UNIVERSITÀ DEGLI STUDI DI SALERNO

DIPARTIMENTO DI MATEMATICA

DOTTORATO DI RICERCA IN MATEMATICA
XIV CICLO - NUOVA SERIE

Performance analysis of queueing systems with resequencing

CANDIDATO: **Caraccio Ilaria**

COORDINATORE: **Chiar.ma Prof.ssa Patrizia Longobardi**

TUTOR: **Chiar.mo Prof. Ciro D'Apice**

COTUTOR: **Prof.ssa Rosanna Manzo**

ANNO ACCADEMICO 2014-2015

Contents

List of Figures

Background, literature review and motivation

The service sector lies at the heart of industrialized nations and continues to serve as a major contributor to the world economy. Over the years, the service industry has given rise to an enormous amount of technological, scientific, and managerial challenges. Among all challenges, operational service quality, service efficiency, and the tradeoffs between the two have always been at the center of service managers' attention and are likely to be so more in the future. Queueing theory attempts to address these challenges from a mathematical perspective. Every service station of a queueing network is characterized by two major components: the external arrival process and the service process. The external arrival process governs the timing of service request arrivals to that station from outside, and the service process concerns the duration of service transactions in that station. These are then fused with a routing process among stations to form the structure of the queueing network. Since the arrival, service, and routing processes are usually stochastic by nature, the study of service networks involves probabilistic analysis, which is the subject of queueing theory. Many distributed applications, such as voice data transmission, remote computations, and database manipulations, information integrity require that data exchanges between different nodes of a system be performed in a specific order. Recently, multipath routing has received some attention in the context of both wired and wireless communication networks. By sending data packets along different paths, multipath routing can potentially help balance the traffic load and reduce congestion levels in the network, thereby resulting in lower sojourn time. Under multipath routing, since consecutive packets travel possibly along different paths from source to destination, they can easily be received out-of-sequence at the destination. If the application requires packets to be processed in the order in which they were sent, then disordered packets have to wait an additional amount of time, known as the resequencing delay, before being consumed. Packet mis-ordering occurs in the

following two transmission scenarios. In the first scenario, multiple (or parallel) routes between the transmitter-receiver pair are utilized to send data packets to increase the data transmission rate. However, a packet transmitted along one route may experience a time delay that is different from that along another route. Consequently, a packet that was sent by the transmitter earlier than another may arrive at the receiver later, resulting in packet mis-ordering at the receiver end. In the second scenario, packets may be lost or erroneously received due to channel degradation, congestion or any network hardware malfunction along a route, in which case they have to be retransmitted for error-free data transmissions via a retransmission scheme, such as the selective repeat automatic repeat request protocol (SR-ARQ). Retransmission of corrupted or lost packets can cause packets to be received out-of-order at the receiver as well. Note that the second scenario happens when there is one single channel between the transmitter and the receiver. In practice, many applications require that the packets are received in the same order from which they were sent. For such applications, the receiver has to buffer the mis-ordered packets in a resequencing buffer, resequence them repeatedly, and deliver them in the corrected order. This process is referred to as packet resequencing. The resequencing issue in simultaneous processing systems, where the order of customers (jobs, units, etc.) upon arrival has to be preserved upon departure, is a crucial theme in the queueing theory. Queueing-theoretic approach to resequencing problem implies that the system under consideration is represented as interconnected queueing systems/ networks. Various analytical methods and models have been proposed to study the impacts of resequencing. A general survey of queueing theoretic methods and early models for the modeling and analysis of parallel and distributed systems with resequencing can be found in [7]. Survey on the resequencing problem that covers period up to 1997 can be found in [8]. In [1] a continuous-time M/M/2 queueing system, with two heterogeneous servers, a *routing policy with variable routing position*, is analyzed with the objective of minimizing the sum of the queueing delay and resequencing delay and of comput-

ing the total expected end-to-end delay (including the resequencing delay). In [2] the effect of fixed delay on the optimal traffic split is studied for a continuous-time system of two end nodes with two parallel M/M/1, in order to minimize the total end to end delay (including the resequencing delay) in a high speed environment. In [3] a continuous-time 2-M/M/1 network is considered and the asymptotic expression of the probability that there are n packets in reordering buffer as n became large is computed, in order to avoid a reduced data throughput caused by overflow of the resequencing queue, a large enough buffer size of the resequencing queue has to be configured. In [4] a distributed system consisting of two parallel heterogeneous single server M/M/1 queues is analyzed. It is assumed that a total number of C different classes arrive at the source node. The resequencing delay when $C = 1$ is evaluated, and the result is then extended to the case of a single class with interfering traffics (that is an additional stream of customers), and in the case of two-class and multiple class systems. In [5] a M/M/2 system is considered, in which servers are parallel, heterogeneous and exponential and the customers are released from the system after service completion according to their arrival order. The customers, which are delayed due to resequencing, have to wait in a resequencing queue. The attention is limited to fixed-position routing policies which route customers to server 1 only from the head of queue Q, and to server 2 only from a fixed position J, $J \geq 2$, where position J means the Jth customer among those in server 1 and in queue Q. The existence of an optimal stationary policy is shown: the faster service is kept active as long as the service queue is not empty. In [6] a virtual circuit from node S to node D connected by m links in parallel, whose arrival process is general and the service times are exponentially distributed, is investigated. An important property of a virtual circuit is that it delivers packets at the destination in the same sequence as they are received at the source. A packet arriving at S has to wait if all links are busy. The distribution of the total delay for the G/M/m queueing system model, the distribution of the resequencing delay for the G/M/m queueing system model, the expectation of the

resequencing delay for the G/M/m queueing system model, the distribution of the total delay for the M/M/m, the M/H_K/∞ and the G/M/∞ queueing system model is obtained. The resequencing has also been studied in system in which the arrivals follow a more complicated process: the Markovian arrival process (MAP). In [9] a MAP/M/2/K queueing model in which messages should leave the system in the order in which they entered into the system is considered. In the case of infinite resequencing buffer, the steady-state probability vector is shown to be of matrix-geometric type. The total sojourn time of an admitted message into the system is shown to be of phase type. Efficient algorithmic procedures for computing various performance measures are given. In [10] a two-server finite capacity queuing model in which messages should leave the system in the order in which they entered the system, is studied. Messages arrive according to a Markovian arrival process and any message finding the buffer full is considered lost. Out-of-sequence messages are stored in a (finite) buffer and may lead to blocking when a processed message cannot be placed in the buffer. Using matrix-analytic methods, the system is analyzed in steady state. It is shown that the stationary waiting time distributions of an admitted message in the queue and in the system as well as the time spent in the service facility follow phase-type distributions. The departure process is characterized as a Batch Markovian Arrival Process. The system performance measures such as system idle probability, server idle and server blocking probabilities, throughput, mean number of messages in primary and in resequencing buffers, rate of departure, average batch size of departure are derived analytically. In [15] a model where the disordering is caused by multipath routing is analyzed. Packets are generated according to a Poisson process. Then, they arrive at a disordering network (DN) modeled by two parallel M/M/1 queues, and are routed to each of the queues according to an independent Bernoulli process. A resequencing buffer follows the DN. In such a model, the packet resequencing delay is known. However, the size of the resequencing queue (RSQ) is unknown. The probability for the large deviations of the queue size is analyzed.

Other systems with resequencing have been studied in [16], [17], [18], [20], [19].

Purpose of the thesis

The main objective of this research is to find the stationary characteristics of M/M/3/∞, M/M/∞/∞ and MAP/PH/2/∞ queueing systems with reordering buffer of infinite capacity.
In the M/M/3/∞ customers in reordering buffer may form two separate queues and focus is given to the study of their size distribution. These two queues are labeled as queue 1 and queue 2. In queue 1 there are customers that are waiting for two customers that are still in service, while in queue 2 there are customers that are waiting for one customer that is still in service. Expressions for joint stationary distribution are obtained both in explicit form and in terms of generating functions. When the parameter of service μ is equal to one and the parameter of arrival λ is between 0.1 and 2.5, numerical examples are given for the mean number of customers in reordering buffer (RB) (queue 1 and queue 2), for the variance of number of customers in RB (queue 1 and queue 2), the coefficient of correlation between queue 1 and queue 2, between queue 1 and RB, between queue 2 and RB.
In the M/M/∞/∞ we propose a new problem statement for systems with resequencing that are modeled by multiserver queues followed with infinite resequencing buffer. Focus is given to the study of joint stationary distribution of the total number of customers in queue and total number of customers in reordering buffer. Using developed analytical methods there was obtained the system of equilibrium equations which allows recursive computation of joint stationary distribution of the total number of customers in buffer and servers and total number of customers in RB.
In MAP/PH/2/∞ we have a queueing system with 2 servers, in which the capacity of the collecting buffer and the reordering buffer is infinite. The type distribution of both two servers is "the phase distribution" (PH), while the arrivals follow Markovian arrival process. We introduce a recurrent algorithm to calculate the simultaneous stationary distribution of the number of the requests at servers, in the collecting buffer and in the reordering buffer. The

stationary distribution of the arrival time in the system and in the reordering buffer are calculated with the Laplace-Stieltjes transform.

Chapter 1

About queueing system

For an accurate description of a queueing system, we need to provide its following basic elements:

1. The input process. It refers to the arrivals to the system. It describes the distribution and dependencies of the interarrival times. The most common input process is the Poisson process.

2. The service mechanism. The basic characteristics of the service mechanism include the number of parallel servers, their identity (homogeneous or heterogeneous, their service speed etc.) and the distribution and dependencies of the service times.

3. The system capacity. It concerns the number of customers that can wait at any given time in a queueing system.

4. The queueing discipline. It is the rule followed by the server(s) for choosing customers for service. The most common queue disciplines are the "first-come, first-served" (FCFS), the "last-come, first-served" (LCFS), and the "service in random order" (SIRO). There are many other queueing disciplines which have been introduced for the efficient operation of computers and communication systems.

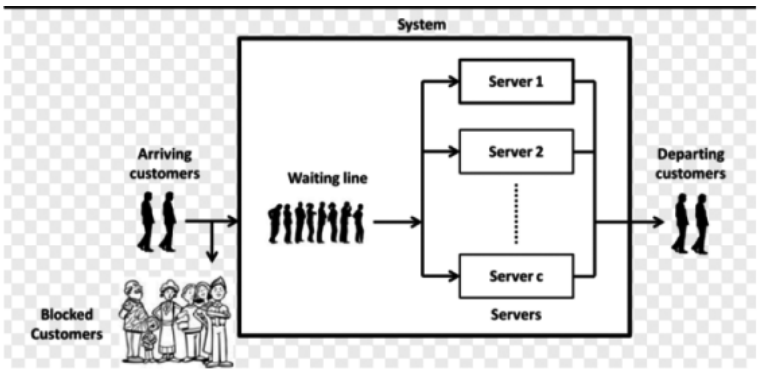


Figure 1.1 A queueing system.

The basic classification-notation that is currently used in queueing theory was introduced by Kendall. According to Kendall's notation, a queue is described by a sequence of five letter combinations - numbers A/B/s/c ( ): input process/service times/number of servers/capacity (discipline). For instance, M is used for exponential (memoryless-Markovian), D for constant (deterministic), Ek for Erlang-k, G or GI for general (independent) interarrival-service times, MAP (Markovian arrival process) and PH phase distibution in the positions A and B of Kendall's notation. In the context of a queueing system there are several processes that concern customers in system:

1. Queue Length Process $\{N(t)\}$: $N(t)$ denotes the number of customers in system at time t, $t > 0$.

2. The Sojourn Time Process is the time from the customer's arrival till his departure.

3. The Waiting Time is the time from customer's arrival till the beginning of service.

Under certain conditions a stochastic process may settle down to what is commonly called steady state or state of equilibrium, in which its distribution properties are time-independent. In this work of thesis we have studied three different systems in steady

state condition: M/M/3/∞, M/M/N/∞ and MAP/PH/2/∞. The arrival and the service processes presented in the first two chapters are well known and are often used in literature. More attention should be paid to the third queueing system in which the arrival process is Markovian and the service process follows a PH distribution. In order to better understand the last system, we introduce these two types of process [14].

First we describe the Markovian arrival process (MAP): let $\nu(t)$ be the number of customers arriving over the time interval $[0, t)$ and $\tau_1, \tau_2, ...$,be the instants of their arrivals. We assume that there also exists a Markov process $\{\xi(t), t \geq 0\}$ defined on the finite state set $I = \{1, 2, ..., l\}$. We assume that $\eta(t) = (\xi(t), \nu(t))$. The process state set $\{\eta(t), t \geq 0\}$ is representable as $\cup_{k=0}^{\infty} I_k$ where $I_k = \{(i, k), i = 1, ..., l\}, k \geq 0$. Therefore, the process $\{\eta(t), t \geq 0\}$ is in the state $(i, k), i = 1, ..., l,\ k \geq 0$, if k customers arrived at the instant t and the process $\{\xi(t), t \geq 0\}$ at time t is in the state i. The customer flow $\{\tau_j, j \geq 1\}$ will be said to be the Markov flow (relative to the process $\{\xi(t), t \geq 0\}$) if the random process $\{\eta(t), t \geq 0\}$ is a homogeneous Markov process and its matrix A of transition intensities is of the block form

$$A = \begin{pmatrix} \Lambda & N & 0 & 0 & . & . \\ 0 & \Lambda & N & 0 & . & . \\ 0 & 0 & \Lambda & N & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \end{pmatrix}$$

where Λ and N are square matrices of the order l. We note that $\Lambda + N$ is the matrix of transition intensities of the Markov process $\{\xi(t), t \geq 0\}$. Obviously, for $j \neq m$ the elements Λ_{jm} of the matrix Λ define the transition intensities of the process $\{\eta(t), t \geq 0\}$ which are not related with customer arrivals, and the elements N_{jm} of the matrix N are the transition intensities accompanied by arrivals of customers. Understandably, if $l = 1, \Lambda_{11} = -\lambda$ and $N_{11} = \lambda$, then we get the ordinary Poisson flow. It is known that if $\{\xi(t), t \geq 0\}$ is a stationary Markov process, then the Markov flow generated by the process $\{\eta(t), t \geq 0\}$ is stationary.

The PH-distribution can describe both the recurrent arrival flow and the customer service times, in our case, we will discuss the phase-type service time. The idea of fictitious phases belongs to A.K. Erlang who used them to Markovize the Erlang distribution. We present a brief description of the main notions for the PH-distributions. The distribution function $F(x)$ of a non-negative random variable is called the phase-type distribution or PH-distribution if it is representable as $F(x) = 1 - \vec{f}^T e^{-Gx}\vec{1}, x > 0$ where $\vec{f}$ is the m-dimensional vector for which $\sum_{j=1}^{m} f_j \leq 1$, $f_j \geq 0, j = 1, ..., m$ and G is $m \times m$ matrix for which $\sum_{j=1}^{m} G_{ij} \leq 0; G_{ij} \geq 0, i \neq j; G_{ij} < 0, i, j = 1, ..., m$, and at least for one i, $\sum_{j=1}^{m} G_{ij} < 0$. The pair $(\vec{f}, G)$ is called the PH-representation of the order m of the distribution function $F(x)$. The distribution function of the PH type admits probabilistic interpretation based on the concept of phase. Let $\nu_1, ..., \nu_m$ be some real numbers, $\nu_i \geq -G_{ii}, i = 1, ..., m$, the numbers $\theta_{ij}, i, j = 1, ..., m$, obey the formula

$$\theta_{ij} = \begin{cases} 1 + \frac{G_{ii}}{\nu_i}, & if\ i = j; \\ \frac{G_{ii}}{\nu_i}, & if\ i \neq j. \end{cases}$$

Then $\sum_{j=1}^{m} \theta_{ij} \leq 1, \theta_{ij} \geq 0, i, j = 1, ..., m$. Let us consider now an open queueing network consisting of m nodes where at most one customer sojourns at each time instant, that is, the arriving flow is blocked if there is a customer in the network. The arriving customer is sent to the node $i, i = 1, ..., m$, with probability f_i and with the complementary probability $f_0 = 1 - \sum_{j=1}^{m} f_j$ immediately departs from the network by passing all nodes. The time of customer service in the node i is distributed exponentially with the parameter ν_i. Upon leaving the node i, the customer travels to the node $j, j = 1, ..., m$, with probability θ_{ij} and with the complementary probability $\theta_{i0} = 1 - \sum_{j=1}^{m} \theta_{ij}$ departs from the network.

Chapter 2

Three-server queueing system with poisson input and exponential service times

2.1 Problem statement

We consider the M/M/3/∞ queueing system (QS) with three servers, infinite capacity buffer, incoming Poisson flow of customers (of intensity λ) and exponential distribution of service time at each server (with parameter μ) and resequencing buffer (RB) of infinite capacity. Customer in reordering buffer may form two separate queues. The most convenient way to explain how queues are separated in resequencing buffer is giving an example. Consider a queueing system with three servers, infinite capacity main buffer and reordering buffer. Let the state of the system at some instant be as depicted in Fig. 1. In squares one can see customers' sequential numbers. White (black) squares in Fig. 1 mean that customers with these sequential numbers have received (have not yet received) service. Here one can distinguish two queues: one which is formed by customers awaiting customer n. 18 (queue #1), another is formed by customers awaiting customer n. 15 (queue

#2). Three cases need to be considered.

1. If customer n. 21 is next to complete its service then it joins queue 1 and stays there until service of customer n. 18 is complete. Customer n. 22 joins idle server.

2. If customer n. 15 is next to complete its service then it goes through queue 1 without waiting and joins queue 2. Meanwhile customer n. 22 joins idle server. As there is no customer in the system with sequential number smaller than any sequential number in queue 2, then all customers from queue 2 leave the system. Resequencing buffer "sees", that queue 2 is empty and moves its contents to queue 2. Now there are three options.

   (a) if customer n. 18 is next to complete service, then it goes through queue 1 without waiting and joins queue 2. Customer n. 23 joins idle server. Again there is no customer in the system with sequential number smaller than any sequential number in queue 2. Thus all customers from queue 2 leave the system. Resequencing buffer becomes empty. Now if customer n. 21 is next to complete service, it leaves the system. If customer n. 22 is next to complete service, it goes through queue 1 without waiting and joins queue 2 where it waits for customer n. 21. Finally, if customer n. 23 is next to complete service, it joins queue 1 and does not proceed to queue 1 because it needs customer n. 22 to complete its service before both of them may join queue 2.

   (b) if customer n. 21 is next to complete service, then it goes through queue 1 again without waiting, joins queue 2 and waits there with other customers for service completion of customer n. 18.

   (c) if customer n. 22 is next to complete service, then customer n. 23 joins idle server, customer n. 22 joins queue 1 and stops there because "sees" gap between

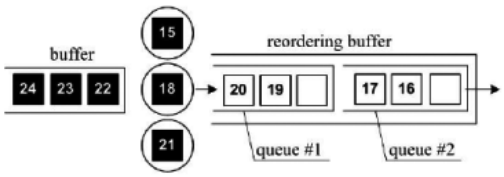
its sequential number and largest sequential number in queue 2. It waits there for customer n. 21.

