} \tag{14}$$

In addition, in virtue of Lemma 1,

$$\begin{gathered}
\sum_{s=j}^{n} h_s \cdot \left(q + min\left(\frac{i}{m}, 1\right) + \frac{s}{r\cdot m}\right) \cdot \frac{A_{i,s}}{A_{m+K,s}} \frac{(-1)^{s-j}\cdot s!}{j!\cdot(s-j)!} \\
= q \cdot \sum_{s=j}^{n} h_s \cdot \frac{A_{i-1,s}}{A_{m+K,s}} \frac{(-1)^{s-j}\cdot s!}{j!\cdot(s-j)!} + \sum_{s=j}^{n} h_s \cdot \frac{A_{i+1,s}}{A_{m+K,s}} \frac{(-1)^{s-j}\cdot s!}{j!\cdot(s-j)!} \\
= \left(P_{min(i-1,m),j,max(i-1-m,0)} \cdot q + P_{min(i+1,m),j,max(i+1-m,0)}\right) / P_{m,n,K}
\end{gathered} \tag{15}$$

Thus, (A2) and (A3) imply that relations among steady-state probabilities in (6) satisfy Equation (1), if $0 \le i < m + \mathrm{K}$,$1 \le j \le n$, or $i = m + \mathrm{K}$, $1 \le \mathrm{j} < n$, and the waiting line is empty.

Next, it is easy to see that Equation (7) is equivalent to recursive relation $P_{m,n,K+k} = \rho \cdot P_{m,n,K+k-1}$, which easily follow from respective Equation (1).

The expression of probability $P_{m,n,K}$ for state in which both channels and stalling buffer are only full, is derived after elementary manipulations using Equation (2) and following it remark, Newton binomial theorem and convinced that $\sum_{j=0}^{n} P_{0,j,0} = \frac{P_{m,n,K}}{\frac{(q \cdot m)^m}{m!}} \cdot q^K$. □

Proof of Lemma A3. We can easily see that

$$b_{i,0} = \prod\nolimits_{t=1}^{i} max\left(\frac{m}{t}, 1\right),\ b_{i,1} = \prod\nolimits_{t=1}^{i} max\left(\frac{m}{t}, 1\right),\ i \ge 1, b_{i,j+1} = 0.$$

Indeed, Lemma is true at $i = 0,\ 0 \le j \le n$. Now the relationship for $0 < i < m$, $0 \le j \le n$, follows by virtue of respective equality of Lemma 1:

$$A_{i+1,j,0} = \frac{\left(q \cdot m + \frac{j}{r} + i\right) \cdot A_{i,j,0}}{i+1} - \frac{m \cdot q \cdot A_{i-1,j,0}}{i+1} = \frac{\left(q \cdot m + \frac{j}{r} + i\right) \cdot \sum_{s=0}^{i} b_{i,s} \cdot \frac{{z_j}^{i+1-s} - {t_j}^{i+1-s}}{z_j - t_j}}{i+1} -$$

$$\frac{m \cdot q \cdot \sum_{s=0}^{i-1} b_{i-1,s} \cdot \frac{{z_j}^{i-s} - {t_j}^{i-s}}{z_j - t_j}}{i+1} = \frac{\left(m \cdot \left(z_j + t_j\right) + i - m\right) \cdot \sum\limits_{s=0}^{i} b_{i,s} \cdot \frac{z_j^{i+1-s} - t_j^{i+1-s}}{z_j - t_j}}{i+1} - \frac{m \cdot q \cdot \sum_{s=0}^{i-1} b_{i-1,s} \cdot \frac{z_j^{i-s} - t_j^{i-s}}{z_j - t_j}}{i+1} =$$

$$\frac{m \cdot \left(z_j + t_j\right) \cdot \sum_{s=0}^{i} b_{i,s} \cdot \frac{{z_j}^{i+1-s} - {t_j}^{i+1-s}}{z_j - t_j}}{i+1} - \frac{(m-i) \cdot \sum_{s=0}^{i} b_{i,s} \cdot \frac{{z_j}^{i+1-s} - {t_j}^{i+1-s}}{z_j - t_j}}{i+1} - \frac{m \cdot q \cdot \sum_{s=0}^{i-1} b_{i-1,s} \cdot \frac{{z_j}^{i-s} - {t_j}^{i-s}}{z_j - t_j}}{i+1} =$$

