RAIRO-Oper. Res. 53 (2019) 1097–1107
https://doi.org/10.1051/ro/2018056

RAIRO Operations Research
www.rairo-ro.org

STALLING FOR SOLVING SLOW SERVER PROBLEM

Liudvikas Kaklauskas[1,*], Leonidas Sakalauskas[2] and Vitalijus Denisovas[3]

Abstract. The model of stalling in queueing system (QS) with two heterogeneous severs is considered, the probabilities of steady states by means of Tchebyshev polynomials of second order are derived. The obtained expressions are stable numerically, their complexity does not depend on the number of states, and they enable us to study QS characteristics analytically. Optimization of a stalling buffer is considered as well and it was shown that stalling helps us to solve the slow server problems under an appropriate choice of stalling buffer size, making a slow server usable under various values of system load. Asymptotic conditions of optimal query distribution in servers are established, when the ratio of capacities of fast and slow channels is increasing. Application of the model developed in computer networks is discussed as well.

Mathematics Subject Classification. 60K25, 90B22.

Received November 19, 2017. Accepted July 7, 2018.

1. Introduction

Queuing systems (QS) are widely used in business and technology. The known queueing strategies are developed for systems with equal capacity channels, however, in practice, systems have heterogeneous (different capacity) channels or servers [11, 21, 24]. For instance, at the shop every tellers efficiency depends on his experience in this field, therefore his service speed can be several times higher than that of the new teller who is working for several days. The similar situation is encountered in computers where computation is done by CPU and GPU. Since their capacity may differ up to several ten times, therefore it is necessary to schedule them at work optimally [5, 12]. Many practical scenarios, for example, communications network supporting communication channels of different transmission rates, multiprogramming computer system, etc, are modeled by systems or networks of heterogeneous servers. Note, in a heterogeneous environment, resources are autonomous, distributed, dense, and dynamic, hence they should be effectively scheduled so that maximum utilization of the resources is possible [11]. If the capacities of processors or servers vary with a little difference, then the queueing discipline in a system might be the First Come First Served (FCFS), and the query goes straight to the free channel or waits in a queue until some channel clears up. It is noted that under a high difference of capacities, this service discipline slows up the system work [21], moreover, in some situations it is better to discharge a slow channel [1]. Therefore, it was proposed to install a stalling buffer, as sometimes it is more efficient to wait until

Keywords. Heterogeneous severs, stalling buffer, queueing system.

[1] Shiauliai State College, Shiauliai, Lithuania.
[2] Shiauliai University, Shiauliai, Lithuania.
[3] Klaipeda University, Klaipeda, Lithuania.
*Corresponding author: liukak@gmail.com

Article published by EDP Sciences

© EDP Sciences, ROADEF, SMAI 2019

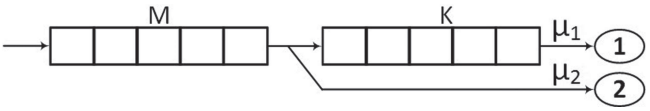

FIGURE 1. Scheme of QS with stalling buffer.

the fast channel gets free even though the slow channel becomes free. Larsen introduced this problem [15] by conjecturing optimal routing using threshold policy for particular server and stalling buffer. In 1983 Larsen and Agrawala [16] proposed a simple approximation to find the optimal threshold value for the slower server. Lin and Kumar showed that threshold value for fast server is zero and for slow server depends on the queue length [17]. Optimality of a threshold policy for two heterogeneouse servers by using coupling arguments and value iteration was proved by Warland and later by Koole [14, 17]. The systems with stalling were discussed in [24], he proposed an algorithm to obtain the optimal threshold value for the slow server [24]. In 1988 Viniotis and Ephremides [27] extended the validity of some results of the optimal control of two-server queuing models with service times of unequal distribution, operating in continuous or discrete time. Xu [28] proved the optimality of a threshold policy for each server in M/M/2 system, as well as the existence of a threshold queue length for denying admission to the system. Rykov [22] attempted to generalize the slow server problem of a threshold policy optimality to the case of more than two heterogeneous servers, but Zhou [6] showed that this proofs are incomplete. Rosberg and Makowski in 1990 analyzed set of heterogeneous and exponential servers [23]. In 2016 Efrosinin and Rykov analyzed M/M/K queue and showed, that the optimal threshold levels may depend also on the states of slower servers although this influence is very negligible [9]. The optimality of a threshold control policy for two heterogeneous servers queueing system is analyzed [19]. Efrosinin and Rykov in 2015 analyzed queuing system with K heterogeneous servers by using a heuristic method (he used means of Howard iterations with restrictions) to derive expressions for the optimal threshold levels in explicit form as functions of system parameters. As separate case the steady state probabilities in heterogeneous two servers system have been derived as well, but no analysis of system characteristics or asymptotic done [8]. The problem of stalling in systems with two or more servers has not been solved properly yet.

Recently the relevance of research of systems with heterogeneous servers has been increased by creating multiprocessor systems, combining CPU and GPU processors [12, 13] as well as considering queueing systems with very heterogeneous processors in networks [7]. Since their capacity may vary up to several ten times, there is an actual problem to develop systems, in which the ratio of capacity of used processors is increasing and to derive the appropriated asymptotic expressions.

This paper studies a queueing system with two heterogeneous servers equipped by a stalling buffer and a finite waiting line. Since analytical study of systems with heterogeneous channels is rather complicated, in the paper the asymptotic analysis of two servers system is presented.

