On the Problems of Queues in Mixed Type Queuing Systems with Random Quantity of Sources and Size-Limited Queues

Alexander Kirpichnikov and Anton Titovtsev(✉)

Kazan National Research Technological University,
K. Marksa str. 68, 420015 Kazan, Russia
kirpichnikov@kstu.ru, notna6683@mail.ru
http://www.kstu.ru

Abstract. The article proposes the technique to investigate the behavior of the moments of numerical characteristics of mixed-type queuing system with a random number of sources upon the change of demands input stream intensity and size-limited queues based on the calculation of boundary values of the number of servicing devices at which the mean squared deviation (MSD) of the investigated quantity does not exceed its mathematical expectation. For the first time the linear nature of behavior of boundary values of the number of service facilities with the change of the given intensity of demands input stream is determined numerically. The article also considers various types of queues arising in queuing systems. The concept of an N-th order queue is introduced, and generalized Little's formulas for N-th order queues in queuing systems of various types are presented.

Keywords: Queue · Physical queue · Real queue · Quality of service (QoS) · Queuing system · M/M/m/K · Service facility

1 Introduction

Issues of studying combined models of queuing originate from Cohen's works (Cohen J.W.) [1], where the combination of Erlang models and classical queuing system was considered for the first time. A number of formulae for probabilities of queuing system (QS) steady states, call loss probability, and first moments of demands number in a queue and waiting time in a queue are given in the paper.

Another specific case of a combined model is a mixed system with losses and expectation having some servers and finite memory, presented in the work of H. Takagi [2]. In this case there are two sources of demands in the system, thus demands from the first source will be lost if all servers are busy at the time of their arrival in the system. Demands from the second source are accepted in a queue only if the number of demands in it does not exceed some defined value K. Streams of demands arriving in the system also have a Poisson character. Formulae for probabilistic characteristics of the system and for the moments of

A. Dudin et al. (Eds.): ITMM 2017, CCIS 800, pp. 68–82, 2017.
DOI: 10.1007/978-3-319-68069-9_6

n order of waiting time and common delay time in the system are given in the paper. In the specific case K → ∞, this model is reduced to J. Cohen's model.

A more general model of a queuing system which is a combination of a multi-channel Erlang model, M/M/m/E model, and also multi-channel classical model (M/M/m models) is considered in the work of authors [3]. A complete formula derivation for probabilistic characteristics, and also for the first and second moments of numerical and temporary characteristics of this type of a queuing system is presented in work [4]; a general algorithm of queuing models mathematical formalization taken from monographs [5,6] is used.

A mathematical model of an open multi-channel system of queuing having m service facilities of identical efficiency with exponentially distributed service time is presented in this paper. A demand input stream in this case is a superposition of components'random number h, each of which represents a Poisson stream of demands served in the order of arrival. For each type of demands entering the system from the j-th source there is a specific size-limited queue ε_j where $\varepsilon_0 < \varepsilon_1 < \varepsilon_2 < \cdots < \varepsilon_h$.

A zero (Erlang) component contains demands which are served only if there is at least one free service facility, and they never stand in a queue. In the case, if at the time when the next similar demand arrives in the system there is no free service facility this demand is refused and leaves the system unserved. The model of a queuing system, containing one such component in an input stream, is the Erlang model; therefore we will call this component an Erlang component.

The first component includes demands which are served if there is a free service facility, or they stand in a queue if the number of demands in the queue is fewer than a particular number ε_1. In case when there is already available ε_1 or more demands in a queue, a newly arrived demand from the first source is refused and leaves the system unserved.

The second component contains demands which are served if there is a free service facility, or they stand in a queue if the number of demands in a queue is fewer than a particular number $\varepsilon_2 > \varepsilon_1$. In the case when ε_2 or more demands are already available in the queue, an arrived demand from this source is refused and leaves the system unserved, and so on.

In general, the h-th component includes demands which are served if there is a free service facility, or they stand in a queue if the number of demands in the queue are fewer than a particular number $\varepsilon_h > \varepsilon_{h-1} > \cdots > \varepsilon_1$. In case when there are already ε_h demands in the queue, a newly arrived demand from the h-th source is refused and leaves the system unserved.

Let us accept the following designations:

$\varepsilon_0 = E_0 = 0$; $\varepsilon_1 = E_1$; $\varepsilon_2 = E_1 + E_2$; $\cdots$ $\varepsilon_j = \sum\limits_{i=0}^{j} E_i = \sum\limits_{i=1}^{j} E_i$; a size-limited queue (memory volume) for demands of the j-th component;

$\Lambda_0 = \sum\limits_{j=0}^{h} \lambda_j$; $\Lambda_1 = \sum\limits_{j=1}^{h} \lambda_j$; $\Lambda_2 = \sum\limits_{j=2}^{h} \lambda_j$; $\cdots$ $\Lambda_h = \lambda_h$; where λ_j demand stream intensity of the j-th component;

$R_0 = \sum\limits_{j=0}^{h} \rho_j$; $R_1 = \sum\limits_{j=1}^{h} \rho_j$; $R_2 = \sum\limits_{j=2}^{h} \rho_j$; $\cdots$ $R_h = \rho_h$; $R_i = \frac{\Lambda_i}{\mu}$, where ρ_j is the given demand stream intensity of the j-th component.