$$\frac{m \cdot \sum_{s=0}^{i} b_{i,s} \cdot \frac{{z_j}^{i+2-s} - {t_j}^{i+2-s}}{z_j - t_j}}{i+1} - \frac{(m-i) \cdot \sum_{s=0}^{i} b_{i,s} \cdot \frac{{z_j}^{i+1-s} - {t_j}^{i+1-s}}{z_j - t_j}}{i+1} + \frac{m \cdot q \cdot \sum_{s=0}^{i-1} \left(b_{1,s} - b_{i-1,s}\right) \cdot \frac{{z_j}^{i-s} - {t_j}^{i-s}}{z_j - t_j}}{i+1} =$$

$$\frac{m \cdot \sum_{s=0}^{i+1} b_{i,s} \cdot \frac{{z_j}^{i+2-s} - {t_j}^{i+2-s}}{z_j - t_j}}{i+1} - \frac{(m-i) \cdot \sum_{s=1}^{i+1} b_{i,s-1} \cdot \frac{{z_j}^{i+2-s} - {t_j}^{i+1-s}}{z_j - t_j}}{i+1} +$$

$$\frac{m \cdot q \cdot \sum_{s=2}^{i+1} \left(b_{1,s-2} - b_{i-1,s-2}\right) \cdot \frac{{z_j}^{i+2-s} - {t_j}^{i+2-s}}{z_j - t_j}}{i+1} = \frac{m \cdot \sum_{s=0}^{i+1} b_{i,s} \cdot \frac{{z_j}^{i+2-s} - {t_j}^{i+2-s}}{z_j - t_j}}{i+1} -$$

$$\frac{(m-i) \cdot \sum_{s=0}^{i+1} b_{i,s-1} \cdot \frac{{z_j}^{i+2-s} - {t_j}^{i+1-s}}{z_j - t_j}}{i+1} + \frac{m \cdot q \cdot \sum_{s=0}^{i+1} \left(b_{1,s-2} - b_{i-1,s-2}\right) \cdot \frac{{z_j}^{i+2-s} - {t_j}^{i+2-s}}{z_j - t_j}}{i+1} =$$

$$\frac{\sum_{s=0}^{i+1} \left(m \cdot b_{i,s} - (m-i) \cdot b_{i,s-1} + m \cdot q \cdot \left(b_{1,s-2} - b_{i-1,s-2}\right)\right) \cdot \frac{{z_j}^{i+2-s} - {t_j}^{i+2-s}}{z_j - t_j}}{i+1} = \sum_{s=0}^{i+1} b_{i+1,s} \cdot \frac{{z_j}^{i+2-s} - {t_j}^{i+2-s}}{z_j - t_j}. \square$$

Proof of Theorem A2. The expected number of queries in the QS by means of Lemma A3 is approximated as follows:

$$\bar{N} = q \cdot m + \frac{p_m \cdot q}{\left(1 - q\right)^2} - q^K \cdot p_m$$

$$\frac{\left(\left(K + m \cdot (1 - q) - \frac{p_m \cdot q}{1-q}\right) \cdot \left(\frac{h0 \cdot q}{1-q} - S_1\right) - \frac{h1 \cdot z_1}{z_1 - 1} + \frac{h0 \cdot q}{(1-q)^2} - S_1 \cdot n - S_2\right)}{h0 + q^K \cdot p_m \cdot \left(S_1 - \frac{h0 \cdot q}{1-q}\right)} + O\left(q^{2K}\right).$$

Then, differentiating it with respect to parameter *K*, equating the obtained derivative to zero and solving this equation, the approximate expression (*i*) of optimal stalling buffer size follows.