2. QS WITH A STALLING BUFFER

The analyzed QS consists of two heterogeneous channels, waiting line and one stalling buffer (Fig. 1). Assume that the interarrival time at a system is distributed under Poisson's Law with a parameter λ and the length of service is also distributed under this law with parameters μ_1 and μ_2 (respectively for a fast and low channel, *i.e.*, $\mu_1 > \mu_2$). Assume that the queries are served according to FCFS discipline and explore the following order of service with stalling. If the query after being released into the system finds a free fast channel, it is served immediately, otherwise it goes to a stalling buffer of K length, where it waits until the efficient server gets free. One query is served only by one server without a break. If all places in the stalling buffer are occupied, the arrived query transfers to a slow server. If the slow server is occupied as well, the query waits in a queue at waiting buffer of length M. If all the places in the waiting and stalling buffers are occupied, the query is rejected

and is deemed lost. Of course, this strategy is suitable only when the coefficient of QS utilization $\rho = \frac{\lambda}{\mu_1+\mu_2}$ is less than 1.

3. QS state probabilities

Considering a two-server queuing system, state graph is drawn, in which the vertices represent the states, connected by arrows representing transitions with a non-null probability from one state to another. Denote the states as follows:

$P_{0,0,0}$, a system is free;

$P_{0,1,0}$, only a slow server is occupied;

$P_{1,0,0}$, only a fast server is occupied;

$P_{1,0,k}$, fast server is occupied, a slow server is free, k queries are stalled in the stalling buffer; $P_{1,0,K}$, a fast channel is occupied, a slow channel is free, a stalling buffer is full;

$P_{1,1,k}$, all channels are occupied, k queries are stalled in the stalling buffer;

$P_{1,1,K}$, all channels are occupied, stalling buffer is full;

$P_{1,1,K+k}$– all channels and stalling buffer are occupied, k queries are waiting in a queue;

$P_{1,1,K+M}$, all channels, stalling buffer as well as waiting line are full, next arrived query will be lost.

State graph of the considered system is provided in Figure 2.

Given that this corresponds to a finite Quasi-Birth-Death process [18]. Thus, one can derive steady state equations according to the steady state graph:

$$\begin{aligned}
P_{0,0,0}\cdot\lambda &= P_{1,0,0}\cdot\mu_1 + P_{0,1,0}\cdot\mu_2,\\
P_{1,0,0}\cdot(\lambda+\mu_1) &= P_{1,0,1}\cdot\mu_1 + P_{1,1,0}\cdot\mu_2 + P_{0,0,0}\cdot\lambda,\\
P_{1,0,k}\cdot(\lambda+\mu_1) &= P_{1,0,k+1}\cdot\mu_1 + P_{1,1,k}\cdot\mu_2 + P_{1,0,k-1}\lambda, 1\le k\le K-1,\\
P_{1,0,K}\cdot(\lambda+\mu_1) &= P_{1,1,K}\cdot\mu_2 + P_{1,0,K-1}\cdot\lambda,\\
P_{0,1,0}\cdot(\lambda+\mu_2) &= P_{1,1,0}\cdot\mu_1,\\
P_{1,1,0}\cdot(\lambda+\mu_1+\mu_2) &= P_{1,1,1}\cdot\mu_1 + P_{0,1,0}\cdot\lambda,\\
P_{1,1,k}\cdot(\lambda+\mu_1+\mu_2) &= P_{1,1,k+1}\cdot\mu_1 + P_{1,1,k-1}\cdot\lambda, 1\le k\le K-1,\\
P_{1,1,K}\cdot(\lambda+\mu_1+\mu_2) &= P_{1,1,K+1}\cdot(\mu_1+\mu_2) + P_{1,0,K}\cdot\lambda + P_{1,1,K-1}\cdot\lambda,\\
P_{1,1,K+k}\cdot(\lambda+\mu_1+\mu_2) &= P_{1,1,K+k+1}\cdot(\mu_1+\mu_2) + P_{1,1,K+k-1}\cdot\lambda, 1\le k< M,\\
P_{1,1,K+M}\cdot(\mu_1+\mu_2) &= P_{1,1,K+M-1}\cdot\lambda,
\end{aligned} \tag{3.1}$$

The obtained equations system is complemented by the normalization condition:

$$P_{0,0,0} + P_{0,1,0} + \sum_{i=0}^{K} P_{1,1,i} + P_{1,0,i} + \sum_{i=1}^{M} P_{1,1,K+i} = 1. \tag{3.2}$$

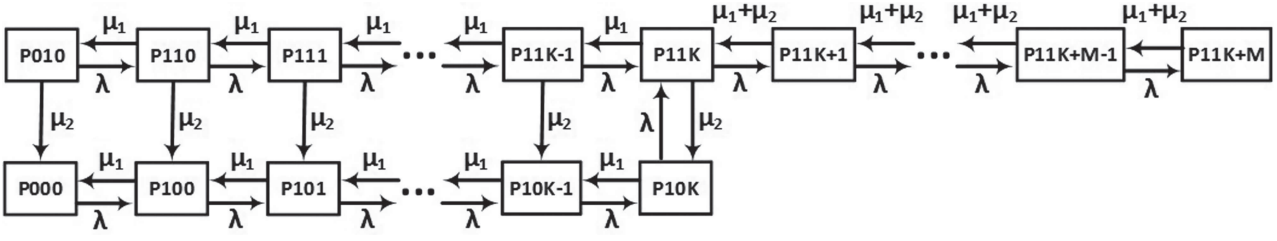

Figure 2. Graph of QS states.