3. If customer n. 18 is the first to complete its service then it joins queue 1 and customer n. 22 joins idle server. Resequencing buffer "sees", that there is no gap in the middle of sequence and moves the content of queue 1 to queue 2 (queue 1 becomes empty). Now there are again three options.

   (a) if customer n. 15 is next to complete service, then it goes through queue 1 without waiting, joins queue 2 and immediately (because the sequence is complete) leaves the system with all other customers of queue 2.

   (b) if customer n. 21 is next to complete service, then it goes through queue 1 again without waiting, joins other customers in queue 2 that wait for service completion of customer n. 15.

   (c) if customer n. 22 is next to complete service, then it joins queue 1 and stops there, because "sees" gap between its sequential number and the largest sequence number in queue 2. The operation of the system proceeds along the line.



Scheme of the model

Figure 2.1 Scheme of the model.

Clearly, when the number of server is n there are (n-1) queues in resequencing buffer. The sum of customers in these (n-1) is the total number of customers in resequencing buffer.

The main contribution of this research are algorithm and probability generating function of joint stationary probabilities of the number of customers in buffer, queue 1 and queue 2.

2.2 Model description and notation

Customers upon entering the system obtain sequential number and join buffer. Without loss of generality we suppose that the sequence starts from 1 and coincides with the row of natural numbers, i.e. the first customer upon entering the (empty) system receives number 1, the second one number 2 and so on and so forth. Customers leave the system strictly in order of their arrival (i.e. in the sequence order). Thus after customer's arrival it remains in the buffer for some time and then receives service when one of the servers becomes idle. If at the moment of its service completion there are no customers in the system or all other customers present at that moment in the queue and the rest two servers have greater sequential numbers, it leaves the system. Otherwise it occupies one place in the RB.

Customer from RB leaves it if and only if its sequential number is less than sequential numbers of all other customers present in system. Thus customers may leave RB in groups.

Let us call "1^{st} level" customer the one which is in service and was the last to enter server; "2^{nd} level" customer is the one which is in service and was the penultimate to enter server; finally, "3^{rd} level" customer is the one which is in service and was the first to enter server. If the number of busy servers is 3, then customers that entered RB between "1^{st} level" and "2^{nd} level" customer form queue #1; customers which entered RB between "2^{nd} level" and "3^{rd} level" customer form queue #2. If the number of busy servers is 2, then customers which entered RB after "1^{st} level" customer form queue #1; customers which entered RB between "1^{st} level" and "2^{nd} level" customer form queue #2. When there is only one busy server all customers in RB form queue #2.

The operation of the considered queueing system can be completely described by a Markov process $\zeta(t) = \{(\xi(t), \eta(t), \upsilon(t)), t \geq 0\}$ with three components: $\xi(t)$ - number of customers in buffer and server at time t, $\eta(t)$ - number of customers in queue #1 of RB at time t, $\upsilon(t)$ - number of customers in queue #2 of RB at time t. In case $\xi(t) = 0$, the second and third component of $\zeta(t)$ are omit-

ted; in case $\xi(t) = 1$, the second is omitted. The state space of $\zeta(t)$ is $\chi = \{0\} \cup \{(1,i), i \geq 0\} \cup \{(n,i,j), n \geq 2, i \geq 0, j \geq 0\}$. Henceforth it is assumed that service and arrival processes are mutually independent and necessary and sufficient condition of stationarity $\frac{\rho}{3} < 1$, where $\rho = \frac{\lambda}{\mu}$, holds for the system.

Note that the total number of customers in servers and buffer of the considered QS with resequencing coincides with the total number of customers in M/M/3/∞ queue. Therefore, its stationary distribution $\{p_i, i \geq 0\}$, has the form:

$$p_0 = \left(\sum_{i=0}^{2} \frac{\rho^i}{i!} + \frac{\rho^3}{2!(3-\rho)} \right)^{-1}, \tag{2.2.1}$$

$$p_i = \frac{\rho^i}{i!} p_0, \quad i = 1, 2, 3, \tag{2.2.2}$$

$$p_i = \frac{\rho^i}{3! 3^{i-3}} p_0 = \tilde{\rho}^{i-3} p_3, \tilde{\rho} = \frac{\rho}{3}, i \geq 4. \tag{2.2.3}$$

Provided that RB is empty when servers are idle, p_0 is also the probability, of the considered system with resequencing, to be empty.

Lets denote by $p_{n;i,j}$, $n \geq 3, i \geq 0, j \geq 0$, stationary probability of the fact that there are n customers in servers and buffer, i customers in queue #1 of RB, j customers in queue #2 of RB. By $p_{n;i}$, $n \geq 3, i \geq 0$, denote stationary probability of the fact that there are n customers in servers and buffer and i customers in queue #1 of RB. Clearly $p_{n;i} = \sum_{j \geq 0} p_{n;i,j}$. Probabilities $p_{2;i,j}$, $i \geq 0, j \geq 0$ and $p_{2;i,}$, $i \geq 0$, are defined by analogy. Finally, let $p_{1;i}$, $i \geq 0$, be stationary probability of the fact that there is only one busy server and i customers reside in queue #2 of RB. Note that distribution p_n, $n \geq 0$, of the total number of customers in servers and buffer (which is defined by 2.2.1-2.2.3) can be expressed as follows:

$$p_1 = \sum_{i \geq 0} p_{1;i}, \qquad p_n = \sum_{i \geq 0} \sum_{j \geq 0} p_{n;i,j}, \quad n \geq 2.$$

2.3 The equilibrium state distribution

The system of equilibrium equations is composed by the following 12 equations:

$$(\lambda + 3\mu)\, p_{n;0} = \lambda p_{n-1;0} + 2\mu p_{n+1}, \quad n \geq 3 \tag{2.3.4}$$

$$(\lambda + 3\mu)\, p_{n;i} = \lambda p_{n-1;i} + \mu p_{n+1;i-1}, \quad n \geq 3, i \geq 1 \tag{2.3.5}$$

$$(\lambda + 2\mu)\, p_{2;0} = \lambda p_1 + 2\mu p_3, \tag{2.3.6}$$

$$(\lambda + 2\mu)\, p_{2;i} = \mu p_{3;i-1}, \quad i \geq 1 \tag{2.3.7}$$

$$(\lambda + \mu)\, p_{1;0} = \lambda p_0 + \mu p_{2;0}, \tag{2.3.8}$$

$$(\lambda + \mu)\, p_{1;i} = \mu p_{2;i} + \mu \sum_{j=0}^{i-1} p_{2;i-j-1,j}, \quad i \geq 1 \tag{2.3.9}$$

$$(\lambda + 3\mu)\, p_{n;0,0} = \lambda p_{n-1;0,0} + \mu p_{n+1;0}, \quad n \geq 3 \tag{2.3.10}$$

$$(\lambda + 3\mu)\, p_{n;0,j} = \lambda p_{n-1;0,j} + \mu p_{n+1;j} + \mu \sum_{k=0}^{j-1} p_{n+1;k,j-k-1}, \quad n \geq 3, j \geq 1 \tag{2.3.11}$$

$$(\lambda + 3\mu)\, p_{n;i,j} = \lambda p_{n-1;i,j} + \mu p_{n+1;i-1,j}, \quad n \geq 3, i \geq 1, j \geq 0 \tag{2.3.12}$$

$$(\lambda + 2\mu)\, p_{2;0,0} = \lambda p_{1;0} + \mu p_{3;0}, \tag{2.3.13}$$

$$(\lambda + 2\mu)\, p_{2;0,j} = \lambda p_{1;j} + \mu p_{3;j} + \mu \sum_{k=0}^{j-1} p_{3;k,j-k-1}, \quad j \geq 1 \tag{2.3.14}$$

$$(\lambda + 2\mu)\, p_{2;i,j} = \mu p_{3;i-1,j}, \quad i \geq 1, j \geq 0. \tag{2.3.15}$$

Now we describe them considering the transition diagram:

- $(\lambda + 3\mu)\, p_{n;0} = \lambda p_{n-1;0} + 2\mu p_{n+1}, \quad n \geq 3$. Analyze the state $(n; 0)$ and the rate in and the rate out of this state. State $(n; 0)$ means that there are 3 busy servers, n customers between buffer and service, 0 customers in queue 1 and we don't know how many customers there are in queue 2. Rate out of this state is $\lambda + 3\mu$ because system can exit it either

through arrival or through service. System can enter this state either: 1) by an arrival if all servers are busy, there are $n-1$ customers between buffer and service and there are no customers in queue 1, 2) by service if all servers are busy, there are $n+1$ customers in buffer and servers, and the 2^{nd} or the 3^{th} level customer could be served. Equating rate-in and rate-out we get equation (2.3.4).

- $(\lambda + 3\mu)\, p_{n;i} = \lambda p_{n-1;i} + \mu p_{n+1;i-1}, \quad n \geq 3, i \geq 1$. Analyze the state $(n;i)$ and the rate in and the rate out of this state. State $(n;i)$ means that there are 3 busy servers, n customers between buffer and service, i customers in queue 1 and we don't know how many customers there are in queue 2. Rate out of this state is $\lambda + 3\mu$ because system can exit it either through arrival or through service. System can enter this state either: 1) by an arrival if all servers are busy, there are $n-1$ customers between buffer and service and there are i customers in queue 1, 2) by service if all servers are busy, there are $n+1$ customers in buffer and servers, $i-1$ customers in queue 1, and the 1^{st} level customer is served. Equating rate-in and rate-out we get equation (2.3.5).

- $(\lambda + 2\mu)\, p_{2;0} = \lambda p_1 + 2\mu p_3$. Analyze the state $(2;0)$ and the rate in and the rate out of this state. State $(2;0)$ means that there are 2 busy servers, 0 customers in queue 1 and we don't know how many customers there are in queue 2. Rate out of this state is $\lambda + 2\mu$ because system can exit it either through arrival or through service. System can enter this state either: 1) by an arrival if only one server is busy, 2) by service if all servers are busy, there are 3 customers in service, and the 2^{nd} or the 3^{th} level customer could be served. Equating rate-in and rate-out we get equation (2.3.6).

- $(\lambda + 2\mu)\, p_{2;i} = \mu p_{3;i-1}, \quad i \geq 1$. Analyze the state $(2;i)$ and the rate in and the rate out of this state. State $(2;i)$ means that there are 2 busy servers, i customers in queue 1 and we don't know how many customers there are in queue 2.

Rate out of this state is $\lambda + 2\mu$ because system can exit it either through arrival or through service. System can enter this state: by service if all servers are busy, there are $i-1$ customers in queue 1, and the 1^{st} level customer is served. Equating rate-in and rate-out we get equation (2.3.7).

- $(\lambda + \mu)\, p_{1;0} = \lambda p_0 + \mu p_{2;0}$. Analyze the state $(1;0)$ and the rate in and the rate out of this state. State $(1;0)$ means that there is 1 busy server, 0 customers in queue 1 and we don't know how many customers there are in queue 2. Rate out of this state is $\lambda + \mu$ because system can exit it either through arrival or through service. System can enter this state either: 1) by an arrival if servers are idle, 2) by service if 2 servers are busy, and the 3^{th} level customer is served. Equating rate-in and rate-out we get equation (2.3.8).

- $(\lambda + \mu)\, p_{1;i} = \mu p_{2;i} + \mu \sum_{j=0}^{i-1} p_{2;i-j-1,j}, \quad i \geq 1$. Analyze the state $(1;i)$ and the rate in and the rate out of this state. State $(1;i)$ means that there is 1 busy server, i customers in queue 1 and we don't know how many customers there are in queue 2. Rate out of this state is $\lambda + \mu$ because system can exit it either through arrival or through service. System can enter this state either: 1) by service if 2 servers are busy, there are i customers in queue 1 and the 3^{th} level customer is served, 2) by service if 2 servers are busy, there are $i-j-1$ customers in queue 1 and j customers in queue 2, and the 2^{nd} level customer is served. Equating rate-in and rate-out we get equation (2.3.9).

- $(\lambda + 3\mu)\, p_{n;0,0} = \lambda p_{n-1;0,0} + \mu p_{n+1;0}, \quad n \geq 3$. Analyze the state $(n;0,0)$ and the rate in and the rate out of this state. State $(n;0,0)$ means that there are 3 busy servers, 0 customers in queue 1 and 0 customers in queue 2. Rate out of this state is $\lambda + 3\mu$ because system can exit it either through arrival or through service. System can enter this state either: 1) by an arrival if 3 servers are busy, there are $n-1$

customers between buffer and servers, 0 customers in queue 1, 0 customers in queue 2, 2) by service if 3 servers are busy, there are $n+1$ customers between buffer and service and 0 customers in queue 1, and the 3^{th} level customer is served. Equating rate-in and rate-out we get equation (2.3.10).

- $(\lambda + 3\mu)\, p_{n;0,j} = \lambda p_{n-1;0,j} + \mu p_{n+1;j} + \mu \sum_{k=0}^{j-1} p_{n+1;k,j-k-1}, \quad n \geq 3, j \geq 1$. Analyze the state $(n;0,j)$ and the rate in and the rate out of this state. State $(n;0,j)$ means that there are 3 busy servers, 0 customers in queue 1 and j customers in queue 2. Rate out of this state is $\lambda + 3\mu$ because system can exit it either through arrival or through service. System can enter this state: 1) by an arrival if 3 servers are busy, there are $n-1$ customers between buffer and servers, 0 customers in queue 1, j customers in queue 2, 2) by service if 3 servers are busy, there are $n+1$ customers between buffer and service and j customers in queue 1, and the 3^{th} level customer is served, 3) by service if 3 servers are busy, there are $n+1$ customers between buffer and service, k customers in queue 1 and $j-k-1$ customers in queue 2, and the 2^{nd} level customer is served. Equating rate-in and rate-out we get equation (2.3.11).

- $(\lambda + 3\mu)\, p_{n;i,j} = \lambda p_{n-1;i,j} + \mu p_{n+1;i-1,j}, \quad n \geq 3, i \geq 1, j \geq 0$. Analyze the state $(n;i,j)$ and the rate in and the rate out of this state. State $(n;i,j)$ means that there are 3 busy servers, i customers in queue 1 and j customers in queue 2. Rate out of this state is $\lambda + 3\mu$ because system can exit it either through arrival or through service. System can enter this state either: 1) by an arrival if 3 servers are busy, there are $n-1$ customers between buffer and servers, i customers in queue 1, j customers in queue 2, 2) by service if 3 servers are busy, there are $n+1$ customers between buffer and service, $i-1$ customers in queue 1, j customers in queue 2 and the 1^{st} level customer is served. Equating rate-in and rate-out we get equation (2.3.12).

- $(\lambda + 2\mu)\, p_{2;0,0} = \lambda p_{1;0} + \mu p_{3;0}$. Analyze the state $(2; 0, 0)$ and the rate in and the rate out of this state. State $(2; 0, 0)$ means that there are 2 busy servers, 0 customers in queue 1 and 0 customers in queue 2. Rate out of this state is $\lambda + 2\mu$ because system can exit it either through arrival or through service. System can enter this state either: 1) by an arrival if 1 server is busy, there are no customers between buffer and servers, 2) by service if 3 servers are busy, there are no customers in queue 1, and the 3^{th} level customer is served. Equating rate-in and rate-out we get equation (2.3.13).

- $(\lambda + 2\mu)\, p_{2;0,j} = \lambda p_{1;j} + \mu p_{3;j} + \mu \sum_{k=0}^{j-1} p_{3;k,j-k-1}, \quad j \geq 1$. Analyze the state $(2; 0, j)$ and the rate in and the rate out of this state. State $(2; 0, j)$ means that there are 2 busy servers, 0 customers in queue 1 and j customers in queue 2. Rate out of this state is $\lambda + 2\mu$ because system can exit it either through arrival or through service. System can enter this state either: 1) by an arrival if 1 server is busy, there are j customers in queue 1, 2) by service if 3 servers are busy, there are j customers in queue 1, and the 3^{th} level customer is served, 3) by service if 3 servers are busy, there are k customers in queue 1 and $j-k-1$ customers in queue 2, and the 2^{nd} level customer is served. Equating rate-in and rate-out we get equation (2.3.14).

- $(\lambda + 2\mu)\, p_{2;i,j} = \mu p_{3;i-1,j}, \quad i \geq 1, j \geq 0$. Analyze the state $(2; i, j)$ and the rate in and the rate out of this state. State $(2; i, j)$ means that there are 2 busy servers, i customers in queue 1 and j customers in queue 2. Rate out of this state is $\lambda + 2\mu$ because system can exit it either through arrival or through service. System can enter this state either: 1) by service if all servers are busy, there are $i - 1$ customers in queue 1, j customers in queue 2 and the 1^{st} level customer is served. Equating rate-in and rate-out we get equation (2.3.15).

The analysis of steady-state equations resulted in the development of simple algorithm for step-by-step computation of stationary joint probabilities $p_{n;i,j}, n \geq 2, i \geq 0, j \geq 0$ and $p_{n;i}, n \geq 1, i \geq 0$.

The algorithm is the following:

Inizialize λ, μ;
for $n \geq 0$ do:
calculate p_n from equation (2.2.1), (2.2.2), (2.2.3);
end for
calculate $p_{2;0}$ from equation (2.3.6);
for $n \geq 3$ do:
calculate $p_{n;0}$ from equation (2.3.4);
end for
for $i \geq 1$ do:
calculate $p_{2;i}$ from equation (2.3.7);
for $n \geq 3$ do:
calculate $p_{n;i}$ from equation (2.3.5);

end for
end for
calculate $p_{1;0}$ from equation (2.3.8);
calculate $p_{2;0,0}$ from equation (2.3.13);
for $n \geq 3$ do:
calculate $p_{n;0,0}$ from equation (2.3.10);
end for
for $i \geq 1$ do:
calculate $p_{2;i,0}$ from equation (2.3.15);
for $n \geq 3$ do:
calculate $p_{n;i,0}$ from equation (2.3.12);
end for
end for
for $i \geq 2$ do:
calculate $p_{1;i}$ from equation (2.3.9);
calculate $p_{2;0,i}$ from equation (2.3.14);
for $n \geq 3$ do:
calculate $p_{n;0,i}$ from equation (2.3.11);

end for
for $j \geq 1$ do:
calculate $p_{2;j,i}$ from equation (2.3.15);
for $m \geq 3$ do:
calculate $p_{m;i,j}$ from equation (2.3.12);

end for
end for
end for

For practical purposes it may be sometimes sufficient to know either only $\pi_{n;i}$ $n \geq 1, i \geq 0$ - stationary probabilities of the fact that total number of customers in servers and in buffer is n and total number of customers in RB (sum of queue #1 and queue #2) is i, or only $\pi_i, i \geq 0$ - stationary probabilities of the fact that there are n customers in total in the whole system (including buffer, servers, RB). These quantities can be calculated from joint probability distribution as follows:

$$\pi_{1;i} = p_{1;i},\ i \geq 0, \quad \pi_{2;i} = \sum_{j=0}^{i} p_{2;j,i-j},\ i \geq 0,$$

$$\pi_{n;i} = \sum_{j=0}^{i} p_{n;j,i-j},\ n \geq 3,\ i \geq 0,$$

$$\pi_0 = p_0,\ \pi_1 = \pi_{1;0},\ \pi_2 = \pi_{1;1} + \pi_{2;0},$$

$$\pi_i = \pi_{1;i-1} + \pi_{2;i-2} + \sum_{j=3}^{i} \pi_{j;i-j}, \quad i \geq 3.$$

2.4 Probability generating functions

Though the calculation of probabilities $p_{n;i,j}, n \geq 2, i \geq 0, j \geq 0$ and $p_{n;i}, n \geq 1, j \geq 0$ is just a matter of computational effort due to obtained above algorithm, performance characteristics (e.g. moments and/or correlation of queue lengths in RB) are not so straightforward to obtain. Below we show that in the considered case one can obtain expressions for probability generating functions (PGF) that ease the computation of various performance characteristics. Let us introduce the following PGF:

$$p_n(z) = \sum_{i=0}^{\infty} z^i p_{n;i}, \quad 0 \leq z \leq 1, n \geq 1,$$

$$p_n(z_1, z_2) = \sum_{i_1=0}^{\infty} \sum_{i_2=0}^{\infty} z_1^{i_1} z_2^{i_2} p_{n;i_1,i_2}, \quad 0 \leq z_1 \leq 1, 0 \leq z_2 \leq 1, n \geq 2,$$

$$P(u, z) = \sum_{n=3}^{\infty} u^{n-3} p_n(z), \quad 0 \leq u \leq 1,$$

$$P(u, z_1, z_2) = \sum_{n=3}^{\infty} u^{n-3} p_n(z_1, z_2), \quad 0 \leq u \leq 1.$$

If one puts $z_1 = z_2 = z$ in $P(u, z_1, z_2)$, then function $P(u, z, z)$ is the double PGF of the total number of customers in buffer and servers and total number of customers in RB when all three servers are busy. By analogy $p_n(z, z), n \geq 2$, is the PGF of the total number of customers: total number of customers in RB and probability of total n customers in servers and buffer. In the following we will make use of PGF $P(u) = \sum_{n=3}^{\infty} u^{n-3} p_n, |u| \leq 1$ which, with respect to (2.2.1)-(2.2.3), equals $P(u) = \frac{p_3}{1-\tilde{\rho}u}$.