The approximate expression (*ii*) of optimal stalling buffer size is obtained in the same way using the expansion with the appropriated number of terms:

$$z_j = 1 + \frac{j}{r \cdot m \cdot (1 - q)} - \frac{q \cdot j}{(r \cdot m \cdot (1 - q))^2} + O\left(\frac{1}{r^3}\right). \square$$

References

1. Kleinrock, L. *Queueing Systems, Volume 1: Theory*; Wiley: New York, NY, USA, 2009.
2. Larsen, R.L. Control of Multiple Exponential Servers with Application to Computer Systems. Ph.D. Thesis, University of Maryland at College Park, College Park, MD, USA, 1981.
3. Rubinovitch, M. The slow server problem: A queue with stalling. *J. Appl. Probab.* **1985**, *22*, 879–892. [CrossRef]
4. De Vericourt, F.; Zhou, Y.P. On the incomplete results for the heterogeneous server problem. *Queueing Syst.* **2006**, *52*, 189–191. [CrossRef]
5. Goswami, V.; Samanta, S.K. Discrete-time bulk-service queue with two heterogeneous servers. *Comput. Ind. Eng.* **2009**, *56*, 1348–1356. [CrossRef]
6. Rykov, V.V.E.; Efrosinin, D.V. On the slow server problem. *Autom. Remote Control* **2009**, *70*, 2013–2023. [CrossRef]
7. Efrosinin, D.; Rykov, V. Heuristic solution for the optimal thresholds in a controllable multi-server heterogeneous queueing system without preemption. In Proceedings of the International Conference on Distributed Computer and Communication Networks, Moscow, Russia, 19–22 October 2015; pp. 238–252.
8. Shi, C.; Li, Y.; Zhang, J.; Sun, Y.; Philip, S.Y. A survey of heterogeneous information network analysis. *IEEE Trans. Knowl. Data Eng.* **2017**, *29*, 17–37. [CrossRef]
9. Kaklauskas, L.; Sakalauskas, L.; Denisovas, V. Stalling for solving slow server problem. *RAIRO Oper. Res.* **2019**, *53*, 1097–1107. [CrossRef]
10. Kalyanaraman, R.; Hellen, A. M/M/2 Heterogeneous Server Queue with Variant Breakdown and with Discouraged Arrivals. *Math. Stat. Eng. Appl.* **2022**, *71*, 1447–1458.
11. Kalyanaraman, R.; Hellen, A. A Two Heterogeneous Server Queue with Variant Breakdown and with Discouraged Arrivals, Balking. *Neuro Quantol.* **2022**, *20*, 4142.
12. Elsayed, E.A. Multichannel queueing systems with ordered entry and finite source. *Comput. Oper. Res.* **1983**, *10*, 213–222. [CrossRef]
13. Rana, P.S. A discrete time queueing problem with S heterogeneous groups of channels. *Microelectron. Reliab.* **1985**, *25*, 455–459.
14. Rana, P.S. A queueing problem with random memory arrivals and heterogeneous servers. *Microelectron. Reliab.* **1985**, *25*, 645–650.
15. Rosberg, Z.; Makowski, A.M. Optimal routing to parallel heterogeneous servers-small arrival rates. *IEEE Trans. Autom. Control* **1990**, *35*, 789–796. [CrossRef]
16. El-Taha, M.; Stidham, S., Jr. Deterministic analysis of queueing systems with heterogeneous servers. *Theor. Comput. Sci.* **1992**, *106*, 243–264. [CrossRef]
17. Chirkova, J.; Mazalov, V.; Morozov, E. Equilibrium in a queueing system with retrials. *Mathematics* **2022**, *10*, 428. [CrossRef]
18. Cabral, F.B. The slow server problem for uninformed customers. *Queueing Syst.* **2005**, *50*, 353–370. [CrossRef]
19. Rykov, V.V. Monotone control of queueing systems with heterogeneous servers. *Queueing Syst.* **2001**, *37*, 391–403. [CrossRef]
20. Efrosinin, D.; Sztrik, J. Performance analysis of a two-server heterogeneous retrial queue with threshold policy. *Qual. Technol. Quant. Manag.* **2011**, *8*, 211–236. [CrossRef]
21. Efrosinin, D.; Sztrik, J. Optimal control of a two-server heterogeneous queueing system with breakdowns and constant retrials. In Proceedings of the International Conference on Information Technologies and Mathematical Modelling, Katun, Russia, 12–16 September 2016; pp. 57–72.
22. Pankratova, E.; Moiseeva, S.; Farkhadov, M. Infinite-Server Resource Queueing Systems with Different Types of Markov-Modulated Poisson Process and Renewal Arrivals. *Mathematics* **2022**, *10*, 2962. [CrossRef]
23. Dudin, A.; Dudina, O.; Dudin, S.; Gaidamaka, Y. Self-service system with rating dependent arrivals. *Mathematics* **2022**, *10*, 297. [CrossRef]
24. Efrosinin, D.; Stepanova, N. Estimation of the optimal threshold policy in a queue with heterogeneous servers using a heuristic solution and artificial neural networks. *Mathematics* **2021**, *9*, 1267. [CrossRef]
25. Efrosinin, D.; Stepanova, N.; Sztrik, J. Algorithmic analysis of finite-source multi-server heterogeneous queueing systems. *Mathematics* **2021**, *9*, 2624. [CrossRef]
26. Hetherington, T.H.; Rogers, T.G.; Hsu, L.; O'Connor, M.; Aamodt, T.M. Characterizing and evaluating a key-value store application on heterogeneous CPU-GPU systems. In Proceedings of the IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), New Brunswick, NJ, USA, 1–3 April 2012; pp. 88–98.
27. Ciznicki, M.; Kierzynka, M.; Kopta, P.; Kurowski, K.; Gepner, P. Benchmarking JPEG 2000 implementations on modern CPU and GPU architectures. *J. Comput. Sci.* **2014**, *5*, 90–98. [CrossRef]
28. Kadjo, D.; Ayoub, R.; Kishinevsky, M.; Gratz, P.V. A control-theoretic approach for energy efficient CPU-GPU subsystem in mobile platforms. In Proceedings of the 52nd Annual Design Automation Conference, San Francisco, CA, USA, 8–12 June 2015; p. 62.
29. Choi, W.; Duraisamy, K.; Kim, R.G.; Doppa, J.R.; Pande, P.P.; Marculescu, D.; Marculescu, R. On-Chip Communication Network for Efficient Training of Deep Convolutional Networks on Heterogeneous Manycore Systems. *IEEE Trans. Comput.* **2018**, *67*, 672–686. [CrossRef]
30. Almusaddar, G.; Naghibijouybari, H. Exploiting Parallel Memory Write Requests for Covert Channel Attacks in Integrated CPU-GPU Systems. *arXiv* **2023**, arXiv:2307.16123.
31. Jin, X.; Min, G. Analytical queue length distributions of GPS systems with long range dependent service capacity. *Simul. Model. Pract. Theory* **2009**, *17*, 1500–1510. [CrossRef]