Assume QS with the stalling buffer to be characterized by the parameters $q = \frac{\lambda}{\mu_1}$ and $r = \frac{\mu_1}{\mu_2}$. Then coefficient of system utilization is as follows: $\rho = \frac{\lambda}{\mu_1+\mu_2} = \frac{q}{1+\frac{1}{r}}$.

Note, that it is possible to obtain closed form expressions for the steady state, which as rather simple. Let as prove a following lemmas.

Lemma 3.1. *The distribution of the number of queries in a fast channel and a stalling buffer together corresponds to probabilities of states in the M/M/1 system with the utilization coefficient q, namely:*

$$P_{1,1,0} + P_{1,0,0} = q \cdot (P_{0,1,0} + P_{0,0,0})$$
$$P_{1,1,k} + P_{1,0,k} = q \cdot (P_{1,1,k-1} + P_{1,0,k-1}) = q^{k+1} \cdot (P_{0,0,0} + P_{0,1,0}), 0 < k \le K.$$

Proof. Summing up corresponding probabilities (3.1) of queries in the fast server and the stalling buffer one can make sure that:

$$(P_{0,1,0} + P_{0,0,0}) \cdot \lambda = (P_{1,1,0} + P_{1,0,0}) \cdot \mu_1$$
$$(P_{1,1,k} + P_{1,0,k}) \cdot (\lambda + \mu_1) = (P_{1,1,k+1} + P_{1,0,k+1}) \cdot \mu_1 + (P_{1,1,k-1} + P_{1,0,k-1}) \cdot \lambda,\ 1 < k < K - 1.$$

The latter system corresponds to steady state equations of the M/M/1 system with the utilization the coefficient q and therefore has a solution given by the lemma [10] ∇ □

In order to obtain explicit formulas of steady state the technique of Tchebyshev polynomials is developed. Note, that Tchebyshev polynomials are used to study properties of finite queues, namely, busy period of M/M/1/N [26], in an analysis of priority queue with a randomized push out mechanism [4].

Thus, let us introduce polynomial functions $A_k(q, r)$ defined in a recurrent manner and applied further to describe probabilities of the number of queries in a slow channel:

$$A_{-1}(q,r) = 1, A_0(q,r) = q + \frac{1}{r}, A_{k+1}(q,r) = \left(q + 1 + \frac{1}{r}\right) \cdot A_k(q,r) - q \cdot A_{k-1}(q,r),$$
$$k = 0, 1, 2, \ldots. \tag{3.3}$$

Let $U_k(x)$ be a Tchebyshev polynomial of the second kind of kth order, here $x = \frac{q+1+\frac{1}{r}}{2\sqrt{q}}$.

Hereinafter, denote, for short $A_k = A_k(q, r)$.

Lemma 3.2. *Functions A_k are presented as:*

$$A_k = \frac{(z-1) \cdot z^{k+1} - (t-1) \cdot t^{k+1}}{z - t},\ 0 \le k \le K, \tag{3.4}$$

where $z = \frac{q+1+\frac{1}{r}+\sqrt{\left(q-1+\frac{1}{r}\right)^2+\frac{4}{r}}}{2}$,

$$t = \frac{q + 1 + \frac{1}{r} - \sqrt{\left(q - 1 + \frac{1}{r}\right)^2 + \frac{4}{r}}}{2}.$$

Proof. Let us prove lemma by induction. Note,

$$A_0 = q + \frac{1}{r} = q^{\frac{1}{2}} \cdot 2 \cdot x - 1 = q^{\frac{1}{2}} \cdot U_1(x) - U_0(x),$$
$$A_1 = \left(q + \frac{1}{r}\right)^2 + \frac{1}{r} = q \cdot \left(4 \cdot x^2 - 1\right) - q^{\frac{1}{2}} \cdot 2 \cdot x = q \cdot U_2(x) - q^{\frac{1}{2}} \cdot U_1(x).$$

Next, by virtue of recurrent relation between Tchebyshev polynomials of the second kind [2]:

$$\begin{aligned} A_{k+1} &= q^{\frac{K+2}{2}} \cdot (2xU_{k+1}(x) - U_k(x)) - q^{\frac{K+1}{2}} (2xU_k(x) - U_{k-1}(x)) \\ &= \left(q + 1 + \frac{1}{r}\right) \cdot A_k - q \cdot A_{k-1}, \end{aligned}$$

which implies:

$$A_k = q^{\frac{K+1}{2}} \cdot U_{k+1}(x) - q^{\frac{K}{2}} \cdot U_k(x), k = 1, 2, \ldots$$

Then lemma follows from the well-known presentation of Tchebyshev polynomials [2].

$$A_k(x) = \frac{\left(x + \sqrt{x^2 - 1}\right)^{k+1} - \left(x - \sqrt{x^2 - 1}\right)^{k+1}}{2 \cdot \sqrt{x^2 - 1}} \nabla \tag{3.5}$$

□

One can easily verify that if $0 \leq q \leq 1$, $r \geq 1$, then: $z > 1$, $0 \leq t < 1$.

Theorem 3.3. *State probabilities of QS with stalling, defined by system (3.1), are as follows, if* $0 \leq \rho < 1$, $K \geq 0$, $M > 0$*:*

$$P_{1,1,k} = A_k \cdot P_{0,1,0}, 0 \leq k \leq K, \tag{3.6}$$

$$P_{1,0,k} = P_{0,1,0} \cdot \left(A_{K+1} \cdot q^{k-K-1} - A_k\right), 0 \leq k \leq K, \tag{3.7}$$

$$P_{1,1,K+k} = \rho^k \cdot A_K \cdot P_{0,1,0}, \ 1 \leq k \leq M, \tag{3.8}$$

$$P_{0,0,0} = \left(\frac{A_{K+1}}{q^{K+2}} - 1\right) \cdot P_{0,1,0}, \tag{3.9}$$

$$P_{0,1,0} = \frac{1}{A_{K+1} \frac{\sum_{i=0}^{K+1} q^i}{q^{K+2}} + A_K \cdot \sum_{i=0}^{M} \rho^i}, \tag{3.10}$$

If $M = 0$, $K \geq 0$, *then* $P_{0,1,0} = \frac{q^{K+2}}{A_{K+1} \cdot \sum_{i=0}^{K+1} q^i}$.