Now we will successively obtain relations for PGF defined above. Start with the following equations:

$$(\lambda + 3\mu) p_{n;0} = \lambda p_{n-1;0} + 2\mu p_{n+1}, \qquad n \geq 3, \tag{2.4.16}$$

$$p_{n;i}(\lambda+3\mu) = p_{n-1;i}\lambda + p_{n+1;i-1}\mu, \qquad n \geq 3, i \geq 1. \tag{2.4.17}$$

We multiply for z^i and sum on i only the equation (4.3.1) because of the equation (2.4.16) does not depend on i :

$$\sum_{i=1}^{\infty} z^i p_{n;i}(\lambda+3\mu) = \sum_{i=1}^{\infty} z^i p_{n-1;i}\lambda + \sum_{i=1}^{\infty} z^i p_{n+1;i-1}\mu \qquad n \geq 3, i \geq 1$$

from which we obtain:

$$(\lambda+3\mu)\left[p_n(z) - p_{n;0}\right] = \lambda\left[p_{n-1}(z) - p_{n-1;0}\right] + \mu z p_{n+1}(z) \tag{2.4.18}$$

Here we sum (2.4.16) with (4.3.5) in order to find $p_n(z)$:

$$p_n(z) = \frac{1}{\lambda+3\mu}\left[2\mu p_{n+1} + \lambda p_{n-1}(z) + \mu z p_{n+1}(z)\right] \qquad n \geq 3 \tag{2.4.19}$$

Then we analize the following equations:

$$p_{2;0}(\lambda+2\mu) = p_1\lambda + p_3 2\mu \tag{2.4.20}$$

$$p_{2;i}(\lambda+2\mu) = p_{3;i-1}\mu \qquad i \geq 1 \tag{2.4.21}$$

Multiplying for z^i and summing on i only the equation (2.4.21), because of the equation (2.4.20) does not depend on i, we obtain:

$$\sum_{i=1}^{\infty} z^i p_{2;i}(\lambda+2\mu) = \sum_{i=1}^{\infty} z^i p_{3;i-1}\mu \qquad i \geq 1$$

$$(\lambda+2\mu)\left[\sum_{i=0}^{\infty} z^i p_{2;i} - p_{2;0}\right] = \mu z \sum_{t=0}^{\infty} z^t p_{3;t}$$

$$(\lambda+2\mu)\left[p_2(z) - p_{2;0}\right] = \mu z p_3(z) \tag{2.4.22}$$

Now we sum (2.4.20) with (2.4.22) in order to find $p_2(z)$:

$$p_2(z) = \frac{1}{\lambda+2\mu}\left[\lambda p_1 + 2\mu p_3 + \mu z p_3(z)\right] \tag{2.4.23}$$

We analize the following equations:

$$p_{1;0}(\lambda+\mu) = p_0\lambda + p_{2;0}\lambda \tag{2.4.24}$$

$$p_{1;i}(\lambda+\mu) = p_{2;i}\mu + \sum_{j=0}^{i-1} p_{2;i-j-1,j}\mu \qquad i \geq 1 \tag{2.4.25}$$

We multiply for z^i and sum on i only the equation (2.4.25) because of the equation (2.4.24) does not depend on i :

$$\sum_{i=1}^{\infty} z^i p_{1;i}(\lambda+\mu) = \sum_{i=1}^{\infty} z^i p_{2;i}\mu + \sum_{i=1}^{\infty} z^i \sum_{j=0}^{i-1} p_{2;i-j-1,j}\mu \qquad i \geq 1$$

$$(\lambda+\mu)\left[\sum_{i=0}^{\infty} z^i p_{1;i} - p_{1;0}\right] = \mu\left[\sum_{i=0}^{\infty} z^i p_{2;i} - p_{2;0}\right] + \sum_{i=1}^{\infty} z^i \sum_{j=0}^{i-1} p_{2;i-j-1,j}\mu$$

$$(\lambda+\mu)\left[p_1(z) - p_{1;0}\right] = \mu\left[p_2(z) - p_{2;0}\right] + \sum_{i=1}^{\infty}\sum_{j=0}^{i-1} z^i p_{2;i-j-1,j}\mu \tag{2.4.26}$$

In order to find $p_1(z)$ we sum (2.4.24) with (2.4.26):

$$p_1(z) = \frac{1}{\lambda+\mu}\left[p_0\lambda + \mu p_2(z) + \mu z \sum_{t=0}^{\infty}\sum_{j=0}^{t} z^t p_{2;t-j,j}\right] \tag{2.4.27}$$

and we observe that for $t - j = k$:

$$\sum_{k+j=0}^{\infty}\sum_{j=0}^{k+j} z^{k+j} p_{2;k,j} = \sum_{k+j=0}^{\infty}\sum_{j=0}^{\infty} z^{k+j} p_{2;k,j} = p_2(z,z)$$

so

$$p_1(z) = \frac{1}{\lambda+\mu}\left[p_0\lambda + \mu p_2(z) + \mu z p_2(z,z)\right] \tag{2.4.28}$$

The next step is to analize the following equations:

$$p_{n;0,0}(\lambda+3\mu) = p_{n-1;0,0}\lambda + p_{n+1;0} \qquad n \geq 3 \tag{2.4.29}$$

$$p_{n;0,j}(\lambda+3\mu) = p_{n-1;0,j}\lambda + p_{n+1;j}\mu + \sum_{k=0}^{j-1} p_{n+1;k,j-k-1}\mu \qquad n \geq 3, j \geq 1 \tag{2.4.30}$$

$$p_{n;i,j}(\lambda+3\mu) = p_{n-1;i,j}\lambda + p_{n+1;i-1,j}\mu \qquad n \geq 3, i \geq 1, j \geq 0 \tag{2.4.31}$$

We multiply for z_2^j and sum on j the equation (2.4.30), then we multiply for z_1^i and z_2^j and sum on i and on j the equation (2.4.31), while the equation (2.4.29) does depend neither on i or on j :

$$\sum_{j=1}^{\infty} z_2^j p_{n;0,j}(\lambda+3\mu) = \sum_{j=1}^{\infty} z_2^j p_{n-1;0,j}\lambda + \sum_{j=1}^{\infty} z_2^j p_{n+1;j}\mu +$$

$$+ \sum_{j=1}^{\infty} z_2^j \sum_{k=0}^{j-1} p_{n+1;k,j-k-1}\mu \quad n \geq 3, j \geq 1$$

and after manipulations:

$$(\lambda+3\mu)\left[\sum_{j=0}^{\infty} z_2^j p_{n;0,j} - p_{n;0,0}\right] = \lambda\left[\sum_{j=0}^{\infty} z_2^j p_{n-1;0,j} - p_{n-1;0,0}\right] + \tag{2.4.32}$$

$$+\mu\left[p_{n+1}(z_2) - p_{n+1;0}\right] + \mu z_2\left[\sum_{t=0}^{\infty}\sum_{k=0}^{t} z_2^t p_{n+1;k,t-k}\right]$$

Here we study the equation (2.4.31):

$$\sum_{j=0}^{\infty} z_2^j \sum_{i=1}^{\infty} z_1^i p_{n;i,j}(\lambda+3\mu) = \sum_{j=0}^{\infty} z_2^j \sum_{i=1}^{\infty} z_1^i p_{n-1;i,j}\lambda +$$

$$+ \sum_{j=0}^{\infty} z_2^j \sum_{i=1}^{\infty} z_1^i p_{n+1;i-1,j}\mu \qquad n \geq 3, i \geq 1, j \geq 0$$

introducing the term for $i = 0$ and after appropriate manipulations, we find:

$$(\lambda+3\mu)\left[p_n(z_1,z_2) - \sum_{j=0}^{\infty} z_2^j p_{n;0,j}\right] = \lambda\left[p_{n-1}(z_1,z_2) - \sum_{j=0}^{\infty} z_2^j p_{n-1;0,j}\right] + \tag{2.4.33}$$

$$+\mu z_1 p_{n+1}(z_1,z_2)$$

In order to find $p_n(z_1, z_2)$ we sum (2.4.29), (2.4.32) and (2.4.33):

$$p_{n;0,0}(\lambda + 3\mu) + (\lambda + 3\mu)\left[\sum_{j=0}^{\infty} z_2^j p_{n;0,j} - p_{n;0,0}\right] + (\lambda + 3\mu)$$

$$\left[p_n(z_1, z_2) - \sum_{j=0}^{\infty} z_2^j p_{n;0,j}\right] = p_{n-1;0,0}\lambda + p_{n+1;0} +$$

$$+\lambda\left[\sum_{j=0}^{\infty} z_2^j p_{n-1;0,j} - p_{n-1;0,0}\right] + \mu\left[p_{n+1}(z_2) - p_{n+1;0}\right] +$$

$$+\mu z_2\left[\sum_{t=0}^{\infty}\sum_{k=0}^{t} z_2^t p_{n+1;k,t-k}\right] + \lambda\left[p_{n-1}(z_1, z_2) - \sum_{j=0}^{\infty} z_2^j p_{n-1;0,j}\right] +$$

$$+\mu z_1 p_{n+1}(z_1, z_2)$$

observing that $t = k + s$ we obtain:

$$\sum_{k+s=0}^{\infty}\sum_{k=0}^{\infty} z_2^{k+s} p_{n+1;k,s} = p_2(z_2, z_2)$$

so

$$p_n(z_1, z_2) = \frac{1}{(\lambda + 3\mu)}\left[\mu p_{n+1}(z_2) + z_2\mu p_2(z_2, z_2) + \right.$$
$$\left. +\lambda p_{n-1}(z_1, z_2) + \mu z_1 p_{n+1}(z_1, z_2)\right]$$

Here we analize the following equations:

$$p_{2;0,0}(\lambda + 2\mu) = p_{1;0}\lambda + p_{3;0}\mu \tag{2.4.34}$$

$$p_{2;0,j}(\lambda + 2\mu) = p_{1;j} + p_{3;j}\mu + \sum_{k=0}^{j-1} p_{3;k,j-k-1}\mu \qquad j \geq 1 \tag{2.4.35}$$

$$p_{2;i,j}(\lambda+2\mu) = p_{3;i-1,j}\mu \qquad i \geq 1, j \geq 0 \tag{2.4.36}$$

Now we multiply for z_2^j and sum on j the equation (2.4.35), then we multiply for z_1^i and z_2^j and sum on i and on j the equation (2.4.36), while the equation (2.4.34) does depend neither on i or on j :

$$\sum_{j=1}^{\infty} z_2^j p_{2;0,j}(\lambda+2\mu) = \lambda \sum_{j=1}^{\infty} z_2^j p_{1;j} + \sum_{j=1}^{\infty} z_2^j p_{3;j}\mu + \sum_{j=1}^{\infty} z_2^j \sum_{k=0}^{j-1} p_{3;k,j-k-1}\mu$$

$$j \geq 1$$

$$(\lambda+2\mu)\left[\sum_{j=0}^{\infty} z_2^j p_{2;0,j} - p_{2;0,0}\right] = \lambda p_1(z_2) - \lambda p_{1;0} + \mu\left[p_3(z_2) - p_{3;0}\right] + \tag{2.4.37}$$

$$+\mu z \sum_{t=0}^{\infty} z_2^t \sum_{k=0}^{t} p_{3;k,t-k}$$

$$\sum_{j=0}^{\infty} z_2^j \sum_{i=1}^{\infty} z_1^i p_{2;i,j}(\lambda+2\mu) = \sum_{j=0}^{\infty} z_2^j \sum_{i=1}^{\infty} z_1^i p_{3;i-1,j}\mu \qquad i \geq 1, j \geq 0$$

as in the previous case, we introduce the term for $i = 0$ and after appropriate substitutions we find:

$$(\lambda+2\mu)\left[p_2(z_1,z_2) - \sum_{j=0}^{\infty} z_2^j p_{2;0,j}\right] = \mu z_1 p_3(z_1,z_2) \tag{2.4.38}$$

In the next step we sum (2.4.34), (2.4.37) and (2.4.38) in order to find $p_2(z_1, z_2)$:

$$p_{2;0,0}(\lambda+2\mu) + (\lambda+2\mu)\left[\sum_{j=0}^{\infty} z_2^j p_{2;0,j} - p_{2;0,0}\right] + (\lambda+2\mu)$$

$$\left[p_2(z_1,z_2) - \sum_{j=0}^{\infty} z_2^j p_{2;0,j}\right] = p_{1;0}\lambda + p_{3;0}\mu + \lambda p_1(z_2) - \lambda p_{1;0} +$$

$$+\mu\left[p_3(z_2)-p_{3;0}\right]+\mu z\sum_{t=0}^{\infty}z_2^t\sum_{k=0}^{t}p_{3;k,t-k}+\mu z_1p_3(z_1,z_2)$$

we observe that for $t=a+k$ we get:

$$\sum_{a+k=0}^{\infty}z_2^{a+k}\sum_{k=0}^{\infty}p_{3;k,a}=p_3(z_2,z_2)$$

so

$$p_2(z_1,z_2)=\frac{1}{\lambda+2\mu}\left[\lambda p_1(z_2)+\mu p_3(z_2)+\mu z_2p_3(z_2,z_2)+\right. \quad (2.4.39)$$

$$\left.+\mu z_1p_3(z_1,z_2)\right]$$

We find $P(u,z)$:

$$\sum_{n=3}^{\infty}u^{n-3}(\lambda+3\mu)p_n(z)=\sum_{n=3}^{\infty}u^{n-3}\left[2\mu p_{n+1}+\lambda p_{n-1}(z)+\mu z p_{n+1}(z)\right]$$

$$(\lambda+3\mu)\sum_{n=3}^{\infty}u^{n-3}p_n(z)=\sum_{n=3}^{\infty}\left[u^{n-3}2\mu p_{n+1}+u^{n-3}\lambda p_{n-1}(z)+\right.$$

$$\left.+u^{n-3}\mu z p_{n+1}(z)\right]\ (\lambda+3\mu)P(u,z)=\sum_{n=3}^{\infty}u^{n-3}2\mu p_{n+1}+$$

$$+\sum_{n=3}^{\infty}u^{n-3}\lambda p_{n-1}(z)+\sum_{n=3}^{\infty}u^{n-3}\mu z p_{n+1}(z)$$

Thanks to the notations introduced at the beginning of this chapter, we can write a better expression using $P(u),P(u,z)$:

$$(\lambda+3\mu)P(u,z)=\frac{2\mu}{u}\left[\sum_{n=3}^{\infty}u^{n-3}p_n-p_3\right]+\lambda\sum_{n=3}^{\infty}u^{n-3}p_{n-1}(z)+$$

$$+\mu z\sum_{n=3}^{\infty}u^{n-3}p_{n+1}(z)$$

$$(\lambda+3\mu)P(u,z)=\frac{2\mu}{u}\left[P(u)-p_3\right]+\lambda\sum_{n=3}^{\infty}u^{n-3}p_{n-1}(z)+$$

$$+\mu z\sum_{n=3}^{\infty}u^{n-3}p_{n+1}(z)$$

$$(\lambda+3\mu)P(u,z)=\frac{2\mu}{u}\left[P(u)-p_3\right]+\lambda\left[u\sum_{n=3}^{\infty}u^{n-3}p_n(z)+p_2(z)\right]+$$

$$+\mu z\sum_{n=3}^{\infty}u^{n-3}p_{n+1}(z)$$

$$(\lambda+3\mu)P(u,z)=\frac{2\mu}{u}\left[P(u)-p_3\right]+\lambda\left[uP(u,z)+p_2(z)\right]+$$

$$+\mu\frac{z}{u}\left[\sum_{n=3}^{\infty}u^{n-3}p_n(z)-p_3(z)\right]$$

$$(\lambda+3\mu)P(u,z)=\frac{2\mu}{u}\left[P(u)-p_3\right]+\lambda\left[uP(u,z)+p_2(z)\right]+$$

$$+\mu\frac{z}{u}\left[P(u,z)-p_3(z)\right]$$

$$u(\lambda+3\mu)P(u,z)=2\mu\left[P(u)-p_3\right]+\lambda u^2P(u,z)+$$

$$\lambda up_2(z)+\mu zP(u,z)-\mu zp_3(z)$$

Finally we obtain:

$$P(u,z)=\frac{\mu zp_3(z)-\lambda\mu p_2(z)-2\mu\left[P(u)-p_3\right]}{\lambda u^2+\mu z-u(\lambda+3\mu)} \tag{2.4.40}$$