Demand streams arriving from each source are Poisson and have intensity λ_j; in this case total streams with intensities Λ_j also have, as we know, a Poisson character. Let us designate the mean intensity of demand service by one service facility as μ. In this case the intensity of an output stream of served demands before the m-th states is multiple μ and depends on the number of busy channels. After the m-th state the intensity of served demand stream is equal to $m\mu$. The served demand stream is also Poisson.

With accepted designations and assumptions taken into account, we will obtain a continuous-time Markov chain.

2 Probabilistic Characteristics of a Queuing System in a Steady-State Mode

We make up a set of Kolmogorov-Chapman equations for probabilities of QS states in a steady-state mode of its functioning. Adding the normalization condition $\sum\limits_{i=0}^{m+\varepsilon_h} P_i = 1$, to this set of equations, we obtain a system that has a unique solution

$$P_0 = \left[e_m\left(R_0\right) + \frac{R_0^m}{m!} \sum_{g=1}^{h} \prod_{j=0}^{g-1} \left(\frac{R_j}{m}\right)^{E_j} \times \begin{Bmatrix} \frac{R_g}{m-R_g}\left(1-\left(\frac{R_g}{m}\right)^{E_g}\right), & R_g \neq m \\ E_g, \quad R_g = m \end{Bmatrix} \right]^{-1}; \qquad (1)$$

$$P_i = \begin{cases} \frac{R_0^i}{i!} P_0, \quad 0 < i \leq m, \\ \left(\frac{R_{j+1}}{m}\right)^{i-m-\varepsilon_j} \prod\limits_{g=0}^{j} \left(\frac{R_g}{m}\right)^{E_g} \frac{R_0^m}{m!} P_0, \quad m+\varepsilon_j \leq i \leq m+\varepsilon_{j+1}, \\ \qquad 0 \leq j \leq h-1, \end{cases} \qquad (2)$$

where the designation $e_m\left(R_0\right) = \sum\limits_{i=0}^{m} \frac{R_0^i}{i!}$ is accepted - a non-complete exponential function. The solution (1) and (2) defines expressions for probabilities of all possible QS states of this type in a steady-state mode of its functioning.

For further calculations it is convenient to introduce the following basic probabilistic characteristics of QS of this type through which all other quantities are expressed:

- basic probability 1

$$P_{B1} = \sum_{i=m}^{m+\varepsilon_1-1} P_i = \frac{1-\left(\frac{R_1}{m}\right)^{E_1}}{1-\frac{R_1}{m}} \frac{R_0^m}{m!} P_0;$$

- basic probability 2

$$P_{B2} = \sum_{i=m+\varepsilon_1}^{m+\varepsilon_2-1} P_i = \frac{1-\left(\frac{R_2}{m}\right)^{E_2}}{1-\frac{R_2}{m}} \left(\frac{R_1}{m}\right)^{E_1} \frac{R_0^m}{m!} P_0;$$

$$\vdots$$

- basic probability h

$$P_{Bh} = \sum_{i=m+\varepsilon_{h-1}}^{m+\varepsilon_h-1} P_i = \frac{1-\left(\frac{R_h}{m}\right)^{E_h}}{1-\frac{R_h}{m}} \prod_{g=1}^{h-1} \left(\frac{R_g}{m}\right)^{E_g} \frac{R_0^m}{m!} P_0;$$

- congestion probability of the system

$$P_{m+\varepsilon_h} = \prod_{g=1}^{h} \left(\frac{R_g}{m}\right)^{E_g} \frac{R_0^m}{m!} P_0. \tag{3}$$

As a result, a general formula for basic probability is written in the form

$$P_{Bi} = \prod_{g=0}^{i-1} \left(\frac{R_g}{m}\right)^{E_g} \frac{R_0^m}{m!} P_0 \begin{cases} \frac{m}{m-R_i}\left(1-\left(\frac{R_i}{m}\right)^{E_i}\right), & R_i \neq m \\ E_i, \quad R_i = m \end{cases}. \tag{4}$$

By means of the expression (4) it is possible to present traditional probabilistic characteristics of a queuing system in the most compact form:

- probability of a newly arrived demand service expectation in the queue

$$P_W = \frac{\Lambda_1}{\Lambda_0} \sum_{i=m}^{m+\varepsilon_1-1} P_i + \frac{\Lambda_2}{\Lambda_0} \sum_{i=m+\varepsilon_1}^{m+\varepsilon_2-1} P_i + \frac{\Lambda_3}{\Lambda_0} \sum_{i=m+\varepsilon_2}^{m+\varepsilon_3-1} P_i + \cdots$$
$$+ \frac{\Lambda_h}{\Lambda_0} \sum_{i=m+\varepsilon_{h-1}}^{m+\varepsilon_h-1} P_i = \frac{1}{R_0} \sum_{i=1}^{h} R_i P_{Bi};$$