Proof. Let us rewrite the equations of states of a slow channel in the following manner:

$$\begin{aligned} P_{1,1,0} \cdot \mu_1 &= P_{0,1,0} \cdot (\lambda + \mu_2), P_{1,1,k+1} \cdot \mu_1 \\ &= P_{1,1,k} \cdot (\lambda + \mu_1 + \mu_2) - P_{1,1,k-1} \cdot \lambda, 1 \leq k \leq K - 1. \end{aligned}$$

Thus, the latter equation and (3.3) imply (3.6). Next, whenever one inserts the last equation in (3.1) into the previous one and does the same further, the formulas

$$P_{1,1,K+k} \cdot (\mu_1 + \mu_2) = P_{1,1,K+k-1} \cdot \lambda, 1 \leq k \leq M,$$

are obtained, that imply probabilities (3.8), describing the states of the waiting queue.

Next, inserting formula (3.8) into equation in (3.1), which describes the probability $P_{1,1,K}$, and using Lemma 3.1, the following equation is derived:

$$P_{1,1,K} \cdot (\lambda + \mu_1 + \mu_2) - P_{1,1,K-1} \cdot \lambda = P_{1,1,K} \cdot \lambda + P_{1,0,K} \cdot \lambda.$$

Then, this equation, (3.6), and (3.3) imply equation (3.9). Next, (3.7) is conclusion of (3.6), Lemma 3.1 and (3.9). Since now all the probabilities of states expressed have to be proportional to probability of only one query in slow channel, *e.g.*, $P_{0,1,0}$, for calculating the latter probability formula (3.10) follows from (3.6)–(3.9) and normalization equation (3.2)∇ □

4. Queueing characteristics

Using probabilities (3.6)–(3.10) the characteristics of QS with stalling are computed as follows.
Occupancy probability of the fast channel is equal to:

$$P_I = \sum_{k=0}^{K}(P_{1,0,k} + P_{1,1,k}) + P_{1,1,K} \cdot \sum_{i=1}^{M} \rho^i = 1 - P_{0,1,0} - P_{0,0,0} = 1 - \frac{A_{K+1}}{q^{K+2}} \cdot P_{0,1,0} \tag{4.1}$$

Also occupancy probability of the second channel is calculated:

$$P_{II} = \sum_{k=0}^{K} P_{1,1,k} + P_{0,1,0} + P_{1,1,K} \cdot \sum_{i=1}^{M} \rho^i = \left(1 + \sum_{k=0}^{K} A_k + A_K \cdot \sum_{k=1}^{M} \rho^k\right) \cdot P_{0,1,0} \tag{4.2}$$

Further, average number of serviced queries in system is calculated:

$$\begin{aligned} \bar{N}_S &= \sum_{i=0}^{K} (P_{1,0,i} + 2 \cdot P_{1,1,i}) + P_{0,1,0} + 2 \cdot P_{1,1,K} \cdot \sum_{i=1}^{M} \rho^i \\ &= \left(1 + \sum_{k=0}^{K} A_k + \frac{A_{K+1}}{q^{K+2}} \cdot \sum_{k=0}^{K+1} q^k + 2 \cdot A_K \cdot \sum_{k=1}^{M} \rho^k\right) \cdot P_{0,1,0}. \end{aligned} \tag{4.3}$$

Also the average number of waiting queries correspondingly in the stalling buffer and queue is obtained as well:

$$\begin{aligned} \bar{N}_K &= \sum_{k=1}^{K} k \cdot (P_{1,0,k} + P_{1,1,k}) + P_{1,1,K} \cdot K \cdot \sum_{i=1}^{M} i \cdot \rho^i \\ &= \left(\frac{A_{K+1}}{q^{K+2}} \cdot \sum_{k=1}^{K+1} k \cdot q^k + K \cdot A_K \cdot \sum_{i=1}^{M} i \cdot \rho^i\right) \cdot P_{0,1,0} \end{aligned} \tag{4.4}$$

$$\bar{N}_W = P_{1,1,K} \cdot \sum_{i=1}^{M} i \cdot \rho^i = A_K \cdot P_{0,1,0} \cdot \sum_{i=1}^{M} i \cdot \rho^i. \tag{4.5}$$

The queries are lost when stalling buffer and queue are full, therefore the probability of the loss of query is equal to

$$P_{\text{loss}} = A_K \cdot \rho^M \cdot P_{0,1,0}. \tag{4.6}$$

5. Optimization of stalling buffer

Average number of queries in the system is most important, because it describes the whole system capacity and the efficiency of the chosen queueing strategy. Thus, the total average amount of serviced, stalled, and waiting queries in the system is as follows:

$$\begin{aligned} \bar{N} &= \sum_{i=0}^{K} ((i+1) \cdot P_{1,0,i} + (i+2) \cdot P_{1,1,i}) + +P_{0,1,0} + P_{1,1,K} \cdot \sum_{i=1}^{M} \rho^i \cdot (i + K + 2) \\ &= \frac{1 + \sum_{i=0}^{K} A_i + \frac{A_{K+1}}{q^{K+2}} \cdot \sum_{i=1}^{K+1} i \cdot q^i + A_K \cdot \sum_{i=1}^{M} (i + K + 2) \cdot \rho^i}{A_{K+1} \frac{\sum_{i=0}^{K+1} q^i}{q^{K+2}} + A_K \cdot \sum_{i=1}^{M} \rho^i}. \end{aligned} \tag{5.1}$$