The next step is to find $P(u,z_1,z_2)$:

$$(\lambda+3\mu)\sum_{n=3}^{\infty}u^{n-3}p_n(z_1,z_2)=\sum_{n=3}^{\infty}u^{n-3}\left[\mu p_{n+1}(z_2)+\right.$$

$$\left.+z_2\mu p_{n+1}(z_2,z_2)+\lambda p_{n-1}(z_1,z_2)+\mu z_1p_{n+1}(z_1,z_2)\right]$$

$$(\lambda+3\mu)P(u,z_1,z_2)=\lambda\sum_{n=3}^{\infty}u^{n-3}p_{n-1}(z_1,z_2)+\sum_{n=3}^{\infty}u^{n-3}\mu p_{n+1}(z_2)+$$

$$+\sum_{n=3}^{\infty} u^{n-3}\mu z_2 p_{n+1}(z_2,z_2)+\sum_{n=3}^{\infty} u^{n-3}\mu z_1 p_{n+1}(z_1,z_2)$$

$$(\lambda+3\mu)P(u,z_1,z_2)=\lambda\left[u\sum_{n=3}^{\infty} u^{n-3}p_n(z_1,z_2)+p_2(z_1,z_2)\right]+$$

$$+\frac{\mu}{u}\left[\sum_{n=3}^{\infty} u^{n-3}p_n(z_2)-p_3(z_2)\right]+\frac{\mu}{u}z_2\left[\sum_{n=3}^{\infty} u^{n-3}p_n(z_2,z_2)-p_3(z_2,z_2)\right]+$$

$$+\frac{\mu}{u}z_1\left[\sum_{n=3}^{\infty} u^{n-3}p_n(z_1,z_2)-p_3(z_1,z_2)\right]$$

$$(\lambda+3\mu)P(u,z_1,z_2)=\lambda\left[uP(u,z_1,z_2)+p_2(z_1,z_2)\right]+\frac{\mu}{u}\left[P(u,z_2)+\right.$$

$$\left.-p_3(z_2)\right]+\frac{\mu}{u}z_2\left[P(u,z_2,z_2)-p_3(z_2,z_2)\right]+\frac{\mu}{u}z_1\left[P(u,z_1,z_2)-p_3(z_1,z_2)\right]$$

$$P(u,z_1,z_2)=\frac{1}{\lambda u^2+\mu z_1-(\lambda+3\mu)u}\left[-\lambda u p_2(z_1,z_2)+\right. \quad (2.4.41)$$

$$\left.-\mu\left[P(u,z_2)-p_3(z_2)\right]-\mu z_2\left[P(u,z_2,z_2)-p_3(z_2,z_2)\right]+\mu z_1 p_3(z_1,z_2)\right]$$

Assuming $z_1=z_2=z$ we find $P(u,z,z)$ and $p_2(z,z)$:

$$P(u,z,z)=\frac{1}{\lambda u^2+2\mu z-(\lambda+3\mu)u}\left[-\lambda u p_2(z,z)+\right. \quad (2.4.42)$$

$$\left.-\mu\left[P(u,z)-p_3(z)\right]+2\mu z p_3(z,z)\right]$$

$$p_2(z,z)=\frac{1}{(\lambda+2\mu)}\left[\lambda p_1(z)+\mu p_3(z)+2\mu z p_3(z,z)\right] \quad (2.4.43)$$

In order to find solution of the denominator $P(u,z,z)$ we consider:

$$f_m(u,z)=\lambda u^2+m\mu z-(\lambda+3\mu)u \qquad m=1,2$$

and study:

$$f_m(u,z)=0\Rightarrow\lambda u^2+m\mu z-(\lambda+3\mu)u=0$$

from which

$$u_m = \frac{\lambda + 3\mu - \sqrt{(\lambda + 3\mu)^2 - 4m\lambda\mu z}}{2\lambda}$$

and

$$\hat{u}_m = \frac{\lambda + 3\mu + \sqrt{(\lambda + 3\mu)^2 - 4m\lambda\mu z}}{2\lambda} = \frac{\lambda + 3\mu}{\lambda} - u_m$$

If $z = 0$:

$$u_m = \frac{\lambda + 3\mu - \sqrt{(\lambda + 3\mu)^2}}{2\lambda} = 0$$

If $z = 1$:

$$\hat{u}_m = \frac{\lambda + 3\mu + \sqrt{(\lambda + 3\mu)^2 - 4m\lambda\mu}}{2\lambda} = 1$$

Now we rewrite $P(u, z)$ and $P(u, z, z)$:

$$P(u, z) = \frac{\mu z p_3(z) - \lambda\mu p_2(z) - 2\mu\left[P(u) - p_3\right]}{f_1(u, z)} \tag{2.4.44}$$

$$P(u, z, z) = \frac{1}{f_2(u, z)}\left[-\lambda u p_2(z, z) - \right. \tag{2.4.45}$$
$$\left. \mu\left[P(u, z) - p_3(z)\right] + 2\mu z p_3(z, z)\right]$$

Denominator in (2.4.44) and (2.4.45) is zero at points $(u_1, z) = (u_1(z), z)$ and $(u_2, z) = (u_2(z), z)$. Since PGF $P(u, z, z)$ is analytic function in the domain $0 \le z \le 1$ then numerator must be zero at these points too. This leads to the following equations:

$$\mu z p_3(z) - \lambda u_1 p_2(z) - 2\mu\left[P(u_1) - p_3\right] = 0 \tag{2.4.46}$$

$$2\mu z p_3(z, z) - \lambda u_2 p_2(z, z) - \mu\left[P(u_2, z) - p_3(z)\right] = 0 \tag{2.4.47}$$

Firstly we find PGF $P(u, z)$. Solution of equations (2.4.23) and (2.4.46):

$$\begin{cases} (\lambda + 2\mu)p_2(z) - \mu z p_3(z) = \lambda p_1 + 2\mu p_3 \\ \mu z p_3(z) - \lambda u_1 p_2(z) = 2\mu\left[P(u_1) - p_3\right] \end{cases}$$

$$\begin{cases} (\lambda+2\mu)p_2(z) - [2\mu\,[P(u_1)-p_3] + \lambda u_1 p_2(z)] = \lambda p_1 + 2\mu p_3 \\ p_3(z) = \frac{1}{\mu z}\,[2\mu\,[P(u_1)-p_3] + \lambda u_1 p_2(z)] \end{cases}$$

$$\begin{cases} p_2(z) = \frac{1}{2\mu+\lambda-\lambda u_1}\,[\lambda p_1 - 2\mu P(u_1)] \\ p_3(z) = \frac{1}{\mu z}\,[2\mu\,[P(u_1)-p_3] + \lambda u_1 p_2(z)] \end{cases}$$

Now we substitute $p_2(z)$ and $p_3(z)$ in $P(u,z)$:

$$P(u,z) = \frac{1}{f_1(u,z)}\left[\mu z\frac{1}{\mu z}\,[\lambda u_1 p_2(z) + 2\mu\,[P(u_1)-p_3]] + \right.$$

$$\left. -(\lambda\mu)\,\frac{\lambda p_1 + 2\mu P(u_1)}{\lambda - \lambda u_1 + 2\mu} - 2\mu\,[P(u)-p_3]\right]$$

$$= \frac{1}{f_1(u,z)}\left[\lambda u_1 p_2(z) + 2\mu\,[P(u_1)-p_3] - (\lambda\mu)\,\frac{\lambda p_1 + 2\mu P(u_1)}{\lambda - \lambda u_1 + 2\mu} + \right.$$

$$\left. -2\mu\,[P(u)-p_3]\right]$$

$$= \frac{1}{(u-u_1)\,(u-\hat{u}_1)}\left[\lambda u_1\frac{\lambda p_1 + 2\mu P(u_1)}{\lambda - \lambda u_1 + 2\mu} + 2\mu\,[P(u_1)-P(u)] + \right.$$

$$\left. -\lambda u\frac{\lambda p_1 + 2\mu P(u_1)}{\lambda - \lambda u_1 + 2\mu}\right]$$

$$= \frac{1}{(u-u_1)\,(u-\hat{u}_1)}\left[p_2(z)(\lambda u_1 - \lambda u) + 2\mu\left(\frac{p_3}{1-\tilde{\rho}u_1} - \frac{p_3}{1-\tilde{\rho}u}\right)\right]$$

$$= \frac{1}{(u-u_1)\,(u-\hat{u}_1)}\,[p_2(z)(\lambda u_1 - \lambda u) +$$

$$\left. +2\mu\left(\frac{p_4}{\tilde{\rho}(1-\tilde{\rho}u_1)} - \frac{p_4}{\tilde{\rho}(1-\tilde{\rho}u)}\right)\right]$$

$$= \frac{1}{(u-u_1)\,(u-\hat{u}_1)}\left[\lambda p_2(z)(u_1-u) + 2\mu\frac{p_4\tilde{\rho}(u_1-u)}{\tilde{\rho}(1-\tilde{\rho}u)(1-\tilde{\rho}u_1)}\right]$$

substitute $p_2(z)$:

$$= \frac{1}{(u-\hat{u}_1)}\left[-\lambda p_2(z) - \frac{2\mu p_4}{(1-\tilde{\rho}u)(1-\tilde{\rho}u_1)}\right]$$

$$= \frac{1}{(u-\hat{u}_1)}\left[-\lambda\frac{\lambda p_1 + 2\mu P(u_1)}{\lambda - \lambda u_1 + 2\mu} - \frac{2\mu p_4}{(1-\tilde{\rho}u)(1-\tilde{\rho}u_1)}\right]$$

substitute $P(u_1)$:

$$= \frac{1}{(u-\hat{u}_1)}\left[-\lambda\frac{\lambda p_1 + 2\mu\frac{p_3}{1-\tilde{\rho}u_1}}{\lambda - \lambda u_1 + 2\mu} - \frac{2\mu p_4}{(1-\tilde{\rho}u)(1-\tilde{\rho}u_1)}\right]$$

$$= \frac{\lambda}{(\hat{u}_1 - u)\,(1-\tilde{\rho}u_1)}\left[\frac{[\lambda+2\mu]p_{2;0} - \lambda p_1\tilde{\rho}u_1}{(\lambda - \lambda u_1 + 2\mu)} + \frac{2\mu p_4}{(1-\tilde{\rho}u)}\right] \tag{2.4.48}$$

Now we find the expression for $P(u,z,z)$. Solving system of equations (2.4.28), (2.4.43) and (2.4.47), one obtains the following expression for PGFs $p_1(z), p_2(z,z)$ and $p_3(z,z)$:

$$p_1(z) = \frac{1}{\lambda+\mu}\left[\lambda p_0 + \mu p_2(z) + \mu z p_2(z,z)\right]$$

$$p_3(z,z) = \frac{1}{2\mu z}\left[\lambda u_2 p_2(z,z) + \mu\left[P(u_2,z) - p_3(z)\right]\right]$$

$$p_2(z,z) = \frac{1}{\lambda - \lambda u_2 + 2\mu - \frac{\lambda\mu z}{\lambda+\mu}}\left[\frac{1}{\lambda+\mu}\left[\lambda p_0 + \mu p_2(z)\right] + \mu P(u_2,z)\right]$$

If one substitutes expression for $p_1(z), p_2(z,z)$ and $p_3(z,z)$ into (2.4.45) then, after collecting the common terms, one finds $P(u,z,z)$. The last PGF to find is $P(u,z_1,z_2)$. Denominator in (2.4.41) is zero at point $(u_1(z_1), z_1)$. Since PGF $P(u,z_1,z_2)$ is an analytic function in the domain $0 \le z_1 \le 1, 0 \le z_2 \le 1$ then numerator must vanish at this point. Hence it holds:

$$-\lambda u_1(z_1)p_2(z_1,z_2) - \mu\left[P(u_1(z_1),z_2) - p_3(z_2)\right] + \tag{2.4.49}$$

$$-\mu z_2\left[P(u_1(z_1),z_2,z_2) - p_3(z_2,z_2)\right] + \mu z_1 p_3(z_1,z_2) = 0.$$

From relation (2.4.39) it follows that:

$$\mu z_1 p_3(z_1,z_2) = (\lambda+2\mu)p_2(z_1,z_2) - \lambda p_1(z_2) - \mu p_3(z_2) - \mu z_2 p_3(z_2,z_2).$$

Substitution of $\mu z_1 p_3(z_1,z_2)$ into (2.4.49), leads to the expression for $p_2(z_1,z_2)$:

$$p_2(z_1,z_2) = \frac{1}{[\lambda + 2\mu - \lambda u_1(z_1)]}\left[\lambda p_1(z_2) + \mu P(u_1(z_1),z_2) + \right.$$

$$+\mu z_2 P(u_1(z_1), z_2, z_2)]$$

Thus we have obtained all the unknown quantities in PGF $P(u_1, z_1, z_2)$ and it is determined completely.

2.5 Numerical results

There are several quantities related to the number of customers in the system that may be of interest. They are mean and variance of the number of customers in queue 1 and queue 2, correlation between queue size in buffer and queue 1, between queue size in buffer and queue 2 and between queue 1 and queue 2. These quantities are calculated considering $\lambda = 2.5$, $\mu = 1$ and $n = 100$. In the Figure 2.2 the upper, the middle and the lower line represents respectively the mean number of customers in RB, queue 2 of RB and queue 1 of RB:

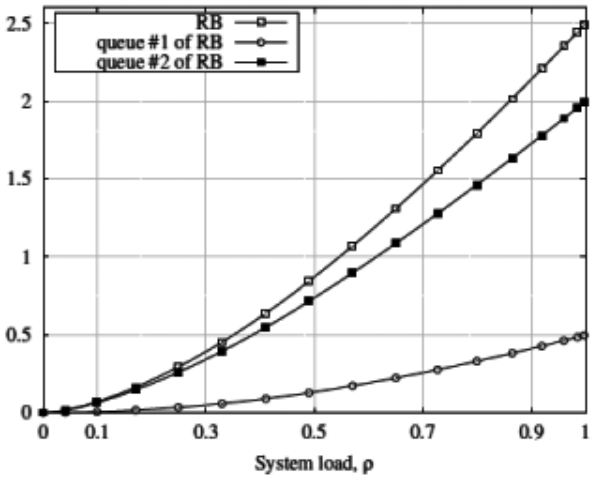


Figure 2.2 Mean number of customers.

We observe that the mean number of customers in RB is the sum of the mean number of customers in queue 1 of RB and queue 2 in RB. In the Figure 2.3 the upper, the middle and the lower line represents respectively the variance of the number of customers in RB, queue 2 of RB and queue 1 of RB.

We observe that the variance of the number of customers in RB is almost the sum of the variance of the number of customers in queue 1 of RB and queue 2 of RB, and this is strange because it is well known that $Var(X+Y) = Var(X)+Var(Y)+Cov(X,Y)$,

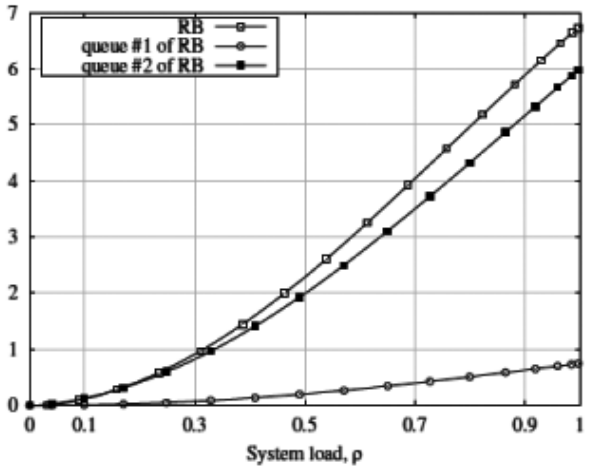


Figure 2.3 Variance of the number of customers.

where X is the number of customers in queue 1 of RB and Y is the number of customers in queue 2 of RB, so the $Cov(X, Y) \sim 0$. In fact if we observe the Figure 2.4 where the upper, the middle and the lower line represents respectively the correlation between buffer and queue 1 of RB, between buffer and queue 2 of RB, between queue 1 and queue 2 of RB, we notice that the correlation between queue 1 and queue 2 of RB is almost equal to zero, so the number of customers in queue 1 of RB and the number of customers in queue 2 of RB are almost uncorrelated, so the $Var(X + Y)$ is almost equal to $Var(X) + Var(Y)$.

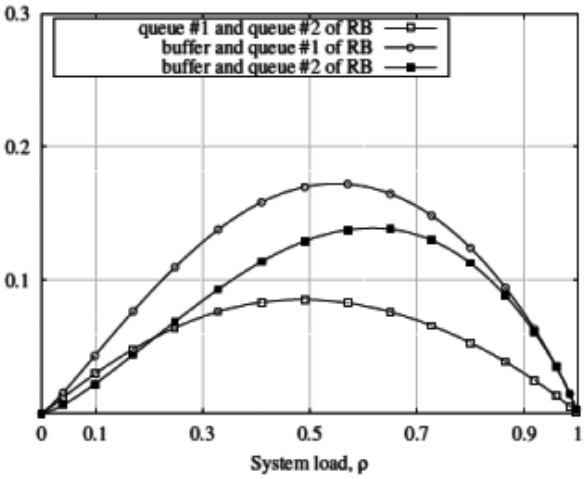


Figure 2.4 Coefficients of correlation.

Chapter 3

N-server queueing system with poisson input and exponential service times

3.1 Problem statement

We consider a queueing system with $3 < N < \infty$ servers, infinite capacity buffer, incoming Poisson flow of customers (of intensity λ), exponential distribution of service time in each server (with parameter μ) and resequencing buffer (RB) of infinite capacity. Customers upon entering the system obtain sequential number and join buffer. Without loss of generality we suppose that the sequence starts from 1 and coincides with the row of natural numbers, i.e. the first customer upon entering the (empty) system receives number 1, the second one number 2 and so on and so forth. Customers leave the system strictly in order of their arrival. Thus after customer's arrival it remains in the buffer for some time and then receives service when one of the servers becomes idle. If at the moment of its service completion there are no customers in the system or all other customers present at that moment in the queue and in all other servers have greater sequential numbers, it leaves the system. Otherwise it occupies one place in the RB. Customer from RB leaves it if and only if its sequential number

is less than sequential number of all other customers present in system i.e. customers may leave RB in groups.