- probability of a newly arrived demand service refusal (probability of demand loss)

$$P_L = \frac{\Lambda_0-\Lambda_1}{\Lambda_0} \sum_{i=m}^{m+\varepsilon_1-1} P_i + \frac{\Lambda_0-\Lambda_2}{\Lambda_0} \sum_{i=m+\varepsilon_1}^{m+\varepsilon_2-1} P_i + \frac{\Lambda_0-\Lambda_3}{\Lambda_0} \sum_{i=m+\varepsilon_2}^{m+\varepsilon_3-1} P_i + \cdots$$
$$+ \frac{\Lambda_0-\Lambda_h}{\Lambda_0} \sum_{i=m+\varepsilon_{h-1}}^{m+\varepsilon_h-1} P_i + P_{m+\varepsilon_h} = \frac{1}{R_0} \sum_{i=1}^{h} (R_0-R_i) P_{Bi} + P_{m+\varepsilon_h}$$
$$= \sum_{i=1}^{h} P_{Bi} - P_W + P_{m+\varepsilon_h} = 1 - P_{IS} - P_W.$$

The probability of an immediate service of a newly arrived demand has, apparently, a form

$$P_{IS} = \sum_{i=0}^{m-1} P_i = e_{m-1}(R_0) P_0. \tag{5}$$

3 Numerical Characteristics of a Queuing System

By means of probabilistic characteristics of the system found above, it is possible to express all main features characterizing a steady-state mode of a queuing system functioning. So, through put capacity of a queuing system is a number of demands passing through the system per unit of time $A = \Lambda_0 q = \Lambda_0 \left(1 - P_L\right) = \Lambda_0 \left(P_{IS} + P_W\right)$. This number includes all demands from a general input stream except refused demands and those that did not get into the system. Relative through put capacity of the system, thus, is a share of demands passing through a queuing system from a general input stream of demands $q = 1 - P_L$. The average number of demands under service at the same time (or, that is the same, an average number of busy channels) with formulae (2)–(5) taken into account has a form

$$\begin{aligned}\bar{n} &= \sum_{i=1}^{m-1} iP_i + m \sum_{i=m}^{m+\varepsilon_h} P_i = R_0 P_0 e_{m-2}(R_0) + m\left(P_W + P_L\right)\\ &= R_0 P_0 e_{m-2}(R_0) + m\left(\sum_{i=1}^{h} P_{Bi} + P_{m+\varepsilon_h}\right).\end{aligned}$$

The second initial moment of demands number under service is

$$\begin{aligned}\overline{n^2} &= \sum_{i=1}^{m-1} i^2 P_i + m^2 \sum_{i=m}^{m+\varepsilon_h} P_i\\ &= R_0 P_0 e_{m-2}(R_0) + R_0^2 P_0 e_{m-3}(R_0) + m^2\left(\sum_{i=1}^{h} P_{Bi} + P_{m+\varepsilon_h}\right).\end{aligned}$$

An average demands number in a queue (average queue length) are

$$\begin{aligned}\bar{l} &= \sum_{i=m+1}^{m+\varepsilon_h} \left(i - m\right) P_i\\ &= \sum_{i=1}^{h} \left\{\begin{array}{ll} \frac{R_i}{m-R_i}\left[P_{Bi} - E_i P_{m+\varepsilon_i}\right], & R_i \neq m\\ \frac{E_i(E_i+1)}{2} P_{m+\varepsilon_{i-1}}, & R_i = m \end{array}\right\} + \sum_{i=2}^{h} \varepsilon_{i-1} \frac{R_i}{m} P_{Bi}.\end{aligned}$$

The second initial moment of demands number in a queue is

$$\begin{aligned}\overline{l^2} &= \sum_{i=m+1}^{m+\varepsilon_h} \left(i - m\right)^2 P_i\\ &= \sum_{i=1}^{h} \left[\varepsilon_{i-1}^2 P_{Bi} + \left\{\begin{array}{l} \frac{R_i}{m-R_i}\left(\frac{m+R_i}{m-R_i} + 2\varepsilon_{i-1}\right) P_{Bi} -\\ -\frac{mE_i}{m-R_i}\left(E_i + \frac{2R_i}{m-R_i} + 2\varepsilon_{i-1}\right) P_{m+\varepsilon_i}, R_i \neq m\\ \left(E_i - 1\right) E_i \left(\frac{2E_i-1}{6} + \varepsilon_{i-1}\right) P_{m+\varepsilon_i}, \quad R_i = m \end{array}\right\}\right]\\ &\quad + \varepsilon_h^2 P_{m+\varepsilon_h}.\end{aligned}$$

Further, in the considered queuing system,the queue is possible only when all service facilities are busy. Thus, the total stream of served demands of the whole system consists of service streams of each channel and has $m\mu$ intensity. In this case, the probability that the system serves i demands during t time in the event of queue, will be recorded in the form $B_i\left(t\right) = \frac{(m\mu\, t)^i}{i\,!}e^{-m\mu\, t}$.