Note, that sums in (5.1) might be expressed analytically [2]:

$$\sum_{i=0}^{K+1} q^i = \frac{1-q^{K+1}}{1-q}, \tag{5.2}$$

$$\sum_{i=1}^{K+1} i\cdot q^i = \frac{[1-(K+2)\cdot q^{K+1}+(K+1)\cdot q^{K+2}]\cdot q}{(1-q)^2}, \tag{5.3}$$

$$S_1 = \sum_{i=1}^{M} \rho^i = \frac{\rho-\rho^{M+1}}{1-\rho}, \tag{5.4}$$

$$S_2 = \sum_{i=1}^{M} i\cdot \rho^i = \frac{\rho-(M+1)\cdot\rho^{M+1}+M\cdot\rho^{M+2}}{(1-\rho)^2}. \tag{5.5}$$

And by virtue of (3.5):

$$1+\sum_{i=0}^{K} A_i = \frac{z^{K+2}-t^{K+2}}{z-t}. \tag{5.6}$$

Thus, after some simple transformations (5.1) is rewritten by means of (3.4) and (5.2)–(5.6) in an analytical shape:

$$\begin{aligned}\bar{N} = \frac{q}{1-q} + q^{K+2}\frac{V}{W} = \frac{q}{1-q} + q^{K+2} \\ \times \frac{\frac{z^{K+2}-t^{K+2}}{(z-t)\cdot A_{K+1}} - \frac{z\cdot(K+2)}{1-q} + \frac{A_K}{A_{K+1}}\cdot\left(S_1\cdot\left(K+2-\frac{q}{1-q}\right)+S_2\right)}{\frac{z}{1-q}+q^{K+2}\cdot\left(\frac{A_K}{A_{K+1}}\cdot S_1 - \frac{z}{1-q}\right)}.\end{aligned} \tag{5.7}$$

Note, this explicit expression is stable numerically, no sums, and, thus, it helps us to study QS characteristics analytically. For instance, using (5.7) one can differentiate $\bar{N}$ with respect to parameter K:

$$\frac{\mathrm{d}\bar{N}}{\mathrm{d}K} = q^{K+2}\cdot\left(\ln(q)\cdot\frac{V}{W} + \frac{\frac{\mathrm{d}V}{\mathrm{d}K}\cdot W - \frac{\mathrm{d}W}{\mathrm{d}K}\cdot V}{W^2}\right). \tag{5.8}$$

One can easily make sure also that:

$$\begin{aligned}\frac{z^{K+2}-t^{K+2}}{A_{K+1}} &= \frac{1}{z-1}\left(1-\left(\frac{t}{z}\right)^{K+2}\cdot\frac{(z-t)}{z-1-(t-1)\cdot\left(\frac{t}{z}\right)^{K+2}}\right) \\ &= \frac{1}{z-1}+O\left(\left(\frac{t}{z}\right)^{K+2}\right)\end{aligned} \tag{5.9}$$

$$\frac{A_K}{A_{K+1}} = \frac{1}{z}+\left(\frac{t}{z}\right)^{K+1}\frac{(1-t)\cdot(z-t)}{(z-1)\cdot z^2-(t-1)\cdot t^2\cdot\left(\frac{t}{z}\right)^{K+1}} = \frac{1}{z}+O\left(\left(\frac{t}{z}\right)^{K+1}\right) \tag{5.10}$$

Theorem 5.1. *Assume,* $0<q<1$, $r\geq 1$, $K, M\geq 0$. *Then:*

1) $$\begin{aligned}\frac{\mathrm{d}\bar{N}}{\mathrm{d}K} = q^{K+2}\cdot\frac{(1-q)\cdot\ln(q)}{z} \\ \times\left[\left(S_1-\frac{z}{1-q}\right)\cdot\left(K+2+\frac{1}{\ln(q)}\right)+\frac{z}{z-1}-S_1\cdot\frac{q}{1-q}+S_2\right]+O\left(q^{2K}\right),\end{aligned}$$

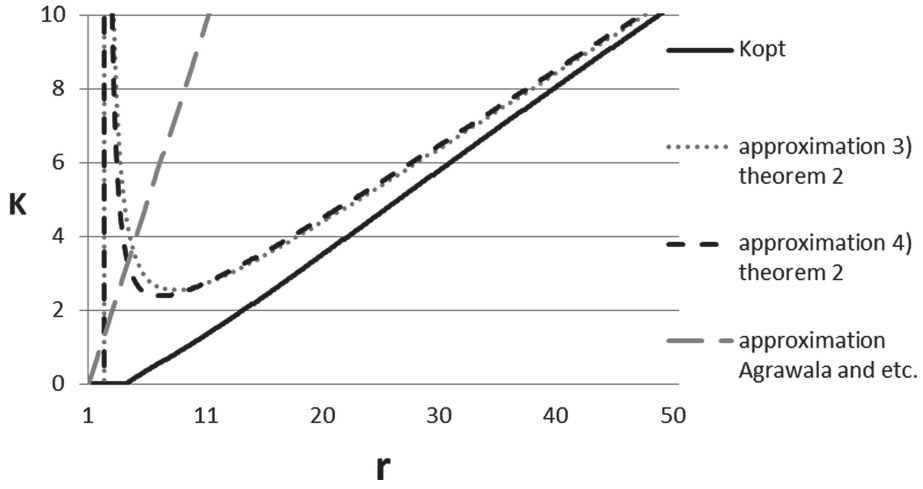

FIGURE 3. Comparison of optimal buffer size and approximations.