3.2 Model description and notation

As it was mentioned in the previous section customers awaiting in RB may form separate (virtual) queues. In order to explain this let us introduce the following notation. If there are n, $n = 1, .., N$ busy servers, we call "1^{st} level" customer the last one (among those n in servers) which joined the system; we call "2^{nd} level" customer the penultimate customer which joined the system (among those n in servers); "3^{rd} level" customer is the one which joined the system before the penultimate customer etc. The customer which was the first (among those n in servers) to join the system is "n^{th}level" customer. If at some instant all servers are busy i.e. $n = N$, then customers which joined RB between "1^{st} level" and "2^{nd} level" customer form queue 1; customers which joined RB between "2^{nd} level" and "3^{rd} level" customer form queue 2 etc. Clearly, customers which joined RB between "$(n-1)^{th}$ level" and "n^{th} level" customer form queue (N-1). Notice that if at some instant $n < N$, then customers which joined RB after "1^{st} level" customer form queue 1; customers which joined RB between "1^{st} level" and "2^{nd} level" customer form queue 2 etc. The operation of the considered queueing system can be completely described by Markov process $\zeta(t) = \{(\xi(t), \eta_1(t), \eta_2(t), ..., \eta_{N-1}(t)), t \geq 0\}$ where $\xi(t)$ is the number of customers in buffer and all servers at time t, $\eta_i(t)$ is the number of customers in queue $\#i$ of RB at time t. In case $\xi(t) = 0$, the all but first component of $\zeta(t)$ are omitted; in case $\xi(t) = n$, $n = 1, ..., N-2$, the last $N-1-n$ components are omitted. The state space of $\zeta(t)$ is:

$$\chi = \{0\} \cup \{(1, i_1), i_1 \geq 0\} \cup \{(2, i_1, i_2), i_1 \geq 0, i_2 \geq 0\} \cup ...$$

$$\cup\{(2, i_1, i_2, ..., i_{N-1}), n \geq N-1, i_1, i_2, ..., i_{N-1} \geq 0\}.$$

It is assumed that service and arrival processes are mutually independent and necessary and sufficient condition of stationarity $\tilde{\rho} = \frac{\rho}{N} < 1$, where $\rho = \frac{\lambda}{\mu}$ holds for the system. Indeed one can notice that the total number of customers in buffer and servers of the considered QS with resequencing coincides with the total number of customers in M/M/∞/∞ queue. Therefore, its stationary

distribution $\{p_n, n \geq 0\}$, has the form:

$$p_n = \frac{\rho^n}{n!} p_0, \ n = 1, ..., N, \tag{3.2.1}$$

$$p_n = \frac{\rho^n}{N! N^{n-N}} p_0, \ n \geq N+1, \tag{3.2.2}$$

$$p_0 = \left(\sum_{n=0}^{N-1} \frac{\rho^n}{n!} + \frac{\rho^N}{(N-1)!(N-\rho)} \right)^{-1}. \tag{3.2.3}$$

Provided that RB is empty when servers are idle, p_0 is also the probability of the considered system with resequencing to be empty. If $n \geq N$ let us denote by $p_{n;i}^{(m)}$, $m = 1, ..., N-1$, $i \geq 0$, stationary probability of the fact that there are total n customers in servers and buffer and total number of customers in first m queues in RB equals i i.e.

$$p_{n;i}^{(m)} = \lim_{t \to \infty} P\left\{\xi(t) = n, \eta_1(t) + ... + \eta_m(t) = i\right\},$$

$$n = 1, ..., N-1, m = 1, ..., n, i \geq 0.$$

By $p_{n;i}^{(m)}$, $m = 1, ..., n$, $i \geq 0$ we denote similar stationary probability when $n = 1, ..., N-1$ i.e.

$$p_{n;i}^{(m)} = \lim_{t \to \infty} P\left\{\xi(t) = n, \eta_1(t) + ... + \eta_m(t) = i\right\},$$

$$n \geq N, m = 1, ..., N-1, i \geq 0.$$

Note that joint stationary distribution $\{p_{n;i}, n \geq 1, i \geq 0\}$ of total number of customers in buffer and servers and total number of customers in RB equals

$$p_{n;i} = p_{n;i}^{(n)}, \ n = 1, ..., N-1, \ i \geq 0,$$

$$p_{n;i} = p_{n;i}^{(N-1)}, \ n \geq N, \ i \geq 0.$$

Moreover distribution $\{p_n, n \geq 0\}$ of the total number of customers in servers and buffer (already defined by 3.2.1, 3.2.2, 3.2.3) can be expressed through probabilities $p_{n;i}$ as follows

$$p_n = \sum_{i \geq 0} p_{n;i}, \ n = 1, ..., N-2, \ p_n = \sum_{i \geq 0} p_{n;i}, \ n \geq N-1.$$

3.3 The equilibrium state distribution

In order to compute joint stationary distribution an algorithm was developed, which led to recursive procedure for computation of all probabilities $p_{n;i}^{(m)}$. For probabilities $p_{n;i}^{(1)}$, $n \geq N$, $i \geq 0$, it holds:

$$p_{n;0}^{(1)}(\lambda + N\mu) = p_{n-1;0}^{(1)}\lambda + p_{n+1}(N-1)\mu, \quad n \geq N, \tag{3.3.4}$$

$$p_{n;i}^{(1)}(\lambda + N\mu) = p_{n-1;i}^{(1)}\lambda + p_{n+1;i-1}^{(1)}\mu, \quad n \geq N, \quad i \geq 1. \tag{3.3.5}$$

Probabilities $p_{N-1;i}^{(1)}$, $i \geq 0$, satisfy the following system of equations:

$$p_{N-1;0}^{(1)}[\lambda + (N-1)\mu] = p_{N-2}\lambda + p_N(N-1)\mu, \tag{3.3.6}$$

$$p_{N-1;i}^{(1)}[\lambda + (N-1)\mu] = p_{N;i-1}^{(1)}\mu, \quad i \geq 1. \tag{3.3.7}$$

Probabilities $p_{n;i}^{(1)}$, $n = \overline{1, N-2}$, $i \geq 0$, are expressed as follows:

$$p_{n;0}^{(1)}(\lambda + n\mu) = p_{n-1}\lambda + p_{n+1;0}^{(1)}n\mu, \quad n = \overline{1, N-2}, \tag{3.3.8}$$

$$p_{n;i}^{(1)}(\lambda+n\mu) = p_{n+1;i}^{(1)}n\mu + p_{n+1;i-1}^{(2)}\mu, \quad n = \overline{1, N-2}, \quad i \geq 1. \tag{3.3.9}$$

Other probabilities $p_{n;i}^{(m)}$, $m = \overline{2, N-1}$, are computed from the following relations:

$$p_{n;0}^{(m)}(\lambda + N\mu) = p_{n-1;0}^{(m)}\lambda + p_{n+1;0}^{(m-1)}(N-m)\mu, \quad n \geq N, \tag{3.3.10}$$

$$p_{n;i}^{(m)}(\lambda + N\mu) = p_{n-1;i}^{(m)}\lambda + p_{n+1;i}^{(m-1)}(N-m)\mu + p_{n+1;i-1}^{(m)}m\mu, \quad n \geq N, \quad i \geq 1, \tag{3.3.11}$$

$$p_{N-1;0}^{(m)}[\lambda + (N-1)\mu] = p_{N-2;0}^{(m-1)}\lambda + p_{N;0}^{(m-1)}(N-m)\mu, \tag{3.3.12}$$

$$p_{N-1;i}^{(m)}[\lambda+(N-1)\mu] = p_{N-2;i}^{(m-1)}\lambda + p_{N;i}^{(m-1)}(N-m)\mu + p_{N;i-1}^{(m)}m\mu, \quad i \geq 1, \tag{3.3.13}$$

$$p_{n;0}^{(m)}(\lambda + n\mu) = p_{n-1;0}^{(m-1)}\lambda + p_{n+1;0}^{(m)}(n-m+1)\mu, \quad n = \overline{m, N-2}, \quad m \neq N-1, \tag{3.3.14}$$

$$p_{n;i}^{(m)}(\lambda+n\mu) = p_{n-1;i}^{(m-1)}\lambda+p_{n+1;i}^{(m)}(n-m+1)\mu+p_{n+1;i-1}^{(m+1)}m\mu, \quad (3.3.15)$$
$$n = \overline{m, N-2}, m \neq N-1, \quad i \geq 1.$$

Now we describe them.

For probabilities $p_{n;i}^{(1)}$, $n \geq N$, $i \geq 0$, it holds:

- $p_{n;0}^{(1)}(\lambda + N\mu) = p_{n-1;0}^{(1)}\lambda + (N-1)p_{n+1}\mu, \quad n \geq N$. Analyze the state $(n;0)$ and the rate in and the rate out of this state. State $(n;0)$ means that there are n customers between buffer and service, 0 customers in queue 1 and we don't know how many customers there are in the other queues. Rate out of this state is $\lambda+N\mu$ because system can exit it either through arrival or through service. System can enter this state either: 1) by an arrival, there are $n-1$ customers between buffer and service and there are no customers in queue 1, 2) by service if all servers are busy, there are $n+1$ customers in buffer and servers. Equating rate-in and rate-out we get equation (3.3.4).

- $p_{n;i}^{(1)}(\lambda + N\mu) = p_{n-1;i}^{(1)}\lambda + p_{n+1;i-1}^{(1)}\mu, \quad n \geq N, \quad i \geq 1$. Analyze the state $(n;i)$ and the rate in and the rate out of this state. State $(n;i)$ means that there are n customers between buffer and service, i customers in queue 1 and we don't know how many customers there are in the other queues. Rate out of this state is $\lambda+N\mu$ because system can exit it either through arrival or through service. System can enter this state either: 1) by an arrival, there are $n-1$ customers between buffer and service and there are i customers in queue 1, 2) by service if all servers are busy, there are $n+1$ customers in buffer and servers, and $i-1$ customers in queue. Equating rate-in and rate-out we get equation (3.3.5).

Probabilities $p_{N-1;i}^{(1)}, \quad i \geq 0$, satisfy the following system of equations:

- $p_{N-1;0}^{(1)}[\lambda + (N-1)\mu] = p_{N-2}\lambda + p_N(N-1)\mu, \; i \geq 0$. Analyze the state $(N-1;0)$ and the rate in and the rate out

of this state. State $(N-1;0)$ means that there are $N-1$ customers between buffer and service, 0 customers in queue 1 and we don't know how many customers there are in the other queues. Rate out of this state is $\lambda+(N-1)\mu$ because system can exit it either through arrival or through service. System can enter this state either: 1) by an arrival, there are $N-2$ customers between buffer and service, 2) by service if all servers are busy, there are N customers in buffer and servers. Equating rate-in and rate-out we get equation (3.3.6).

- $p^{(1)}_{N-1;i}[\lambda+(N-1)\mu] = p^{(1)}_{N;i-1}\mu, \quad i \geq 1$. Analyze the state $(N-1;i)$ and the rate in and the rate out of this state. State $(N-1;i)$ means that there are $N-1$ customers between buffer and service, i customers in queue 1 and we don't know how many customers there are in the other queues. Rate out of this state is $\lambda+(N-1)\mu$ because system can exit it either through arrival or through service. System can enter this state: 1) by service if all servers are busy, there are N customers in buffer and servers. Equating rate-in and rate-out we get equation (3.3.7) .

Probabilities $p^{(1)}_{n;i}, \quad n = \overline{1, N-2}, \quad i \geq 0$, are expressed as follows:

- $p^{(1)}_{n;0}(\lambda+n\mu) = p_{n-1}\lambda + p^{(1)}_{n+1;0}n\mu, \quad n = \overline{1, N-2}$. Analyze the state $(n;0)$ and the rate in and the rate out of this state. State $(n;0)$ means that there are n customers between buffer and service, 0 customers in queue 1 and we don't know how many customers there are in the other queues. Rate out of this state is $\lambda+n\mu$ because system can exit it either through arrival or through service. System can enter this state either: 1) by an arrival, there are $n-1$ customers between buffer and service, 2) by service, there are $n+1$ customers in buffer and servers, 0 customers in queue 1. Equating rate-in and rate-out we get equation (3.3.8).

- $p^{(1)}_{n;i}(\lambda + n\mu) = p^{(1)}_{n+1;i} n\mu + p^{(2)}_{n+1;i-1}\mu, \quad n = \overline{1, N-2}, \quad i \geq 1.$ Analyze the state $(n; i)$ and the rate in and the rate out of this state. State $(n; i)$ means that there are n customers between buffer and service, i customers in queue 1 and we don't know how many customers there are in the other queues. Rate out of this state is $\lambda + n\mu$ because system can exit it either through arrival or through service. System can enter this state either: 1) by service, there are $n + 1$ customers in buffer and servers, i customers in queue 1, 2) by service, there are $n+1$ customers in buffer and servers, $i-1$ customers in queue 1 and queue 2. Equating rate-in and rate-out we get equation (3.3.9).

Other probabilities $p^{(m)}_{n;i}, \quad m = \overline{2, N-1}$, are computed from the following relations:

- $p^{(m)}_{n;0}(\lambda + N\mu) = p^{(m)}_{n-1;0}\lambda + p^{(m-1)}_{n+1;0}(N-m)\mu, \quad n \geq N$: suppose system is in state $n; 0$. It means that there are n customers between buffer and service, 0 customers in the first m queues and we don't know how many customers there are in the other queues. Rate out of this state is $\lambda + N\mu$ because system can exit it either through arrival or through service. Now, system can enter this state either: 1) by an arrival, there are $n - 1$ customers between buffer and service, 0 customers in the first m queues 2) by service if all servers are busy, there are $n + 1$ customers in buffer and servers, 0 customers in the first $m - 1$ queues. Equating rate-in and rate-out we get equation (3.3.10).

- $p^{(m)}_{n;i}(\lambda + N\mu) = p^{(m)}_{n-1;i}\lambda + p^{(m-1)}_{n+1;i}(N-m)\mu + p^{(m)}_{n+1;i-1} m\mu, \quad n \geq N, \quad i \geq 1$: suppose system is in state $n; i$. It means that there are n customers between buffer and service, i customers in the first m queues and we don't know how many customers there are in the other queues. Rate out of this state is $\lambda + N\mu$ because system can exit it either through arrival or through service. Now, system can enter this state: 1) by an arrival,

there are $n-1$ customers between buffer and service, i customers in the first m queues 2) by service if all servers are busy, there are $n+1$ customers in buffer and servers, i customers in the first $m-1$ queues, 3) by service if all servers are busy, there are $n+1$ customers in buffer and servers, $i-1$ customers in the first m queues. Equating rate-in and rate-out we get equation (3.3.11).

- $p^{(m)}_{N-1;0}[\lambda+(N-1)\mu] = p^{(m-1)}_{N-2;0}\lambda + p^{(m-1)}_{N;0}(N-m)\mu$: suppose system is in state $N-1;0$. It means that there are $N-1$ customers between buffer and service, 0 customers in the first m queues and we don't know how many customers there are in the other queues. Rate out of this state is $\lambda+(N-1)\mu$ because system can exit it either through arrival or through service. Now, system can enter this state either: 1) by an arrival, there are $N-2$ customers between buffer and service, 0 customers in the first $m-1$ queues 2) by service if all servers are busy, there are N customers in buffer and servers, 0 customers in the first $m-1$ queues. Equating rate-in and rate-out we get equation (3.3.12).

- $p^{(m)}_{N-1;i}[\lambda+(N-1)\mu] = p^{(m-1)}_{N-2;i}\lambda + p^{(m-1)}_{N;i}(N-m)\mu + p^{(m)}_{N;i-1}m\mu,\ \ i \geq 1$: suppose system is in state $N-1;i$. It means that there are $N-1$ customers between buffer and service, i customers in the first m queues and we don't know how many customers there are in the other queues. Rate out of this state is $\lambda+(N-1)\mu$ because system can exit it either through arrival or through service. Now, system can enter this state: 1) by an arrival, there are $N-2$ customers between buffer and service, i customers in the first $m-1$ queues 2) by service if all servers are busy, there are N customers in buffer and servers, i customers in the first $m-1$ queues, 3) by service if all servers are busy, there are N customers in buffer and servers, $i-1$ customers in the first m queues. Equating rate-in and rate-out we get equation (3.3.13).

- $p^{(m)}_{n;0}(\lambda+n\mu) = p^{(m-1)}_{n-1;0}\lambda + p^{(m)}_{n+1;0}(n-m+1)\mu,\ \ n = \overline{m, N-2}, m \neq$

$N-1$: suppose system is in state $n;0$. It means that there are n customers between buffer and service, 0 customers in the first m queues and we don't know how many customers there are in the other queues. Rate out of this state is $\lambda+n\mu$ because system can exit it either through arrival or through service. Now, system can enter this state either: 1) by an arrival, there are $n-1$ customers between buffer and service, 0 customers in the first $m-1$ queues 2) by service, there are $n+1$ customers in buffer and servers, 0 customers in the first m queues. Equating rate-in and rate-out we get equation (3.3.14).

- $p_{n;i}^{(m)}(\lambda+n\mu) = p_{n-1;i}^{(m-1)}\lambda+p_{n+1;i}^{(m)}(n-m+1)\mu+p_{n+1;i-1}^{(m+1)}m\mu,\ \ n = \overline{m, N-2}, m \neq N-1, \quad i \geq 1$: suppose system is in state $n;i$. It means that there are n customers between buffer and service, i customers in the first m queues and we don't know how many customers there are in the other queues. Rate out of this state is $\lambda+n\mu$ because system can exit it either through arrival or through service. Now, system can enter this state: 1) by an arrival, there are $n-1$ customers between buffer and service, i customers in the first $m-1$ queues 2) by service, there are $n+1$ customers in buffer and servers, i customers in the first m queues, 3) by service, there are $n+1$ customers in buffer and servers, $i-1$ customers in the first $m+1$ queues. Equating rate-in and rate-out we get equation (3.3.15).

3.4 Probability generating functions and numerical results

The analysis of steady-state equations resulted in the development of simple recursive algorithm for step-by-step computation of $p_{n;i}^{(m)}$. The pseudo-code of the algorithm is the following:

Initialize N, λ, μ;
Calculate p_0
for $1 \leq n \leq N$ **do**
  Calculate p_n from $\frac{\rho^n}{n!} p_0$
end for
for $n \geq N+1$ **do**
  Calculate p_n from $\frac{\rho^n}{N!\, N^{n-N}} p_0$
end for
Calculate $p_{N-1;0}^{(1)}$ from

$$p_{N-1;0}^{(1)}[\lambda + (N-1)\mu] = p_{N-2}\lambda + p_N(N-1)\mu,$$

for $n \geq N$ **do**
  Calculate $p_{n;0}^{(1)}$ from

$$p_{n;0}^{(1)}(\lambda + N\mu) = p_{n-1;0}^{(1)}\lambda + p_{n+1}(N-1)\mu,$$

end for

for $i \geq 1$ **do**
  Calculate $p_{N-1;i}^{(1)}$ from

$$p_{N-1;i}^{(1)}[\lambda + (N-1)\mu] = p_{N;i-1}^{(1)}\mu,$$

  for $n \geq N$ **do**
    Calculate $p_{n;i}^{(1)}$ from

$$p_{n;i}^{(1)}(\lambda + N\mu) = p_{n-1;i}^{(1)}\lambda + p_{n+1;i-1}^{(1)}\mu,$$

  end for

end for
for $n = N-2$ to 1 **do**
Calculate $p^{(1)}_{n;0}$ from

$$p^{(1)}_{n;0}(\lambda + n\mu) = p_{n-1}\lambda + p^{(1)}_{n+1;0}n\mu,$$

end for
for $m = 2$ to $N-1$ **do**
Calculate $p^{(m)}_{N-1;0}$ from

$$p^{(m)}_{N-1;0}[\lambda + (N-1)\mu] = p^{(m-1)}_{N-2;0}\lambda + p^{(m-1)}_{N;0}(N-m)\mu$$

for $n \geq N$ **do**

Calculate $p^{(m)}_{n;0}$ from

$$p^{(m)}_{n;0}(\lambda + N\mu) = p^{(m)}_{n-1;0}\lambda + p^{(m-1)}_{n+1;0}(N-m)\mu,$$

end for
for $i \geq 1$ **do**
Calculate $p^{(1)}_{N-m;i}$ from

$$p^{(1)}_{n;i}(\lambda + n\mu) = p^{(1)}_{n+1;i}n\mu + p^{(2)}_{n+1;i-1}\mu$$

if $m \neq 2$ then
for $j = 2$ to $m-1$ **do**
Calculate $p^{(j)}_{N-m+j-1;i}$ from

$$p^{(m)}_{n;i}(\lambda + n\mu) = p^{(m-1)}_{n-1;i}\lambda + p^{(m)}_{n+1;i}(n-m+1)\mu + p^{(m+1)}_{n+1;i-1}m\mu,$$

end for
end if
Calculate $p^{(m)}_{N-1;i}$ from

$$p^{(m)}_{N-1;i}[\lambda + (N-1)\mu] = p^{(m-1)}_{N-2;i}\lambda + p^{(m-1)}_{N;i}(N-m)\mu + p^{(m)}_{N;i-1}m\mu$$

for $n \geq N$ **do**

$$p_{n;i}^{(m)}(\lambda + n\mu) = p_{n-1;i}^{(m-1)}\lambda + p_{n+1;i}^{(m)}(n - m + 1)\mu + p_{n+1;i-1}^{(m+1)}m\mu, \ ,$$

end for
end for
if $m \neq N - 1$ **then**
for $n = N - 2$ to m **do**
Calculate $p_{n;0}^{(m)}$ with formula

$$p_{n;0}^{(m)}(\lambda + n\mu) = p_{n-1;0}^{(m-1)}\lambda + p_{n+1;0}^{(m)}(n - m + 1)\mu,$$

end for
end if
end for

3.5 Numerical example

There are several quantities related to the number of customers in the system that may be of interest. We have calculated mean and variance of the number of customers in RB, correlation between queue size in buffer and RB. These quantities are depicted in the following figure for different number N of servers in the system. Along the x-axis values of system's load (ρ/N) are indicated, along the y-axis we indicate the corresponding value of mean number of customers in reordering buffer (Figure 3.1(a)), variance of the number of customers in reordering buffer (Figure 3.1(b)), correlation on the number of customers in queue and reordering buffer (Figure 3.1(c)). In all examples service rate $\mu = 1$. We can observe that when the number of servers increases, the values of mean number of customers in reordering buffer increases and this happens because using more servers, the service is speeded, but in this way it is not ensured that the customers end the service in the same order of their arrival in service. It is worth noticing that correlation between queue sizes in buffer and RB is almost insignificant. In Figure 3.2 one can see the behaviour of joint stationary distribution $\{p_0, p_{n;i}, n \geq 1, i \geq 0\}$ when the number of servers $N = 5$ and system's load ρ/N takes values 0.5, 0.7 and 0.9. We observe that when the value of ρ/N increases we have more often values of $p_{n;i}$ different from zero.