The function of service waiting time distribution for one demand we will find according to a known dependence $F_W\left(t\right) = 1 - P\left(t_W \geq t\right)$, where $P\left(t_W \geq t\right)$ - the probability that waiting time in a queue for one demand is more than an advanced set time t. As it is easy to see, it is possible, firstly, in case when the queue is absent, but a newly arrived demand finds all service facilities in the system busy, and during t time none of facilities is released. Secondly, in case when one demand is already in a queue and during t time the system serves no more than one demand, or there are two demands in a queue, and during t time no more than two demands are served, and so on. In this case, according to the formula of full probability, we have

$$\begin{aligned} & q\left[1 - F_W\left(t\right)\right] \\ &= \frac{\Lambda_1}{\Lambda}\left[B_0\left(t\right)\sum_{i=m}^{m+\varepsilon_1-1} P_i + B_1\left(t\right)\sum_{i=m+1}^{m+\varepsilon_1-1} P_i + \cdots \right. \\ & \left. \quad + B_{\varepsilon_1-1}\left(t\right)P_{m+\varepsilon_1-1}\right] \\ & + \frac{\Lambda_2}{\Lambda}\left[\sum_{i=0}^{\varepsilon_1} B_i\left(t\right)\sum_{i=m+\varepsilon_1}^{m+\varepsilon_2-1} P_i + B_{\varepsilon_1+1}\left(t\right)\sum_{i=m+\varepsilon_1+1}^{m+\varepsilon_2-1} P_i + \cdots \right. \\ & \left. \quad + B_{\varepsilon_2-1}\left(t\right)P_{m+\varepsilon_2-1}\right] \\ & + \frac{\Lambda_3}{\Lambda}\left[\sum_{i=0}^{\varepsilon_2} B_i\left(t\right)\sum_{i=m+\varepsilon_2}^{m+\varepsilon_3-1} P_i + B_{\varepsilon_2+1}\left(t\right)\sum_{i=m+\varepsilon_2+1}^{m+\varepsilon_3-1} P_i + \cdots \right. \\ & \left. \quad + B_{\varepsilon_3-1}\left(t\right)P_{m+\varepsilon_3-1}\right] + \cdots \\ & + \frac{\Lambda_h}{\Lambda}\left[\sum_{i=0}^{\varepsilon_{h-1}} B_i\left(t\right)\sum_{i=m+\varepsilon_{h-1}}^{m+\varepsilon_h-1} P_i + B_{\varepsilon_{h-1}+1}\left(t\right)\sum_{i=m+\varepsilon_{h-1}+1}^{m+\varepsilon_h-1} P_i + \cdots \right. \\ & \left. \quad + B_{\varepsilon_h-1}\left(t\right)P_{m+\varepsilon_h-1}\right]. \qquad (6) \end{aligned}$$

After a number of intermediate calculations, it is possible to obtain the following expressions for finite-sums sequence in square brackets in the right-hand side of this ratio. As a result, substituting obtained ratios into the right member of a formula (6), we will finally find

$$F_W(t) = 1 - e^{-m\mu t}\frac{P_{m-1}}{q}$$
$$+\left\{\frac{R_1}{m-R_1}\left[e_{\varepsilon_1-1}(R_1\mu t) - \left(\frac{R_1}{m}\right)^{E_1} e_{\varepsilon_1-1}(m\mu t)\right]\right.$$
$$+\sum_{i=2}^{h}\frac{R_i}{m-R_i}\left[\prod_{g=1}^{i-1}\left(\frac{R_g}{m}\right)^{E_g} e_{\varepsilon_{i-1}-1}(m\mu t)\right.$$
$$+\prod_{g=1}^{i-1}\left(\frac{R_g}{R_i}\right)^{E_g}\left[e_{\varepsilon_i-1}(R_i\mu t) - e_{\varepsilon_{i-1}-1}(R_i\mu t)\right]$$
$$\left.\left.-\prod_{g=1}^{i}\left(\frac{R_g}{m}\right)^{E_g} e_{\varepsilon_i-1}(m\mu t)\right]\right\};$$

Hence, the density of a demand waiting time distribution for service in a queue is

$$f_W(t) = \frac{dF_W(t)}{dt} = e^{-m\mu t}\frac{P_{m-1}}{q}$$
$$\times\left\{\Lambda_1 e_{\varepsilon_1-1}(\Lambda_1 t) + \sum_{i=2}^{h}\Lambda_i\prod_{g=1}^{i-1}\left(\frac{R_g}{R_i}\right)^{E_g}\left[e_{\varepsilon_i-1}(\Lambda_i t) - e_{\varepsilon_{i-1}-1}(\Lambda_i t)\right]\right\} \quad (7)$$

and then, mean waiting time of demand service in a queue is

$$\bar{t}_W = \int_0^\infty t f_W(t)\,dt$$
$$= \frac{1}{\Lambda_0 q}\sum_{i=1}^{h}\left\{\frac{R_i}{m-R_i}\left[P_{Bi} - E_i P_{m+\varepsilon_i}\right] + \frac{R_i}{m}\varepsilon_{i-1}P_{Bi}\right\} = \frac{\bar{l}}{\Lambda}$$