2) there exists a value K_0 *such, that* $\frac{\mathrm{d}\bar{N}}{\mathrm{d}K}_{K=K_0} = 0$, $0 \le K_0 < \infty$*;*

$$3)\quad K_0 = \frac{\frac{z}{z-1} - \frac{S_1 \cdot q}{1-q} + S_2}{\frac{z}{1-q} - S_1} - \frac{1}{\ln(q)} - 2 + O\left(q^K\right);$$

$$4)\quad K_0 = r \cdot \frac{(1-q)^2}{1-q+q^{M+1}} - \frac{q^{M+1}}{(1-q+\rho^{M+1})} \times \left(M + 1 - \frac{2 + M \cdot (1-q)^2}{(1-q+\rho^{M+1})}\right) - \frac{1}{\ln(q)} - 2 + O\left(\frac{1}{r}\right).$$

Proof. It follows from (5.7) and (5.8) that $\frac{\mathrm{d}\bar{N}}{\mathrm{d}K} = q^{K+2} \cdot \left(\ln(q) \cdot \frac{V \cdot 1 - q}{z} - \frac{\mathrm{d}W}{\mathrm{d}K} \cdot \frac{(1-q)^2}{z^2}\right) + O(q^{2K+4})$. Thus, 1) is implied by (5.9), (5.10). It is easy to make sure that $\frac{z}{1-q} > S_1$, which implies 2), because *ln(q)<0.* Then 3) is obtained by equating expression 1) to zero and solving a corresponding equation. Formula 4) is the Taylor approximation of expression (3.3)∇ □

Hence, numerical optimization of the stalling buffer size might be directly performed by means of (5.7) by starting from $K = 0$ and consequently selecting K: $K_{\mathrm{opt}} = \arg\min_{K \ge 0} \overline{N}(K)$. The other way for buffer optimization is to solve the equation, following from (5.8): $\frac{\mathrm{d}\bar{N}}{\mathrm{d}K} = 0$ with respect to K, and round the obtained solution. As follows from Theorem 5.1, these approaches provide us finite exact values of optimal stalling buffer size.

Hence, Theorem 5.1 enables us to apply simple analytical approximations 3) and 4). Optimal buffer size (computed according to equation $\frac{\mathrm{d}\bar{N}}{\mathrm{d}K} = 0$), approximations 3) and 4) (Thm. 5.1), and approximation of Agrawala *et al.* [3,23] in a system with infinite waiting line are depicted on Figure 3. One can see the accuracy of considered approximations. Note, approximation, proposed in Agrawala *et al.*, is $K = r - 1$, that is quietly far from optimal.

The Figure 4 presents the average number of queries in system in stalling with optimal buffer size and infinite waiting line under various ratio r. This figure also illustrates approximation 4) in Theorem 5.1.

The Figure 5 presents the ratio of probabilities of processors occupancy $\frac{P_{II}}{P_I}$ in optimal stalling regime. One can see that if system load is low or moderate the slow server is mostly stalled and is involved to work if the load is increased much.

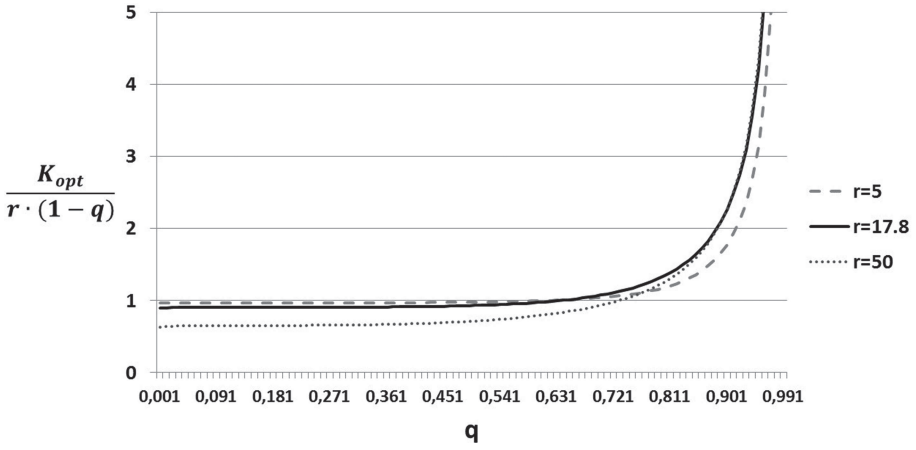

FIGURE 4. Ratio of Optimal Stalling Buffer size.

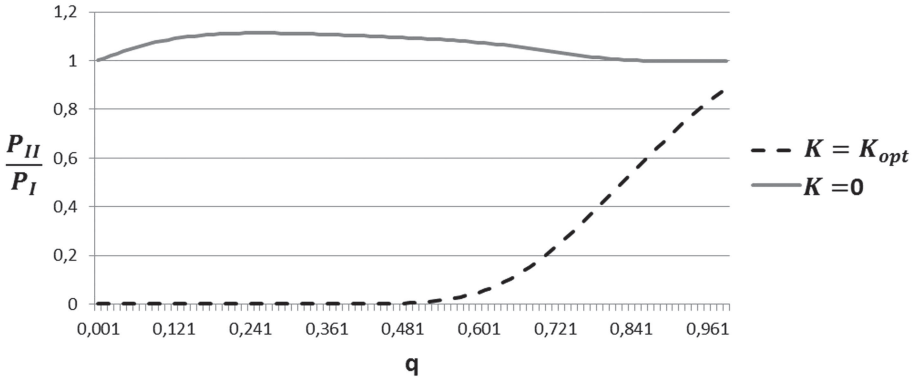

FIGURE 5. Ratio of probabilities of processors occupancy in stalling regime and without stalling.