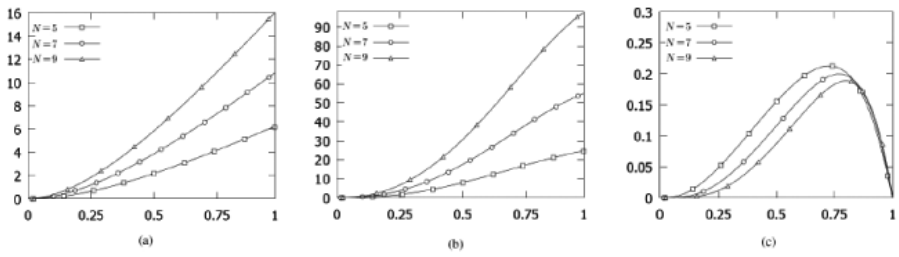


Figure 3.1 Dependence on load ρ/N of (a) mean number of customers in reordering buffer, (b) variance of number of customers in reordering buffer, (c) correlation on the number of customers in queue and reordering buffer.

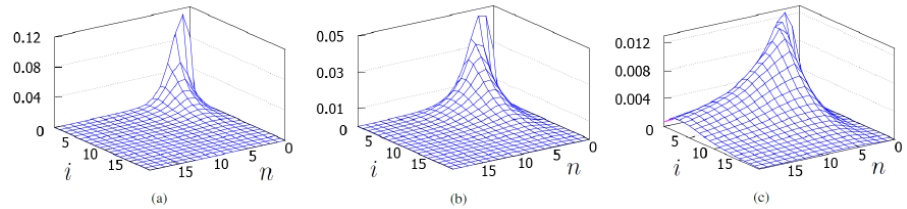


Figure 3.2 Join stationary distribution $p_{\pi;i}$ (a) $\rho/5 = 0.5$, (b) $\rho/5 = 0.7$, (c) $\rho/5 = 0.9$.

Chapter 4

System MAP/PH/2 with resequencing

4.1 Problem statement

We have a queueing system with 2 homogeneous servers, in which the capacity of the collecting buffer and the reordering buffer is infinite. The type distribution of both two servers is "the phase distribution" (PH), while the arrivals follow Markovian arrival process. We have two type of request: the request that, in order to leave the system, have to wait one request that is still in service, and the request that, in order to leave the system, have to wait for two requests still in service. We introduce a recurrent algorithm to calculate the simultaneous stationary distribution of the number of the requests at servers, in the collecting buffer and in the reordering buffer. Then we calculate the stationary distribution of the arrival time of the requests in the buffer and in service, and the stationary distribution of the arrival time of the requests in the reordering buffer in terms of Laplace-Stieltjes transform using geometric-matrix methods.

4.2 Model description and notation

The queueing system studied has 2 homogeneous servers, collecting buffer and reordering buffer with infinite capacity. The arrival process is Markovian with n generation phases. The matrix of the intensity of change of the generation phases of the requests without the arrival of the requests is N, while the matrix of the intensity of change of the generation phases of the requests with the arrival of the requests is Λ. The service time for both servers is the same phase type distribution with parameters $(\vec{f}, G)$, where $\vec{f} = (f_1, \ldots, f_m)$ is a row vector of dimension m, and $G = (g_{ij})_{i,j=\overline{1,m}}$ is a square matrix of dimension m, $1 \leq m < \infty$. The request arrives at the server with probability $f_i \quad i = \overline{1,m}$, and it is served starting from the phase i. If in a certain period of time the server serves the request in the phase i, then in a "small" period of time Δ with probability $g_{ij}\Delta + o(\Delta)$, $i, j = \overline{1,m}$, the phase of the service changes into the $j-th$ phase. The service of the request finishes with probability $g_i^\star \Delta + o(\Delta)$, where

$$g_i^\star = -\sum_{j=1}^{m} g_{ij}.$$

With $\vec{g} = -G\vec{1}$ is indicated a column-vector with coordinate $g_i^\star$, where $\vec{1}$ is a column-vector whose elements are all equal to 1. When the requests arrive in the system, a sequential number is given at each of them. The order given by these serial numbers is kept for requests when they are leaving the queueing system. The requests that have finished the service earlier and interrupted the order, have to wait in the reordering buffer and they can leave the queueing system only after that the requests with lower sequential number have finished the service. In the case that both servers are busy, the server to whom the request arrives earlier is called "primary server" , and the server to whom the request arrives later is called "secondary server". Here λ is the stationary intensity of the arrivals, and μ is the intensity of service:

$$\lambda = \vec{\pi}^{(A)}\Lambda\vec{1}, \qquad \mu = -(\vec{f}G^{-1}\vec{1})^{-1},$$

where $\vec{\pi}^{(A)}$ is the row-vector of the stationary probabilities of the Markovian process with infinitesimal matrix $(N + \Lambda)$. It is supposed that the sufficient and necessary condition $\rho = \lambda/(2\mu) < 1$ for existence of stationary behavior of the queueing system holds true.

4.3 Stationary state probabilities

Let $\xi(t)$ be the number of the requests at servers and in the collecting buffer at the time moment t, $\eta(t)$ the number of the requests in the reordering buffer (RB) at the time t, $\alpha(t)$ the request generation phase at the time t, $\beta_1(t)$ the distribution service phase of the primary server at time t, $\beta_2(t)$ the distribution service phase of the secondary server at time t. Initial suppositions, concerning the input flux and the service process, guarantee that the random process $\zeta(t) = \{(\xi(t), \eta(t), \alpha(t), \beta_1(t), \beta_2(t)),\ t \geq 0\}$ is Markovian. We note that when $\xi(t) = 0$ then the second, the fourth and the fifth component of $\zeta(t)$ is undefined, while when $\xi(t) = 1$ then the fifth component is undefined. The state set of this process has the following form:

$$\mathcal{X} = \left\{(0,i),\ \ i=\overline{1,n}\right\} \cup \left\{(1,k,i,j)\,,\ \ k\geq 0,\ \ i=\overline{1,n},\ \ j=\overline{1,m}\right\} \cup$$

$$\cup \left\{(u,k,i,j,l)\,,\ \ u\geq 2,\ \ k\geq 0,\ \ i=\overline{1,n},\ \ j,l=\overline{1,m}\right\}.$$

Now we want to compute the simultaneous stationary distribution of the number of requests at servers and in the collecting buffer, and the number of the requests in the reordering buffer. We introduce the following notations:

$p_0(i)$ denotes stationary probability of the process $\zeta(t)$ being in the state $(0,i)$, $i=\overline{1,n}$;

$p_{1,k}(i,j)$ denotes stationary probability of the process $\zeta(t)$ being in the state $(1,k,i,j)$, $k\geq 0$, $i=\overline{1,n}$, $j=\overline{1,m}$;

$p_1(i,j)=\sum_{k=0}^{\infty} p_{1,k}(i,j)$ denotes stationary probability of having only one request in the system, generation phase equals to i, and service phase equals to j, $i=\overline{1,n}$, $j=\overline{1,m}$;

$p_{u,k}(i,j,l)$ denotes stationary probability of the process $\zeta(t)$ being in the state (u,k,i,j,l), $u\geq 2$, $k\geq 0$, $i=\overline{1,n}$, $j,l=\overline{1,m}$;

$p_u(i,j,l)=\sum_{k=0}^{\infty} p_{u,k}(i,j,l)$, $u\geq 2$, $i=\overline{1,n}$, $j,l=\overline{1,m}$ denotes stationary probability of the fact that there are u requests at servers and in the collecting buffer, the generation phase is i, the service phase of the primary server is j, while the service phase of the secondary server is l.

We set:
$\vec{p_0}$ be a vector with coordinates $p_{0,i} = p_0(i)$, $i = \overline{1,n}$;
$\vec{p}_{1,k}$, $k \geq 0$ be a vector with coordinates $p_{1,k,z}$, $z = \overline{1,nm}$, where $p_{1,k,z} = p_{1,k}(i,j)$ for $z = (i-1)m + j$, $i = \overline{1,n}$, $j = \overline{1,m}$;
$\vec{p}_1$ be a vector with coordinates $p_{1,z}$, $z = \overline{1,nm}$, where $p_{1,z} = p_1(i,j)$ for $z = (i-1)m + j$, $i = \overline{1,n}$, $j = \overline{1,m}$; $\vec{p}_{u,k}$, $u \geq 2$, $k \geq 0$ denotes vector with coordinates $p_{u,k,z}$, $z = \overline{1,nm^2}$, where $p_{u,k,z} = p_{u,k}(i,j,l)$, when $z = (i-1)m^2 + (j-1)m + l$, $i = \overline{1,n}$, $j,l = \overline{1,m}$;
$\vec{p}_u$, $u \geq 2$ be a vector with coordinates $p_{u,z}$, $z = \overline{1,nm^2}$, where $p_{u,z} = p_u(i,j,l)$ for $z = (i-1)m^2+(j-1)m+l$, $i = \overline{1,n}$, $j,l = \overline{1,m}$.

Before writing the system of the equilibrium equations for stationary probabilities, we introduce some other notations:

$$\Lambda_0 = \Lambda \otimes \vec{f}, \quad \Lambda_1 = \Lambda \otimes E \otimes \vec{f}, \quad \Lambda^* = \Lambda \otimes E \otimes E,$$

$$M_1 = E \otimes \vec{g}, \quad M_{2,1} = E \otimes E \otimes \vec{g}, \quad M_{2,2} = E \otimes \vec{g} \otimes E, \quad M_2 = M_{2,1} + M_{2,2},$$

$$M_1^* = E \otimes E \otimes \vec{g} \otimes \vec{f}, \quad M_2^* = E \otimes \vec{g} \otimes E \otimes \vec{f}, \quad M^* = M_1^* + M_2^*,$$

$$N_1 = N \otimes E + E \otimes G, \quad N^* = N \otimes E \otimes E + E \otimes E \otimes G + E \otimes G \otimes E.$$

where $\otimes$ is the Kronecker product.
The meaning of the matrices is the following:

- $\Lambda_0 = \Lambda \otimes \vec{f}$: this matrix rappresents the passage from the state 0 requests in the buffer and in service to 1 request in the buffer and service. The matrix Λ indicates that there is the arrival of a request and $\vec{f}$ indicates that this request goes immediately in service.

- $\Lambda_1 = \Lambda \otimes E \otimes \vec{f}$: this matrix rappresents the passage from the state 1 request in the buffer and in service to 2 requests in the buffer and service. The matrix Λ indicates that there is the arrival of a request, the matrix E suggests that the primary server goes on serving the customer in service and $\vec{f}$ indicates that the second request goes immediately in service.

- $\Lambda^* = \Lambda \otimes E \otimes E$: this matrix rappresents the passage from the state u request in the buffer and in service to $u+1$ requests in the buffer and service. The matrix Λ indicates that there is the arrival of a request and, because of both servers are busy, this customer has to wait in the buffer, the matrix E suggests that the primary server goes on serving the customer in service and the other matrix E indicates that the secondary server goes on serving the customer in service.

- $M_1 = E \otimes \vec{g}$: this matrix rappresents the passage from the state 1 request in the buffer and in service to 0 request in the buffer and service. The matrix E indicates that happens nothing in the arrival and the vector $\vec{g}$ suggests that the request that was in service has finished the service.

- $M_{2,1} = E \otimes E \otimes \vec{g}$: this matrix rappresents the passage from the state 2 requests in the buffer and in service to 1 request in the buffer and service. The first matrix E indicates that happens nothing in the arrival, the second matrix E suggests that the primary server goes on serving the customer in service and the vector $\vec{g}$ indicates that the request that was in service in the secondary server has finished the service.

- $M_{2,2} = E \otimes \vec{g} \otimes E$: this matrix rappresents the passage from the state 2 requests in the buffer and in service to 1 request in the buffer and service. The first matrix E indicates that happens nothing in the arrival, the vector $\vec{g}$ suggests that the request that was in service in the primary server has finished the service and the last matrix E indicates that the primary server goes on serving the customer in service.

- $M_1^* = E \otimes E \otimes \vec{g} \otimes \vec{f}$: this matrix rappresents the passage from the state u requests in the buffer and in service to $u-1$ request in the buffer and service. The first matrix E indicates that happens nothing in the arrival, the second one suggests that the primary server goes on serving the customer in service, the vector $\vec{g}$ indicates that the request

that was in service in the secondary server has finished the service and $\vec{f}$ says that the request that was waiting in the buffer goes immediately in service.

- $M_2^* = E \otimes \vec{g} \otimes E \otimes \vec{f}$: this matrix rappresents the passage from the state u requests in the buffer and in service to $u-1$ request in the buffer and service. The first matrix E indicates that happens nothing in the arrival, the vector $\vec{g}$ suggests that the request that was in service in the primary server has finished the service, the second matrix E indicates that the secondary server goes on serving the customer in service and $\vec{f}$ says that the request that was waiting in the buffer goes immediately in service.

- $N_1 = N \otimes E + E \otimes G$: this matrix rappresents the passage from the state 1 requests in the buffer and in service to 1 request in the buffer and service. The system can go to the state 1 to 1 in two ways: in the first one (as indicated by matrix N)there are no arrival of the requests in the system and there is only one virtual customer, (as indicated by matrix E) happen nothing in the service; in the second one (as indicated by matrix E) happens nothing in the arrival and (as indicated by matrix G) the only request in service is still in service.

- $N^* = N \otimes E \otimes E + E \otimes E \otimes G + E \otimes G \otimes E$: this matrix rappresents the passage from the state u requests in the buffer and in service to u request in the buffer and service. The system can remain in the state u in three different ways: in the first (as indicated by N) there are no arrival of the requests in the system and there are only u virtual customers, (as indicated by the first matrix E) happens nothing in the primary server, (as indicated by the second matrix E) happens nothing in the secondary server; in the second one (as indicated by the first matrix E) happens nothing in the arrival, (as indicated by the second matrix E) happens nothing in the secondary server, the matrix G says that the requests in service are

still in service; in the third one the matrix E indicates that happens nothing in the arrival, the matrix G suggests that the requests in service are still in service, the last matrix E indicates that happens nothing in the secondary server.