32. Lu, J.; Ma, M. Cross-layer QoS support framework and holistic opportunistic scheduling for QoS in single carrier WiMAX system. *J. Netw. Comput. Appl.* **2011**, *34*, 765–773. [CrossRef]
33. Okorogu, V.N.; Okafor, C.S. Improving Traffic Management in a Data Switched Network Using an Adaptive Discrete Time Markov Modulated Poisson Process. *Eur. J. Sci. Innov. Technol.* **2023**, *3*, 542–562.
34. Hruby, T.; Bos, H.; Tanenbaum, A.S. When Slower Is Faster: On Heterogeneous Multicores for Reliable Systems. In Proceedings of the USENIX Annual Technical Conference, San Jose, CA, USA, 26–28 June 2013; pp. 255–266.
35. Wang, J.; Zhang, Q.; Rong, H.; Xu, G.H.; Kim, M. Leveraging Hardware Probes and Optimizations for Accelerating Fuzz Testing of Heterogeneous Applications. In Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, San Jose, CA, USA, 3–9 November 2023; pp. 1101–1113.
36. Hewlet Packard Enterprice. HPE ProLiant DL380 Gen9 Server. 2019. Available online: https://h20195.www2.hpe.com/v2/getpdf.aspx/c04346247.pdf (accessed on 15 November 2022).
37. Hewlet Packard Enterprice Support Center. HPE ProLiant DL180 G6 Server—Overview. 2019. Available online: https://support.hpe.com/hpsc/doc/public/display?docId=emr_na-c01709693 (accessed on 15 November 2022).
38. Kovalenko, S.M.; Kazancceva, L.V.; Supronenko, D.V. Integrated Indicator of the Performance of Servers. The Online Scientific and Methodological Journal "Herald of MSTU MIREA". 2014. Available online: https://rtj.mirea.ru/upload/medialibrary/3df/10-kovalenko_kazantsev.pdf (accessed on 15 November 2022).
39. Mizaras, S.; Kombaris, N. Economic assessment of multifunctional forests in conservation land areas of Varniai regional park. *Agric. Sci.* **2012**, *19*, 98–105.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.