Thus, optimal stalling enables us to solve slow server problem. Note that, QS with the optimal stalling buffer size, which is finite according to Theorem 5.1, outperforms the system without stalling as well as QS with discharged slow server ($K = \infty$), if the length of waiting line $M$is sufficient.

One can see on Figure 6 that in a case of high traffic without stalling it is better to discharge the slow server and its capacity remains unused. However, under various system loads optimal stalling provides better service than service without stalling or discharged slow server.

Example. Compatibility problems of heterogeneous networks are solved by offering specialized data maintenance solutions, evaluating their combining cases, analyzing errors, links, etc. Optimization of the network node stalling buffer is important for combining networks of different capacity [25]. Web hosting company's forced to change old slow server to new fast server in oreder to serve increasing flow of service. The optimal solution for this problem is to use both fast and slow servers connected to one heterogeneous servers cluster equipped by stalling buffer (Fig. 7). We calculate stalling buffer length changing old server with Intel Xeon E5-2679 v4 @ 2.50 GHz (CPU Benchmarks 4862) into server Intel Xeon X3450 @ 2.67 GHz processor (CPU Benchmarks 25236) [20]. Servers use 32GB working memory. Coefficient of QS utilization $\rho = 0.34$, $r = 5.19$, calculated stalling buffer size for this heterogeneous servers cluster solving the equation $\frac{\mathrm{d}\bar{N}}{\mathrm{d}K} = 0$ is 2.129. By rounding optimal stalling buffer size is obtained: $K = 2$.

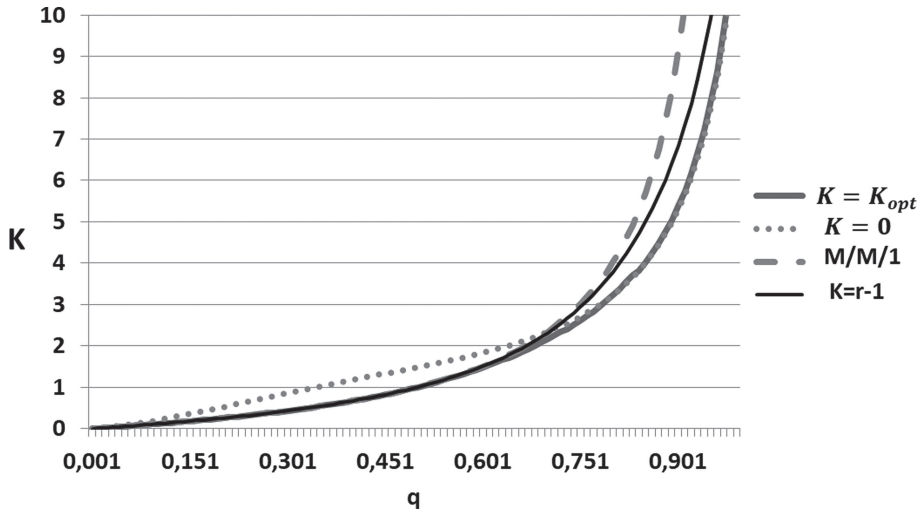

FIGURE 6. Comparison of QS with optimal stalling buffer, without stalling, M/M/1, $K = r - 1$ [3] when $r = 15$.

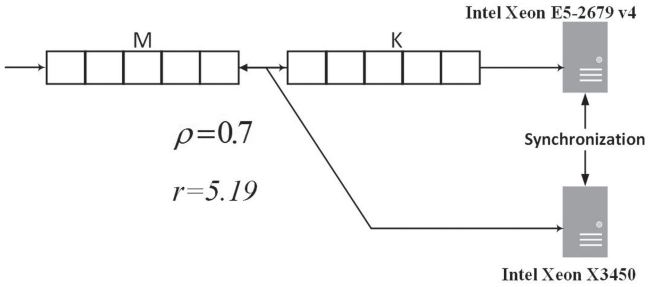

FIGURE 7. Heterogeneous servers cluster with stalling buffer.

Characteristics of the cluster with stalling are as follows: occupancy probability of new server – $P_I = 0.393$, that of old – $P_{II} = 0.064$, number of queries in service – $\bar{N}_S = 0.457$, number of stalled – $\bar{N}_K = 0.194$, number of waiting in line – $\bar{N}_W = 0.01$, probability of loss – $P_{\text{loss}} = 0$, average number of queries in system – $\overline{N} = 0.661$.

6. Conclusions

The model of stalling in QS with two heterogeneous severs is presented deriving the explicit probabilities of steady states by Tchebyshev polynomials of second order. The obtained expressions are numerically stable, their complexity does not depend on the number of states and provides us a way to study QS characteristics analytically. The existence of a finite optimal size of the stalling buffer is proved and numerical approach for buffer size optimization is developed.

The results shown that the appropriate choice of stalling buffer size helps us to solve the slow server problem. The asymptotic approximations of optimal stalling buffer size, when the ratio of capacities of fast and slow servers is increasing, are established, too. Application of model developed in computer networks is discussed as well.