Now we observe that the number of requests at servers and in the collecting buffer in the queueing system (QS) with reordering is equal to the number of the requests in the similar QS without reordering. For this reason the stationary distribution $\vec{p} = (\vec{p}_0, \vec{p}_1, \ldots)$ of the number of requests at servers and in the collecting buffer for the considered system is defined by the same formulas as for a usual QS MAP/PH/2/∞, i.e. this distribution must satisfy the following system of the equilibrium equations:

$$\vec{p}T = \vec{0} \tag{4.3.1}$$

with a normalization condition:

$$\vec{p}\vec{1} = 1, \tag{4.3.2}$$

where the infinitesimal transition matrix T is a tridiagonal block matrix:

$$T = \begin{pmatrix} N & \Lambda_0 & 0 & 0 & 0 & \ldots \\ M_1 & N_1 & \Lambda_1 & 0 & 0 & \ldots \\ 0 & M_2 & N^* & \Lambda^* & 0 & \ldots \\ 0 & 0 & M^* & N^* & \Lambda^* & \ldots \\ 0 & 0 & 0 & M^* & N^* & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{4.3.3}$$

The problem of finding a stationary distribution for Markov process with an infinitesimal matrix of the type T is well studied in [11], [12]. For this reason we do not linger on how the vector $\vec{p}$ is found, we only mention that its components $\vec{p}_i$ are solutions of

the following system:

$$\begin{gathered}\vec{p}_0 N + \vec{p}_1 M_1 = 0,\\ \vec{p}_0 \Lambda_0 + \vec{p}_1 N_1 + \vec{p}_2 M_2 = 0,\\ \vec{p}_1 \Lambda_1 + \vec{p}_2 N^* + \vec{p}_3 M^* = 0,\\ \vec{p}_i = \vec{p}_2 R^{i-2}, \quad i \geq 2,\\ \vec{p}_0 \vec{1} + \vec{p}_1 \vec{1} + \sum_{i=2}^{\infty} \vec{p}_i \vec{1} = 1,\end{gathered}$$

where the matrix R is the minimal non-negative solution of the matrix equation:

$$\Lambda^* + RN^* + R^2 M^* = 0.$$

Introduce the vector $\vec{\pi} = (\vec{\pi}_1, \vec{\pi}_2, \ldots)$, that is composed of vectors $\vec{\pi}_1 = (\pi_{1,1}, \ldots, \pi_{1,m})$, $\vec{\pi}_i = (\pi_{i,1}, \ldots, \pi_{i,m^2})$, $i \geq 2$. Each coordinate $\pi_{i,j}$ of the vector $\vec{\pi}_i$ is the stationary probability of the fact that right after the arrival of the request in the queueing system we have i requests at servers and in the collecting buffer. Here the service phase equals to j when $i = 1$, while when $i \geq 2$ the service phase can be found from the formula $j = (k-1)m+l$, where k and l denote the service phases of the primary and secondary server respectively. The vectors $\vec{\pi}_i$ are defined by the following relations:

$$\vec{\pi}_1 = \frac{1}{\lambda} \vec{p}_0 \Lambda_0 (\vec{1} \otimes E) = \frac{1}{\lambda} \vec{p}_0 \Lambda \vec{1} \otimes \vec{f},$$

$$\vec{\pi}_2 = \frac{1}{\lambda} \vec{p}_1 \Lambda_1 (\vec{1} \otimes E \otimes E) = \frac{1}{\lambda} \vec{p}_1 (\Lambda \otimes E)(\vec{1} \otimes E)(E \otimes \vec{f}),$$

$$\vec{\pi}_i = \frac{1}{\lambda} \vec{p}_{i-1} \Lambda^* (\vec{1} \otimes E \otimes E), \quad i \geq 3.$$

Now let us calculate the simultaneous stationary distribution of the number of requests at servers, in the collecting buffer and in the reordering buffer. It can be easily shown that for probabilities $\vec{p}_{1;i}$, $k \geq 0$ the following equations are satisfied:

$$\vec{p}_{1,0} N_1 + \vec{p}_0 \Lambda_0 + \vec{p}_2 M_{2,2} = 0, \tag{4.3.4}$$

$$\vec{p}_{1,k}N_1 + \vec{p}_{2,k-1}M_{2,1} = 0, \quad k \geq 1. \tag{4.3.5}$$

For probabilities $\vec{p}_{2,j}$, $k \geq 0$, we have

$$\vec{p}_{2,0}N^* + \vec{p}_{1,0}\Lambda_1 + \vec{p}_3 M_2^* = 0, \tag{4.3.6}$$

$$\vec{p}_{2,k}N^* + \vec{p}_{1,k}\Lambda_1 + \vec{p}_{3,k-1}M_1^* = 0, \quad k \geq 1. \tag{4.3.7}$$

And, finally, for probabilities $\vec{p}_{u,k}$, $u \geq 3$, $k \geq 0$, the equations

$$\vec{p}_{u,0}N^* + \vec{p}_{u-1,0}\Lambda^* + \vec{p}_{u+1}M_2^* = 0, \quad u \geq 3, \tag{4.3.8}$$

$$\vec{p}_{u,k}N^* + \vec{p}_{u-1,k}\Lambda^* + \vec{p}_{u+1,k-1}M_1^* = 0, \quad u \geq 3, \quad k \geq 1 \tag{4.3.9}$$

hold true. All equations (4.3.4)–(4.3.9) are based on the global equilibrium principle. Now we describe them. For probabilities $\vec{p}_{1;k}$, $k \geq 0$ it holds:

- $\vec{p}_{1,0}N_1 + \vec{p}_0\Lambda_0 + \vec{p}_2 M_{2,2} = 0, k = 0$. Analyze the state $(1;0)$ and the rate in and the rate out of this state. State $(1;0)$ means that there is one customer in service, the markovian phase is i, the service phase is j and there are no customers in the RB. Rate out of this state is given by the matrix N_1. System can enter this state either: 1) by an arrival, there are 0 customers between buffer and service, the Markovian phase is i, there are no customers in the RB, 2) by service if all servers are busy, so there are 2 customers in servers, the Markovian phase is i, the service phase is j for one server and l for the other one and there are no customers in the RB. Equating rate-in and rate-out we get equation (4.3.4).

- $\vec{p}_{1,k}N_1 + \vec{p}_{2,k-1}M_{2,1} = 0, \quad k \geq 1$. Analyze the state $(1;k)$ and the rate in and the rate out of this state. State $(1;k)$ means that there is one customer in service, the Markovian phase is i, the service phase is j and there are k customers in the RB. Rate out of this state is given by the matrix N_1. Now, system can enter this state only: 1) by service if all servers are busy, so there are 2 customers in servers, the Markovian phase is i, the service phase is j for one server and

l for the other one and there are $k-1$ customers in the RB. The customer that has to be served is that whose sequential number is bigger than the sequential number of customer who is still in service. Equating rate-in and rate-out we get equation (4.3.5).

For probabilities $\vec{p}_{2,j}$ it holds:

- $\vec{p}_{2,0}N^* + \vec{p}_{1,0}\Lambda_1 + \vec{p}_3 M_2^* = 0, k = 0$. Analyze the state $(2;0)$ and the rate in and the rate out of this state. State $(2;0)$ means that there are 2 customers in service, the Markovian phase is i, the service phase is j for one server and l for the other one, there are no customers in RB. Rate out of this state is given by the matrix N^*. System can enter this state either: 1) by an arrival, there is one customer in service, the Markovian phase is i, there are no customers in the RB, 2) by service if all servers are busy, so there are 3 customers between buffer and service, 2 of them are in service, the Markovian phase is i, the service phase is j for one server and l for the other one and there are no customers in the RB. Equating rate-in and rate-out we get equation (4.3.6).

- $\vec{p}_{2,k}N^* + \vec{p}_{1,k}\Lambda_1 + \vec{p}_{3,k-1}M_1^* = 0, \quad k \geq 1$. Analyze the state $(2;k)$ and the rate in and the rate out of this state. State $(2;k)$ means that there are 2 customers in service, the Markovian phase is i, the service phase is j for one server and l for the other one, there are k customers in RB. Rate out of this state is given by the matrix N^*. System can enter this state either: 1) by an arrival, there is one customer in service, the Markovian phase is i, there are k customers in the RB, 2) by service if all servers are busy, so there are 3 customers between buffer and service, 2 of them are in service, the Markovian phase is i, the service phase is j for one server and l for the other one and there are $k-1$ customers in the RB. Equating rate-in and rate-out we get equation (4.3.7).

For probabilities $\vec{p}_{u,k}$, $u \geq 3$, $k \geq 0$ it holds:

- $\vec{p}_{u,0}N^* + \vec{p}_{u-1,0}\Lambda^* + \vec{p}_{u+1}M_2^* = 0, \quad u \geq 3$. Analyze the state $(u;0)$ and the rate in and the rate out of this state. State $(u;0)$ means that there are u customers between buffer and service, the markovian phase is i, the service phase is j for one server and l for the other one, there are no customers in RB. Rate out of this state is given by the matrix N^*. System can enter this state either: 1) by an arrival, there are $u-1$ customers between buffer and service, the markovian phase is i, there are no customers in the RB, 2) by service if all servers are busy, so there are $u+1$ customers between buffer and service, 2 of them are in service, the markovian phase is i, the service phase is j for one server and l for the other one and there are no customers in the RB. Equating rate-in and rate-out we get equation (4.3.8).

- $\vec{p}_{u,k}N^* + \vec{p}_{u-1,k}\Lambda^* + \vec{p}_{u+1,k-1}M_1^* = 0, \quad u \geq 3, \quad k \geq 1$. Analyze the state $(u;k)$ and the rate in and the rate out of this state. State $(u;k)$ means that there are u customers between buffer and service, the markovian phase is i, the service phase is j for one server and l for the other one, there are k customers in RB. Rate out of this state is given by the matrix N^*. System can enter this state either: 1) by an arrival, there are $u-1$ customers between buffer and service, the markovian phase is i, there are k customers in the RB, 2) by service if all servers are busy, so there are $u+1$ customers between buffer and service, 2 of them are in service, the markovian phase is i, the service phase is j for one server and l for the other one and there are $k-1$ customers in the RB. Equating rate-in and rate-out we get equation (4.3.9).

Analyzing the equations we obtain the following simple algorithm that allows to find successively the stationary probabilities $\vec{p}_{u,k}$, $u \geq 1$, $k \geq 0$.

- Setting of $\Lambda, N, G, \vec{f}$.

- Finding $\vec{p}_n$ for $n \geq 0$.

- Finding $\vec{p}_{1,0}$ from formula (4.3.4).
- Finding $\vec{p}_{2,0}$ from formula (4.3.6).
- Finding $\vec{p}_{n,0}$ for $n \geq 3$ from formula (4.3.8).
- Finding $\vec{p}_{1,i}$ for $i \geq 1$ from formula (4.3.5).
- Finding $\vec{p}_{2,i}$ from formula (4.3.7)
- Find $\vec{p}_{n,i}$ for $n \geq 3$ from formula (4.3.9)

4.4 Stationary distribution of the in-service time of a request in the queueing system

From now on we use some notations that are used before, but with a different meaning. First of all let us consider the Markov process $\{\tilde{\gamma}(t), \quad t \geq 0\}$ with a state set $\tilde{\mathcal{Y}} = \tilde{\mathcal{Y}}_0 \cup \tilde{\mathcal{Y}}_1$, where

$$\{\tilde{\mathcal{Y}}_k = (k, i), \quad k = 0, 1, \quad i = \overline{1, m^2}\},$$

the matrix $\tilde{M} = E \otimes \vec{g} \otimes \vec{f} + \vec{g} \otimes E \otimes \vec{f} = (\tilde{m}_{i,j})_{i,j=\overline{1,m^2}}$ of transition intensities when passing from the state $(1, i) \in \tilde{\mathcal{Y}}_1$ to the state $(0, j) \in \tilde{\mathcal{Y}}_0$ and the matrix $\tilde{N} = E \otimes G + G \otimes E = (\tilde{n}_{i,j})_{i,j=\overline{1,m^2}}$ of transition intensities when passing from the state $(1, i) \in \tilde{\mathcal{Y}}_1$ to the state $(1, j) \in \tilde{\mathcal{Y}}_1$.

Let us find a square matrix $\tilde{R}(s)$ of the elements $\tilde{r}_{i,j}(s)$ that are the Laplace-Stieltjes transformations of the time of the first passage from the state $(1, i)$ to the state $(0, j)$ and the probability of the fact that when the process leaves the state set $\tilde{\mathcal{Y}}_1$ at the first time, it passes to the state $(0, j)$.

We note that, that from the point of view of the queueing system with reordering, $\tilde{r}_{i,j}(s)$ is the Laplace-Stieltjes transformation of the service termination time for the first one of the two requests that are at servers, and the probability of the fact that at the same time the service process will pass to the state j, where $j = (k-1)m+l$ and by k and l denote the service phase of the primary server and of the secondary server respectively. We suppose that at the start (zero) time there are no less than two requests in the queueing system and that the service process is in the state i where $i = (u-1)m + v$, and by u and v we denote the service phase of the primary and the secondary server respectively.

We can use the embedded Markov chain generated by moments of change of the state of the process $\tilde{\gamma}(t)$. The matrix $\tilde{Q}^{(0)}$ of transition probabilities from the state $(1, i) \in \tilde{\mathcal{Y}}_1$ to the state

$(0, j) \in \tilde{\mathcal{Y}}_0$ for this Markov chain consists of the elements:

$$\tilde{q}_{i,j}^{(0)} = -\frac{\tilde{m}_{i,j}}{\tilde{n}_{i,i}}, \quad i, j = \overline{1, m^2}.$$

Elements of the matrix $\tilde{Q}^{(1)}$ of transition probabilities from the state $(1, i) \in \tilde{\mathcal{Y}}_1$ to the state $(1, j) \in \tilde{\mathcal{Y}}_1$ of the Markov chain are given by the formulas:

$$\tilde{q}_{i,j}^{(1)} = -\frac{\tilde{n}_{i,j}}{\tilde{n}_{i,i}}, \quad i, j = \overline{1, m^2}, \quad j \neq i,$$

$$\tilde{q}_{i,i}^{(1)} = 0, \quad i = \overline{1, m^2}.$$

Finally, since the time of the Markov process $\tilde{\gamma}(t)$ staying in the state $(1, i) \in \tilde{\mathcal{Y}}_1$ has exponential distribution with the parameter $-\tilde{n}_{i,i}$, then considering the matrix $\tilde{Q}^{(k)}(s)$ $k = 0, 1$, whose element $\tilde{q}_{i,j}^{(k)}(s)$ is the Laplace-Stieltjes transformation of the time of the first passage of the process $\tilde{\gamma}(t)$ from the state $(1, i) \in \tilde{\mathcal{Y}}_1$ to the state $(k, j) \in \tilde{\mathcal{Y}}_0$ and the probability of the fact that, going out of the state $(1, i)$, the process $\tilde{\gamma}(t)$ passes immediately to the state (k, j), we have:

$$\tilde{q}_{i,j}^{(k)}(s) = \frac{\tilde{n}_{i,i}}{\tilde{n}_{i,i} - s} \tilde{q}_{i,j}^{(k)}, \quad k = 0, 1, \quad i, j = \overline{1, m^2}.$$

Now we can write the equation for $\tilde{R}(s)$:

$$\tilde{R}(s) = \tilde{Q}^{(0)}(s) + \tilde{Q}^{(1)}(s)\tilde{R}(s),$$

and after solving it we have:

$$\tilde{R}(s) = [E - \tilde{Q}^{(1)}(s)]^{-1} \tilde{Q}^{(0)}(s).$$

We introduce the vector $\vec{W}(s) = (w_1(s), \ldots, w_{m^2}(s))$, where $w_i(s)$ – is the Laplace-Stieltjes transformation of the stationary distribution of the waiting-in-queue time and the probability of the fact that right after the arrival of a request in the queueing system, there will be no less than two requests, and at the moment

when the service of the distinguished request starts the process, $\tilde{\gamma}(t)$ will be in the state i, i.e.

$$\vec{W}(s) = \sum_{i=2}^{\infty} \vec{\pi}_i \tilde{R}^{i-2}(s). \tag{4.4.10}$$

Further we consider the Markov process $\{\hat{\gamma}(t), \quad t \geq 0\}$ with a state set $\hat{\mathcal{Y}} = \hat{\mathcal{Y}}_0 \cup \hat{\mathcal{Y}}_1$, where

$$\{\hat{\mathcal{Y}}_0 = (0,i), \quad i = \overline{1,m}\},$$

$$\{\hat{\mathcal{Y}}_1 = (1,i), \quad i = \overline{1,m^2}\},$$

with the transition intensity matrix $\hat{M}_0 = E \otimes \vec{g} + \vec{g} \otimes E = (\hat{m}_{0,i,j})_{i=\overline{1,m^2},j=\overline{1,m}}$ of for transactions from the state $(1,i) \in \hat{\mathcal{Y}}_1$ to the state $(0,j) \in \hat{\mathcal{Y}}_0$ and $\hat{N} = \tilde{N} = E \otimes G + G \otimes E = (\hat{n}_{i,j})_{i,j=\overline{1,m^2}}$ is transition intensity matrix for transactions from the state $(1,i) \in \hat{\mathcal{Y}}_1$ to the state $(1,j) \in \hat{\mathcal{Y}}_1$.

Let us find the matrix $\hat{R}(s) = (\hat{r}_{i,j})_{i=\overline{1,m^2},j=\overline{1,m}}$, whose element $\hat{r}_{i,j}(s)$ are the Laplace-Stiltjes transformation of the time of the first passage from the state $(1,i) \in \hat{\mathcal{Y}}_1$ to the state $(0,j) \in \hat{\mathcal{Y}}_0$ and the probability of the fact that right after the first exit from the state set $\hat{\mathcal{Y}}_1$ the process will pass to the state $(0,j)$. From the point of view of the initial queueing system with reordering, the element $\hat{r}_{i,j}(s)$ is the Laplace-Stieltjes transformation of the service termination time for the first one of the two requests that are at servers and the probability of the fact that at the same time the service phase of the request at the other server will be equal to j. We suppose that at the initial moment of time there are no less than two requests at the queueing system and the service process is in the state i , where $i = (u-1)m + v$ and by u and $v-$ we denote the service phase of the primary and secondary server respectively.

We can use the embedded Markov chain generated by moments of the state changes of the process $\hat{\gamma}(t)$. Elements of matrix $\hat{Q}^{(0)}$ of transition probabilities from the state $(1,i) \in \hat{\mathcal{Y}}_1$ to the state

$(0,j) \in \hat{\mathcal{Y}}_0$ of this Markov chain are defined by the following relations:

$$\hat{q}_{i,j}^{(0)} = -\frac{\hat{m}_{i,j}}{\hat{n}_{i,i}}, \quad i = \overline{1,m^2}, \quad j = \overline{1,m}.$$

Elements of the matrix $\hat{Q}^{(1)}$ of transition probabilities from the state $(1,i) \in \hat{\mathcal{Y}}_1$ to the state $(1,j) \in \hat{\mathcal{Y}}_1$ of the considered Markov chain can be found from formulas:

$$\hat{q}_{i,j}^{(1)} = -\frac{\hat{n}_{i,j}}{\hat{n}_{i,i}}, \quad i,j = \overline{1,m^2}, \quad j \neq i,$$

$$\hat{q}_{i,i}^{(1)} = 0, \quad i = \overline{1,m^2}.$$

Finally, since the time of the Markov process $\hat{\gamma}(t)$ staying in the state $(1,i) \in \hat{\mathcal{Y}}_1$ has exponential distribution with the parameter $-\hat{n}_{i,i}$, then, introducing the matrix $\hat{Q}^{(k)}(s)$, $k = 0,1$, whose element $\hat{q}_{i,j}^{(k)}(s)$ is the Laplace-Stieltjes transformation of the time of the first passage of the process $\hat{\gamma}(t)$ from the state $(1,i) \in \hat{\mathcal{Y}}_1$ to the state $(k,j) \in \hat{\mathcal{Y}}_0$ and the probability of the fact that going out from the state $(1,i)$, the process $\hat{\gamma}(t)$ immediately passes to the state (k,j), we have:

$$\hat{q}_{i,j}^{(0)}(s) = \frac{\hat{n}_{i,i}}{\hat{n}_{i,i} - s}\hat{q}_{i,j}^{(k)}, \quad i = \overline{1,m^2}, \quad j = \overline{1,m},$$

$$\hat{q}_{i,j}^{(1)}(s) = \frac{\hat{n}_{i,i}}{\hat{n}_{i,i} - s}\hat{q}_{i,j}^{(1)}, \quad i,j = \overline{1,m^2}.$$

The equation for $\hat{R}(s)$ takes the form:

$$\hat{R}(s) = \hat{Q}^{(0)}(s) + \hat{Q}^{(1)}(s)\hat{R}(s),$$

and its solution is given by formula:

$$\hat{R}(s) = [E - \hat{Q}^{(1)}(s)]^{-1}\hat{Q}^{(0)}(s).$$

Assuming that at the initial point of time the distribution of the request service phase was $\vec{f}$, the Laplace-Stieltjes transformation of the in-service time of the request takes the following well known form:

$$\varphi(s) = -\vec{f}(sE - G)^{-1}G\vec{1}.$$

Now we can write the formula for the Laplace-Stieltjes transformation $V(s)$ of the stationary distribution of the total in-service time of the request in the queueing system (including the time spent in the reordering buffer). This time consists of two terms, the first one is the time from the moment of the request arrival in the queueing system to the moment of its entry at the server. While the second term is the time from the moment of the request entry at the server to its exit from the system, that, in its turn, is the sum of the in-service time of this distinguished request and also of the in-service time of the request served by the second server at the moment of arrival of this distinguished request (it is evident that when a request arrives at the empty queueing system, the total time of this request is only its own in-service time). So

$$V(s) = -\vec{\pi}_1(sE - G)^{-1}G\vec{1} - \vec{W}(s)\hat{R}(s)(sE - G)^{-1}G\vec{1}. \quad (4.4.11)$$

At the end of this section we will write down formulas for the Laplace-Stieltjes transformation $\psi(s)$ of the stationary distribution of the arrival time of the request in the reordering buffer.