in compliance with J. Littl's formulae. In the same way the second initial moment of a demand waiting time in a queue is

$$\overline{t_W^2} = \int_0^\infty t^2 f_W(t)\,dt$$
$$= \frac{1}{\Lambda_0 q}\sum_{i=1}^{h} R_i\left\{\begin{array}{l} \frac{2\left(P_{Bi}-E_iP_{m+\varepsilon_i}\right)}{\mu(m-R_i)^2}\left[1+\frac{\varepsilon_{i-1}}{m}(m-R_i)\right] \\ \frac{P_m}{3m^2\mu}\prod_{g=0}^{i-1}\left(\frac{R_g}{R_i}\right)^{E_g} \end{array}\right.$$
$$\left.\begin{array}{ll} +\frac{\varepsilon_{i-1}(\varepsilon_{i-1}+1)P_{Bi}}{m^2\mu} - \frac{E_i(E_i+1)P_{m+\varepsilon_i}}{m\mu(m-R_i)}, & R_i \neq m \\ \times\left[\varepsilon_i(\varepsilon_i+1)(\varepsilon_i+2) - \varepsilon_{i-1}(\varepsilon_{i-1}+1)(\varepsilon_{i-1}+2)\right], & R_i = m \end{array}\right\}.$$

Let us note that the ratio (7) gives a possibility to calculate moments of any order as a demand waiting time in a queue for service.

4 Numerical Investigation of Queue Parameters Behavior in QS

In actual conditions of objects operating according to the principle of queuing systems, the problem of queues and delays in service is always topical. It naturally causes desire to organize the process of their exploitation in such a way that the operation of these objects and systems would proceed in more stable modes. It should be borne in mind that a single parameter which could be changed more or less quickly in actual practice for multi-channel devices in practice is the number of homogeneous service facilities m working in parallel. Therefore, we will set the task to study the work of QS in the following way.

Let us investigate the nature of behavior of the moments of queue length and waiting time of the demand in queue with the change of the number of service facilities m. For this purpose, let us formally replace factorial dependences m in formulas for probabilistic characteristics [7] through which the moments of the number of demands in the queue and waiting time are expressed with corresponding gamma-functions $\mathrm{G}(m + 1)$; m is conditionally regarded as a continuous quantity. Dependencies of mathematical expectation and variance of demands number waiting for service in the queue on the number of service facilities show that there is some boundary value m corresponding to a cross point of the moments of demands number in the queue which divides the axis m into two parts. The first part is the area in which the mean squared deviation (MSD) of the queue length is within the limits of mathematical expectation; the second part is the area in which the dispersion of demands number in the queue exceeds the mean value. The system functioning mode at which MSD of the queue length does not exceed its mean value is pretty stable and predictable from the point of view of operation.

In this case it is interesting to trace the dynamics of m change that is boundary when the given intensity components of demands input stream change and the queue length for corresponding components of input stream is limited.

A special program was developed to conduct a series of computational experiments to calculate m boundary according to the mathematical model with known as initial data of given intensity components of the demands input stream and corresponding size-limited queues. Varying the given components intensity of demands input stream ρ_i within 1 to 12, we found values *m1* boundary for the moments of queue length and *m2* boundary for the moments of demand servicing-waiting moments in a queue at various values of step size between queue length limitations for various components of demands input stream $E_i = 1;\ 2;\ 5;\ 10$.

As an example let us consider the queuing model with a two-component demand input stream and two queue length limits for each component. For this purpose let us set $\lambda_0 = 0; \quad \mu = 1; \quad E_0 = 0; \quad h = 2$ in the program. As $\lambda_0 = 0$, the zero (Erlang) component in this model is absent. Here $\varepsilon_1 = E_1$ is queue length limit for demands of the first component of the input stream with the given intensity ρ_1, and $\varepsilon_2 = E_1 + E_2$ is queue length limit for demands of the second component of the input stream with the given intensity ρ_2.

The behavior of *m1* and *m2* boundary with the change of given intensity ρ_1 and ρ_2 is linearly increasing. We will call obtained straight lines limits of stability. Each point lying on the stability boundary corresponds to equal values of mathematical expectation and MSD of the queue length (for *m1*), and waiting time to service the demand in the queue (for *m2*) at a definite value of the given intensity of demand input stream. The coefficients of variation of the queue length and waiting time in the queue are equal to the unity. In fact, it is the border above which MSD exceeds mathematical expectation. The area below the straight line corresponds to the stable mode of system operation at which the mean squared deviation is within mathematical expectation.

When $\rho_1 > 1$ obtained straight lines divide the coordinate plane into 3 areas: the upper one corresponds to an unstable mode of system operation both according to the queue length and waiting time; the middle one corresponds to the stable mode as for the queue and unstable as for waiting time; the lower – to the stable mode on the queue and waiting time as well. It turns out that the set of values of the number of service facilities corresponding to the stable mode of system operation is limited from above by the stability boundary for waiting time. Both straight lines form a multiplicative strip of instability in regard to waiting time; its width enhances upon increasing of the given stream intensity ρ_1.