Investigation of developed model enables us to conclude that stalling is universal solution, which can be implemented into any heterogeneous system. The findings done in this article for two severs QS might be useful for investigation of systems with any number of heterogeneous channels or servers.

References

[1] M.O. Abou-El-Ata and A.I. Shawky, A simple approach for the slower server problem. *Commun. Faculty Sci. Ankara Univ. Ser. A* **1** (1999) 1–6.

[2] M. Abramowitz and I.A. Stegun, Handbook of mathematical functions with formulas, graphs, and mathematical tables. In Vol. 55 of *National Bureau of Standards, Applied Mathematics Series*. (1983).

[3] A.K. Agrawala, E.G. Coffman Jr., M.R. Garey and S.K. Tripathi, A stochastic optimization algorithm minimizing expected flow times on uniform processors. *IEEE Trans. Comput.* **33** (1984) 351–356.

[4] K.E. Avrachenkov, G.L. Shevlyakov and N.O. Vilchevskii, Randomized push-out disciplines in priority queueing. *J. Math. Sci.* **122** (2004) 3336–3342.

[5] B.R. Bilel, N. Navid and M.S.M. Bouksiaa, Hybrid cpu-gpu distributed framework for large scale mobile networks simulation. In: *Proceedings of the 2012 IEEE/ACM 16th International Symposium on Distributed Simulation and Real Time Applications*. IEEE Computer Society (2012, October) 44–53.

[6] F. De Vericourt and Y.P. Zhou, On the incomplete results for the heterogeneous server problem. *Queue. Syst.* **52** (2006) 189–191.

[7] D. Efrosinin and J. Sztrik, Performance analysis of a two-server heterogeneous retrial queue with threshold policy. *Qual. Technol. Quant. Manage.* **8** (2011) 211–236.

[8] D. Efrosinin and V. Rykov, Heuristic solution for the optimal thresholds in a controllable multi-server heterogeneous queueing system without preemption. In: *International Conference on Distributed Computer and Communication Networks*. Springer, Cham (2015) 238–252.

[9] D. Efrosinin and J. Sztrik, Optimal control of a two-server heterogeneous queueing system with breakdowns and constant retrials. In: *International Conference on Information Technologies and Mathematical Modelling*. Springer, Cham (2016) 57–72.

[10] G. Giambene, Queuing Theory and Telecommunications. Springer US (2005)

[11] V. Goswami and S.K. Samanta, Discrete-time bulk-service queue with two heterogeneous servers. *Comput. Ind. Eng.* **56** (2009) 1348–1356.

[12] T.H. Hetherington, T.G. Rogers, L. Hsu, M. O'Connor and T.M. Aamodt, Characterizing and evaluating a key-value store application on heterogeneous CPU-GPU systems. In: *IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*. IEEE (2012) 88–98.

[13] D. Kadjo, R. Ayoub, M. Kishinevsky and P.V. Gratz, A control-theoretic approach for energy efficient CPU-GPU subsystem in mobile platforms. In: *Proceedings of the 52nd Annual Design Automation Conference*. ACM (2015) 62.

[14] G. Koole, A simple proof of the optimality of a threshold policy in a two-server queueing system. *Syst. Control Lett.* **26** (1995) 301–303.

[15] R.L. Larsen, *Control of multiple exponential servers with application to computer systems*. Ph.D. thesis, University of Maryland, College Park, MD, USA (1981).

[16] R.L. Larsen and A.K. Agrawala, Control of a heterogeneous two-server exponential queueing system. *IEEE Trans. Softw. Eng.* **4** (1983) 522–526.

[17] W. Lin and P. Kumar, Optimal control of a queueing system with two heterogeneous servers. *IEEE Trans. Autom. Control* **29** (1984) 696–703.

[18] M.F. Neuts, Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach. Baltimore, Johns Hopkins University Press (1981).

[19] E. Ozkan and J.P. Kharoufeh, Optimal control of a two-server queueing system with failures. *Probab. Eng. Info. Sci.* **28** (2014) 489–527.

[20] PassMark software, CPU Benchmarks. Interactyve: Available at: https://www.cpubenchmark.net/high_end_cpus.html (2017).

[21] V.V.E. Rykov and D.V. Efrosinin, On the slow server problem. *Autom. Remote Control* **70** (2009) 2013–2023.

[22] V.V. Rykov, Monotone control of queueing systems with heterogeneous servers. *Queue. Syst.* **37** (2001) 391–403.

[23] Z. Rosberg and A.M. Makowski, Optimal routing to parallel heterogeneous servers-small arrival rates. *IEEE Trans. Autom. Control* **35** (1990) 789–796.

[24] M. Rubinovitch, The slow server problem: a queue with stalling. *J. Appl. Probab.* **22** (1985) 879–892.

[25] C. Shi, Y. Li, J. Zhang, Y. Sun and S.Y. Philip, A survey of heterogeneous information network analysis. *IEEE Trans. Knowl. Data Eng.* **29** (2017) 17–37.

[26] H. Takagi and A.M. Tarabia, Explicit probability density function for the length of a busy period in an M/M/1/K queue. In: Advances in Queueing Theory and Network Applications. Springer, New York (2009) 213–226.

[27] I. Viniotis and A. Ephremides, Extension of the optimality of the threshold policy in heterogeneous multiserver queueing systems. *IEEE Trans. Autom. Control* **33** (1988) 104–109.

[28] S.H. Xu, A duality approach to admission and scheduling controls of queues. *Queue. Syst.* **18** (1994) 273–300.