We introduce $\vec{W}(0) = (w_1(0), \ldots, w_{m^2}(0))$ − this is a vector whose components are $w_i(0)$,where $w_i(0) = w_{(u-1)m+v}(0)$ − are the stationary probabilities that at the beginning of the service of a request there are not less than two requests in the system, while the primary and the secondary request in the servers will be served in the phases u and v.

We look at the Markov process $\{\check{\gamma}(t), \quad t \geq 0\}$ with multiple states $\check{\mathcal{Y}} = \check{\mathcal{Y}}_0 \cup \check{\mathcal{Y}}_1 \cup \check{\mathcal{Y}}_2$, where

$$\{\check{\mathcal{Y}}_k = (k, i), \quad k = 0, 1, \quad i = \overline{1, m}\},$$

$$\{\check{\mathcal{Y}}_2 = (2, i), \quad i = \overline{1, m^2}\},$$

the matrix $\check{M}_0 = E \otimes \vec{g} = (\check{m}_{0,i,j})_{i=\overline{1,m^2},j=\overline{1,m}}$ is the matrix of the intensity of transactions from the state $(2, i) \in \check{\mathcal{Y}}_2$ to the state $(0, j) \in \check{\mathcal{Y}}_0$, the matrix $\check{M}_1 = \vec{g} \otimes E = (\check{m}_{1,i,j})_{i=\overline{1,m^2},j=\overline{1,m}}$ is the matrix of the intensity of transactions from the state $(2, i) \in \check{\mathcal{Y}}_2$ to the state $(1, j) \in \check{\mathcal{Y}}_1$ and the matrix $\check{N} = \hat{N} = \tilde{N} = E \otimes G +$

$G \otimes E = (\check{n}_{i,j})_{i,j=\overline{1,m^2}}$ is the matrix of the intensity of transactions from the state $(2,i) \in \check{\mathcal{Y}}_2$ to the state $(2,j) \in \check{\mathcal{Y}}_2$. The matrix $\check{R} = (\check{r}_{i,j})_{i=\overline{1,m^2},j=\overline{1,m}}$, is composed of elements $\check{r}_{i,j}$ that indicate the passage of the process $\check{\gamma}(t)$ from the initial state $(2,i)$ to the finale state $(0,j)$. From the point of view of the queueing system with reordering, $\check{r}_{i,j}$ is the probability that at the beginning the request in the secondary server is served and at the end of this service the service phase of the request in the primary server will be j, at the initial moment there are not less than two requests in the system and the service process was in the state i, where $i = (u-1)m + v$, with u and v the service phase in the primary and secondary server. We use the Markov chain given by moments of changes of states of the process $\hat{\gamma}(t)$. The matrix $\check{Q}^{(0)}$ of the transient probabilities in the Markov chain from the state $(2,i) \in \check{\mathcal{Y}}_2$ to the state $(0,j) \in \check{\mathcal{Y}}_0$ is composed by the following elements

$$\check{q}_{i,j}^{(0)} = -\frac{\check{m}_{0,i,j}}{\check{n}_{i,i}}, \quad i = \overline{1,m^2}, \quad j = \overline{1,m}.$$

The elements of the matrix of the transient probabilities $\check{Q}^{(2)}$, given in the Markov chain, from the state $(2,i) \in \check{\mathcal{Y}}_2$ to the state $(2,j) \in \check{\mathcal{Y}}_2$ are given by the formula:

$$\check{q}_{i,j}^{(2)} = -\frac{\check{n}_{i,j}}{\check{n}_{i,i}}, \quad i,j = \overline{1,m^2}, \quad j \neq i,$$

$$\check{q}_{i,i}^{(2)} = 0, \quad i = \overline{1,m^2}.$$

The equation of the matrix $\check{R}$ is:

$$\check{R} = \check{Q}^{(0)} + \check{Q}^{(2)}\check{R},$$

whose solution is:

$$\check{R} = [E - \check{Q}^{(2)}]^{-1}\check{Q}^{(0)}.$$

Then the Laplace-Stieltjes form (PLS) $\psi(s)$ of the stationary distribution of the arrival time of the request in the reordering buffer is:

$$\psi(s) = -\vec{W}(0)\check{R}(sE - G)^{-1}G\vec{1},$$

while the mean time of the arrival in the reordering buffer is:

$$-\psi'(0) = -\vec{W}(0)\check{R}G^{-1}\vec{1}.$$

The PLS T(s) of the stationary distribution of the arrival time of the request in the buffer and in server is the same of that in MAP/PH/2 system without reordering:

$$T(s) = -\vec{\pi}_1(sE - G)^{-1}G\vec{1} - \vec{W}(s)(\vec{1} \otimes E)(sE - G)^{-1}G\vec{1},$$

and its mean value is given by the formula:

$$-T'(0) = -\vec{\pi}_1 G^{-1}\vec{1} - \vec{W}'(0)\vec{1} - \vec{W}(0)(\vec{1} \otimes E)G^{-1}\vec{1}.$$

We observe that differentiating the formula (4.3.11) we can find moments of differents order of stationary distruibution of the arrival time of the request in the system. The stationary mean time $\bar{v}$ of the arrival of the request in the system is given by the formula:

$$\bar{v} = -V'(0) = -\vec{\pi}_1 G^{-1}\vec{1} - \vec{W}'(0)\vec{1} - \vec{W}(0)\hat{R}'(0)\vec{1} - \vec{W}(0)\hat{R}(0)G^{-1}\vec{1}. \tag{4.4.12}$$

4.5 Numerical examples

We introduce a recurrent algorithm that allows to calculate the simultaneous stationary distribution of the number of the requests in the buffer and in service, and the number of the requests in the reordering buffer. We have also calculated in terms of PLS the stationary distribution of the waiting in-queue time of the requests, the stationary distribution of the total in-service time of the requests in the queueing system with its moments, the stationary distribution of the arrival time of the request in buffer and server with its moments, and the stationary distribution of the arrival time of the requests in the reordering buffer with its moments. In all example parameters of the phase distribution of the service time are the same, in order to intensify the service of the requests on each server the service parameter is equal to one. The curves on the Figure 4.1 show the mean and the variance of the number of requests in the reordering buffer, and finally the coefficient of correlation of the number of the requests in the buffer and in the reordering buffer from the time of the arrival in the system, that in this case coincide with the intensification of the parameter λ. From the graphic is evident that this two quantities are almost uncorrelated, because of, with the definition chosen for the parameter of the system and of the arrival, the system is stationary. As it is evident in [13] an alike result is observed in the case when the arrivals follow a Poisson process and the service time is exponential on each server. In Figure 4.2 we can see the mean time of arrival of requests in the system, the mean time of arrival in the buffer and in server, the mean time of reordering from the arrival in the reordering buffer, and finally the variance of time of arrival of requests in the system. When the requests increase, we have a queue in the buffer, and the mean and the variance of the waiting time from the beginning of the service. The mean time of the arrival in the reordering buffer (and moments of bigger order) is limited, this is evident from the formula of Little. We can observe that the variation of the parameter of the service $(\vec{f}, G)$ affect the number of the requests waited in the reordering buffer, whereas

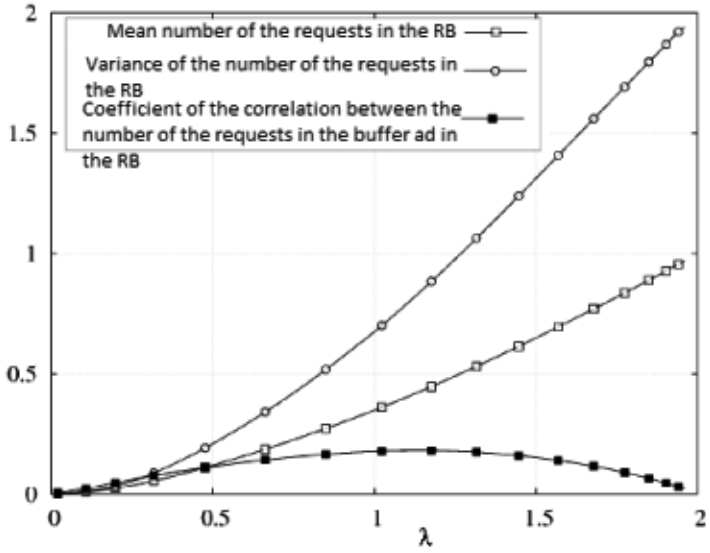


Figure 4.1 Mean and variance of the number of the requests in the reordering buffer and the coefficient of correlation of the number of the requests in the reordering buffer and in the buffer.

the parameter of the MAP arrival process don't affect it. In fact, as example we consider two systems: in the fist one the arrival process is defined by the couple (N_1, Λ_1), while in the second one the couple is (N_2, Λ_2) where

$$N_1 = \begin{pmatrix} -2 & 2 \\ 0 & -2 \end{pmatrix} \qquad \Lambda_1 = \begin{pmatrix} 0 & 0 \\ 2 & 0 \end{pmatrix}$$

$$N_2 = \begin{pmatrix} -0.5452 & 0 \\ 0 & -16.3565 \end{pmatrix} \qquad \Lambda_2 = \begin{pmatrix} 0.0818 & 0.4634 \\ 15.6477 & 0.7088 \end{pmatrix}$$

For both systems the mean time of arrival in the system is equal to $\frac{1}{\lambda} = 1$, while the values of the variance of the time of arrival and of the coefficient of correlation between the mean time of arrival in the buffer and the mean time of the arrival in the reordering buffer are in the table 1.

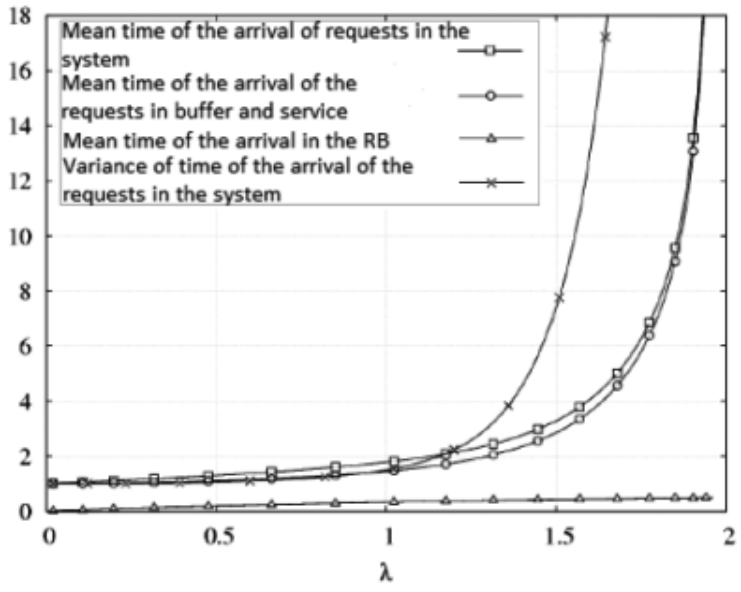


Figure 4.2 Moments of the arrival time of the requests in the system and in the reordering buffer.

Table 1

	(N_1, Λ_1)	(N_2, Λ_2)
Variance	0.5	2.566
Coefficient of correlation	0	-0.246

We observe that $\rho = \frac{1}{2\mu}$ and for both systems the parameter of the service are the same. From the Figure 4.3 we can observe that for $\rho > 0.2$ the mean time of the arrival in both systems start to diverge, but the difference of the MAP process don't affect on the mean time of the arrival of he requests in the reordering buffer that only in the range $0.2 < \rho < 0.8$ are different, while for value of ρ out of this range tend to the same value.

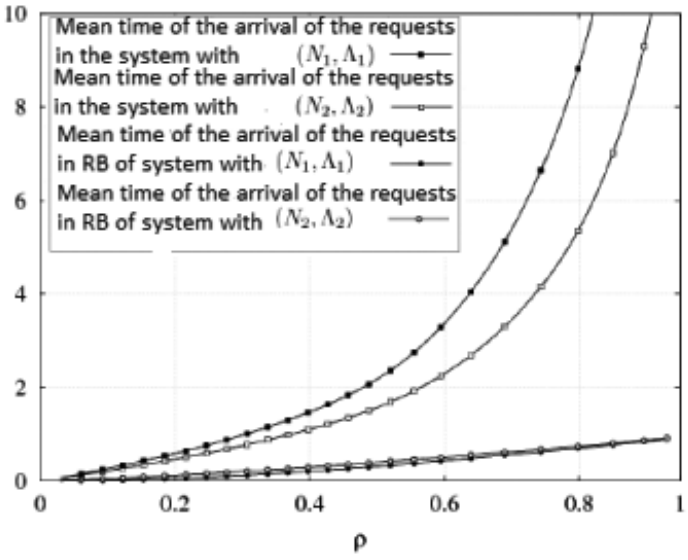


Figure 4.3 Moments of the distribution of the mean time of the arrival of the requests in the system and in the reordering buffer for both systems with a different values of the parameters of the MAP process and the same values for the parameters of the service process.

Conclusions

In this work we have studied three different queueing systems characterized by resequencing. First of all we have analyzed the M/M/3/∞, of infinity capacity, with resequencing. We have noticed that customer in reordering buffer may form two separate queues, so focus is given to the study of their size distribution. We present results of the thorough analysis of joint stationary distribution (both explicit and in terms of generating functions). We have shown numerically that, for the all possible range of load values, correlation between any queues that are formed in the system is almost insignificant. Then we have generalized the M/M/3/∞ introducing a number of servers bigger than 3 and less than infinity. The analysis of steady-state equations resulted in the development of simple recursive algorithm for step-by-step computation of stationary probability of the fact that there are n customers in servers and buffer and total number of customers in first m queues in reordering buffer (RB) is equals i. We study mean and variance of the number of customers in RB, correlation between queue size in buffer and RB. Further research will be devoted to analysis of joint stationary distribution of number of customers in all queues (i.e. buffer and N - 1 queues in RB) and study of behaviour of different performance characteristics. For example in [13]for the case N = 3 we have shown numerically that queues in RB and queues of RB and buffer are almost uncorrelated. Case N $>$ 3 remains an open issue. In the last studied system we generalize the arrival and the service flow introducing MAP and PH process. We found a recurrent algorithm that allows to calculate the simultaneous stationary distribution of the number of the requests in the buffer and in service, and the number of the requests in the reordering buffer. We have also calculated in terms of PLS the stationary distribution of the waiting in-queue time of the requests,the stationary distribution of the total in-service time of the requests in the queueing system with its moments, the stationary distribution of the arrival time of the request in buffer and server with its moments, and the stationary distribution of the arrival time of

the requests in the reordering buffer with its moments. Further study will be devoted to the analysis of stationary distribution of the number of the requests in the buffer and in service, and the number of the requests in the reordering buffer in more complex system with possibly arbitrary number of servers.

Bibliography

[1] Nitin R. G. and Shivendra S. P., Assigning customers to two parallel servers with resequencing. IEEE COMMUNICATIONS LETTERS. 1999. Vol. 3, no. 4.

[2] Nitin G. and Shivendra S. P., On a Resequencing Model for High Speed Networks. Department of Electrical Engineering Polytechnic University 6 Metrotech Center, Brooklyn, NY 11201. 1994.

[3] Jun L., Yifeng Z., Lamont L., Minyi H., Zhao Y.Q., Probabilistic Analysis of Resequencing Queue Length in Multipath Packet Data Networks. IEEE Global Telecommunications Conference (GLOBECOM 2010). P. 1-5.

[4] Iliadis I. and Yeong-Chang L., Resequencing in distributed systems with multiple classes. INFOCOM Networks: Evolution or Revolution, Proceedings. Seventh Annual Joint Conference of the IEEE Computer and Communcations Societies, IEEE. 1988. P. 881-888.

[5] Ayoun S. and Rosberg Z., Optimal routing to two parallel hetergeneous servers with resequencing. IEEE Trans. Autom. Control. 1991. Vol. 36, no. 12, P. 1436-1449.

[6] Chowdhury S., Distribution of the total delay of packets in virtual circuits. Proceedings of the Tenth Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 1991). Vol. 2. P. 911-918.

[7] Boxma O., Koole G. and Liu Z., Queueing-theoretic solution methods for models of parallel and distributed systems // Performance Evaluation of Parallel and Distributed Systems Solution Methods. 1994. P. 1-24.

[8] Dimitrov B., Queues with resequencing. A survey and recent results. Proceedings 2-ndWorld Congress on Nonlinear Analysis, Theory, Methods, Applications. 1997. Vol. 30, no. 8. P. 5447-5456.

[9] Chakravarthy S., Chukova S. and Dimitrov B., Analysis of MAP/M/2/K Queueing Model with Resequencing. Performance Evaluation. 1998. Vol. 31. P. 211-228.

[10] Chakravarthy S. and Chukova S., A Finite Capacity Resequencing Model with Markovian Arrivals // Asia-Pacific Journal of Operational Research. 2005. Vol. 22. P. 409-443.

[11] Neuts M., Matrix-geometric solutions in stichastic models. An algorithmic approach. Baltimore: The John Hopkins University Press, 1981.

[12] Latouche G. and Ramaswami V., Introduction to Matrix Analytic Methods in Stochastic Modeling. ASA-SIAM Series on Statistics and Applied Probability Series, 1999.

[13] Pechinkin A., Caraccio I. and Razumchik R., Joint Stationary Distribution Of Queues In Homogenous M—M—3 Queue With Resequencing. Proc. of the 28th European Conference on Modelling and Simulation. 2014. P. 558-564.

[14] Bocharov P.P., D'Apice C., Pechinkin A.V. and Salerno S., Queueing Theory, Modern Probability and Statistics, VSP, The Netherlands, 2004.

[15] Xia Y., D.N.C, On the large deviations of resequencing queue size: 2-M/M/1 Case. IEEE Transactions on information theory, 2008. Vol. 54, no. 9. P. 4107-4118.

[16] Bilgen S., Altintas O., An Approximate Solution for the Resequencing Problem in Packet-Switching Networks. IEEE TRANSACTIONS ON COMMUNICATIONS, 1994. Vol 42, no. 21314.

[17] Harrus G., Plateau B., Queueing Analysis of a Reordering Issue. IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, 1982, Vol.8, no. 2.

[18] Chowdhury S.,The Mean Resequencing Delay for M/HK/oo Systems.IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, 1989, Vol. 15, no. 12.

[19] Lelarge M., Packet reordering in networks with heavy-tailed delays. Math Meth Oper Res, 2008, Vol 67. P. 341-371

[20] Alain J. M., Levent G., Parallel Queues with Resequencing. Journal of the Association for Cumputmg M chmety, 1993, Vol 40, no 5. P. 1188-1208.