When the step between queue length limits for demands of different components is $E_1 = E_2 = 2$, stability boundaries on the queue length and waiting time when the given intensity of the first component of the stream is changed ρ_1, form a multiplicative instability strip of the system according to waiting time. In case the given intensity of the second component of the stream changes, ρ_2 form the additive instability strip of the system as for waiting time; its width does not practically change with the increase of ρ_2.

When the step between queue length limits for demands of different components is $E_1 = E_2 = 5$, the further narrowing of instability strips with regard to waiting time both for multiplicative at increase of the given intensity of the first component of stream ρ_1 and additive is observed when the given intensity of the second component ρ_2 changes.

Finally, when the step between queue length limits for demands of different components is $E_1 = E_2 = 10$, instability strips on waiting time practically disappear turning into a single boundary of the stability area both in queue and waiting time as well.

In case of a two-component service model with two queue length limits for each component of the demands input stream with intervals between limits $E_1 \geq 10$ and $E_2 \geq 10$, boundary values of the number of service facilities (inside of which MSD queue lengths and waiting time meet corresponding mathematical expectations) practically coincide. They are approximately equal to the sum of given intensity of all components of demands input stream. Also boundaries of stability on queue length and waiting time are straight lines and at $E_1 \geq 10$ and $E_2 \geq 10$ their slope angle makes 45°.

For a two-component queuing model with queue limits there is an opportunity to investigate behavior of *m1* and *m2* boundary at simultaneous change of the given intensity of both components of demands input stream ρ_1 and ρ_2.

Having conducted a cycle of corresponding computational experiments at the step between queue length limits for demands of different components $E_1 = E_2 = 5$, we obtain hypersurfaces of stability on queue length and waiting time of the demand in a queue, very close to planes.

Obtained hypersurfaces break a coordinate space into 3 parts: upper is the space of system instability on queue and waiting time; low is the space of system behavior stability both on the queue length and demands waiting time in the queue; middle – the layer corresponding to an unstable operation mode of the system only on waiting time.

5 Higher Orders Queues

An N-th order queue will be called the queue calculated in case when there are N claims in the system as minimum, and some of them are in the memory. If $N = 0$ we have a usual mathematical queue, when $N = m$ where m - the number of channels in the service facility, we have a physical queue which is explicitly studied in work [8]. At $N = m + 1$ we have the so-called real queue [5], [6]; at all values $N > m + 1$ we have consequently higher orders queues [9].

Apparently, T. Saaty was the first to state the issue of real queues in his classical monograph [10]; it specified the value for the M/M/m system representing itself as an average number of demands which stay in the queue for some time to be served.

The physical sense of the real queue defined in the above-stated sense is that in this case a newly arrived into the system claim finds busy all service channels (all devices) and, at least, one more claim in the queue waiting for the service. Thus, the minimum mean length of a real queue (in case the intensity of an input stream of claims tends to zero) is unity but not zero as a general and well-studied mathematical queue has. As we see, the real queue is understood as the situation when there is at least one claim in the queue for the service on a par.

However, this numerical characteristic is not the only one to characterize real queues in queuing systems.

Along with real queues in the sense explained above, it is possible to consider another numerical characteristic of QS which, for example, in the standard report of the GPSS simulation system has the name “a queue without zero inputs”. Here, zero input is understood as such arrival of the claim in the system at which there is, at least, one free service channel in the multi-channel device, and in this case the claim is served immediately. Let’s emphasize that unlike the situation considered above, in this case we imply the situation when at the time of a new claim arrival in the system all service channels of the service facility are occupied, but the queue, as such, can be absent. In the latter case, the claim expecting service has no other service waiting claims before it; it is just before the service facility in which all channels are busy at that time. Thus defined the "queue without zero inputs" is calculated considering only those claims which really expected service, and without taking into account claims which did not have to wait as at the time of their arrival in the system one serving channel was free at least. Queue mean length without zero inputs is, apparently, longer than mean length known for all and more habitual mathematical queue but, in its turn, it is less than mean length of the real queue considered above. It is clear, that the minimum mean length of such queue is zero, as well as the usual mathematical queue is, i.e. on average such a queue, as well as a mathematical queue, can have any number of claims.

Thus, if the usual mathematical queue is calculated as the average for all claims which visited the system, then the queue without zero inputs is calculated as the average value minus those claims which were served immediately as they got into the system when, at least, one of service channels was free. The so-called real queue in this case is calculated as the average minus both those claims which were served without a queue, and those ones which found all service channels occupied but were the first in the service waiting list as there were no other claims in the system at this moment. In work [8] it was proposed to call the queues calculated without zero inputs as physical queues.

It is clear, that this result can be generalized if the concept of higher orders queues of systems with queues is introduced in the following way.

Let the queuing system have m serving channels with identical service intensity μ. In this case we will call the queue of a 0-th order the average queue calculated on condition that when a new claim enters the system, there can be any number of claims including the case when there are no claims at all, i.e. the system can be the completely free from claims. In this case we will call the queue of the 1-st order the average queue calculated on condition that when a new claim enters the system, it already contains at least one claim, and so on. It is clear, that upon this the physical queue means an average queue in all those cases that when the claim enters the system, there are at least m claims in it; thus according to this nomenclature, the physical queue is a queue of the m-th order, Then the real queue is a queue of the $m + 1$-th order, etc.

Thus, the N-th order queue is the average queue calculated on condition that when a new claim enters the system there are already Nclaims in it, and some of them can be in the memory. At the same time the case $N = 0$ corresponds to a usual mathematical queue; for $N = m$ we have a physical queue; let us remind that in the system of GPSS World simulation modeling this characteristic has the name "a queue without zero inputs". In case $N = m + 1$ we have a real queue; for those cases when $N > m + 1$ we have higher orders queues. In case all serving channels are busy, a newly arrived claim will have to expect service, the minimum quantity of claims in the physical queue is equal to zero in the memory; for the real queue it is equal to unity, and so on. It should be noted that physical and real queues have the greatest deviations from the known mathematical queue at small values of the intensity of claims stream entering the system.

As it is known, mean processing time of one claim in the system $\bar{t}_S$, mean staying time of claims in the queue $\bar{t}_W$ and the common mean staying time of the claim in the system in general $\bar{t}_T = \bar{t}_S + \bar{t}_W$ for Markov queuing systems are bound to corresponding discrete characteristics of QS by the following three formulas [5,6]:

$$\bar{t}_S = \bar{n}/A; \quad \bar{t}_W = \bar{l}/A; \quad \bar{t}_T = \bar{k}/A; \tag{8}$$

where A is throughout capacity of the system, i.e. an average number of claims served by the system in unity of time. Discrete characteristics of the system are understood respectively as an average number of busy channels $\bar{n}$, mean length of the queue $\bar{l}$, and an average number of claims in the system in general $\bar{k} = \bar{n} + \bar{l}$. Sometimes, these formulas are written in the form

$$\bar{t}_S = \bar{n}/\lambda; \quad \bar{t}_W = \bar{l}/\lambda; \quad \bar{t}_T = \bar{k}/\lambda,$$

when the total intensity of claims stream λ coming into the system is in the denominator.

In fact, however, the denominator of these formulas should not be made of the total intensity of claims stream but of that part only which corresponds to those claims that are really transferred through the system (more precisely, through the service facility), i.e. absolute throughout capacity of the system A.

Formulas (8) are commonly called Little's formulas. At first, the result which engineers used for a long time existed as several empirical formulas, i.e. in the form of some kind of "folkloric theorem", as it is said. Apparently, J. D. C. Little was the first person who gave it a strict formulation in 1961. The intuitive proof of Little's formulas comes to the fact that in a steady state mode the next demand entering the system finds in it the same average number of demands which remains in the system when this demand leaves it. This quantity is just equal to the product of claims stream intensity transferred through the system (or its any subsystem) multiplied by the mean time of their staying in this system (subsystem):

$$\bar{n} = A\,\bar{t}_S; \quad \bar{l} = A\,\bar{t}_W; \quad \bar{k} = A\,\bar{t}_T. \tag{9}$$

Direct mechanical analog of formulas (9) is a well-known relation for the way passed at a steady movement s based on moving velocity v and travel time t.

$$s = v\,t.$$

The case is somewhat different with QS numerical characteristics concerning a real queue and higher orders queues in these systems. Let us remind that the N-th order queue we have called the average queue calculated on condition that when a new claim enters the system there are already $2N$ claims in it, and some of them can be in the memory.

At the same time $N = 0$ corresponds to a usual mathematical queue; for $N = m$ we have a physical queue which in the system of GPSS World simulation modeling has the name "a queue without zero inputs".

In case $N = m + 1$ we have a real queue; for those cases when $N > m + 1$ we respectively have higher orders queues.

For a physical queue, as it is shown in work [8], the corresponding ratio has the form quite similar to (8):

$$\bar{t}_{Wphys} = \bar{l}_{phys}/A \tag{10}$$

It is possible to ascertain that the relation (10) is applicable for all types of queues from a mathematical to a physical queue, including the latter one, however for a real queue and higher orders queues this formula becomes unfair.

Somewhat different is the situation with numerical characteristics of QS concerning real queues and higher orders queues in regard to a real queue in these systems. In works [5,6] it was found out that the following relation is performed for the systems of M/M/m and M/M/m/E classes (however, all numerical characteristics of the first ones can be obtained by ultimate passing from numerical characteristics of the second ones)

$$\bar{t}_{Wreal} = \bar{l}_{real}/m\mu \tag{11}$$

as the real queue moves with velocity $m\mu$ to serve demands by the multi-channel device. It is possible to show that the same dependence will remain fair for all types of higher orders queues for which $N > m + 1$:

$$\bar{t}_{WN} = \bar{l}_{N}/m\mu \tag{12}$$

Relations (8)–(12) connect parameters of usual mathematical, physical and real queues in open queuing systems and parameters of higher orders queues in these systems as well. It is clear that these relations will be absolutely similar for close-loop queuing systems. At the same time the obtained system of formulas (8)–(12) may be called as *generalized Little*'s *formulas.*

As we see, all higher orders queues in queuing systems of various types from the point of view of claims traveling velocity in these queues can be divided into two unequal classes. In this case, the first class will include all types of queues from mathematical to physical inclusive, which move with a transferring velocity of claims through system A. Thus $m + 1$ types of queues of various orders from zero to m-th are in the first class. The second, a more extensive class, includes a real queue and all higher orders queues in regard to a real queue for which, according to the definition, we have $N > m + 1$. All these queues move with the service velocity $m\mu$. The number of queues of various orders in this class is not limited.

Further, the work [8] provides formulas obtained for the mean length of a physical queue for queuing systems of various types. In particular, the expression for a mean length of a real queue of the system with an unlimited memory volume (within M. Kendall's symbolism – M/M/m model) is the following

$$\bar{l}_{phys} = \frac{\rho}{m - \rho}.$$

But for the M/M/m model $A = \lambda$, and then according to formulas (10) and (11) we have

$$\begin{aligned}\bar{t}_{Wphys} &= \tfrac{\bar{l}_{phys}}{\lambda} = \tfrac{1}{\mu(m-\rho)};\\ \bar{t}_{Wreal} &= \tfrac{\bar{l}_{real}}{m\mu} = \tfrac{1}{\mu(m-\rho)}.\end{aligned}$$

i.e. for the model with an unlimited queue the mean staying time of one claim in a physical queue coincides with the mean staying time of the claim in a real queue: $\bar{t}_{Wphys} = \bar{t}_{Wreal}$. The obtained result can be called the theorem on physical and real queues in queuing systems with an unlimited memory volume.

6 Conclusion

Generalizing data of all computational experiments submitted in the work it is possible to draw the following conclusion.

In queuing systems of multicomponent streams stable operation modes of the system on the queue length and waiting time of demands are possible. Boundaries of these modes correspond to single coefficients of queue length variation and demands servicing-waiting in system. Regardless the number of components in demands input stream and values of the step between queue length limits for various components of the stream, boundary values of the number of service facilities depending on the given intensity of various stream components form straight lines described by the equation $m\left(\rho_i\right) = a + b\rho_i$ where ρ_i- given intensity of the i-th component of demands input stream. When the step between queue length limits for various components of demands input stream is $E_i \geq 10$, coefficients a and b accept values $a = \sum\limits_{j \neq i} \rho_j$, $b = 1$. Thus, at $E_i \geq 10$ the boundary value of the number of service facilities is numerically equal to the sum of the given intensity of all input stream components. If above this limit, the operation mode of the system will be unstable both on the queue length and demand waiting time.

The proposed results of the work can be used to project and operate quite a wide class of objects and systems to assess their efficiency, and also to develop projects of modernization or construction of various technical objects working according to the principle of queuing systems.

References

1. Cohen, J.W.: Certain delay problems for a full availability trunk group loaded by two sources. Commun. News **16**(3), 105–113 (1956)
2. Takagi, H.: Explicit delay distribution in first-come first-served M/M/m/K and M/M/m/K/n queues and mixed loss-delay system. Int. J. Pure Appl. Math. **40**(2), 185–200 (2007)
3. Kirpichnikov, A.P., Titovtsev, A.S.: Open systems of multicomponent flows differentiated service. Ciência e Técnica Vitivinícola **29**(7), 108–122 (2014)

4. Titovtsev, A.: Sistemy differentsirovannogo obsluzhivaniya polikomponentnykh potokov. Modeli i kharakteristiki [Systems of differentiated services multicomponent flows. Models and specifications]. LAP LAMBERT Academic Publishing GmbH & Co. KG Publ., Saarbrücken (2012). (in Russian)
5. Kirpichnikov, A.P.: Prikladnaya teoriya massovogo obsluzhivaniya [Applied queuing theory]. Publishing office of KSU Publ., Kazan (2008). (in Russian)
6. Kirpichnikov, A.P.: Metody prikladnoy teorii massovogo obsluzhivaniya. Publishing office of KSU Publ., Kazan (2011). (in Russian)
7. Kirpichnikov, A., Titovtsev, A.: Mathematical model of a queuing system with arbitrary quantity of sources and size-limited queue. Int. J. Pure Appl. Math. **106**(2), 649–661 (2016)
8. Kirpichnikov, A., Titovtsev, A.: Physical and mathematical queues in the applied queuing theory. Int. J. Pure Appl. Math. **108**(2), 409–418 (2016)
9. Titovtsev, A.: The concept of higher orders queues in the queuing theory. Int. J. Pure Appl. Math. **109**(2), 451–457 (2016)
10. Saaty, T.L.: Elements of Queueing Theory with Applications. McGRAW-HILL book company Inc., New York, Toronto, London (1961)