UNIVERSITÀ DEGLI STUDI DI SALERNO
Dipartimento di Fisica "E. R. Caianiello" e Dipartimento di Matematica

in convenzione con

UNIVERSITÀ DEGLI STUDI DELLA CAMPANIA
"LUIGI VANVITELLI"
Dipartimento di Matematica e Fisica

Dottorato di Ricerca
"Matematica, Fisica ed Applicazioni"
XXIX Ciclo
Curriculum Matematica

Tesi di dottorato

MAP/PH/1 systems with group service: performance analysis under different admission strategies

Candidato: **Brugno Arianna**

Coordinatore:	**Prof. Sandro Pace**
Tutor:	**Prof.ssa Giovannina Albano**
Cotutor:	**Prof. Alexander Dudin**
Responsabile Scientifico:	**Prof. Ciro D'Apice**

Anno Accademico 2015-2016

Contents

List of Figures

Introduction

Everyone needs time, time to work, to eat, to sleep, to devote to relationships and to accomplish all the daily chores of living. Most of the time of everyone's life is spent in waiting for something.
Any system in which a customer has to wait to receive the service represents a *queueing system*. It could be a telephone exchange, a supermarket, a petrol station, a computer system, etc.
The first studies about queueing systems are due to A. K. Erlang, an employee of the Copenhagen Telephone Company, at the beginning of the 20th century. His results regarded, in detail, the waiting time until a call could be handled by an operator in a telephone network, and the probability of a call blocking due to the unavailability of free lines. These studies have increased the interest of many scientists towards the field of queueing theory, especially in order to give a major contribute to the numerous challenges raised by the industrial sector: improvement of operational service quality and service efficiency, analysis of loss probability of a product or informations, etc. Queueing theory attempts to address these challenges from a mathematical perspective.

Every queueing system is characterized by two major components: the external arrival process and the service one. The external arrival process governs the timing of service request arrivals to that system from outside, and the service process concerns the duration of service transactions in that system. Since the arrival and service processes are usually stochastic by nature, the study of queueing systems involves probabilistic analysis.

On the other hand, since every queueing system requires its own special research tool, and analytical or numerical calculation is often not possible and appropriate approximations are not available, a serious competing tool is becoming established, the simulation modeling.

Queueing systems can be classified on the basis of several characteristics, some of which are: modality of access to the system and service processing, number of servers, discipline which rules the assignment of a customer to the server.

The main objective of the research reported in this work is to analyse some $MAP/PH/1$ queueing systems, introducing novel strategies of admission, with the aim to apply the results to real life communication systems. Such queueing systems can be used to model multi-rate transmission technologies.

Exponentiality of the involved distributions is a common assumption when modelling the performance of communication systems through queueing models. The interest of the exponential law is justified by the mathematical tractability of the resulting Markovian models. It has been shown that the Markovian Arrival Process and the phase type distribution provide alternatives to generalize, respectively, the Poisson flow and the exponential distribution, and still keep the tractability for modelling purposes. For these reasons, in the queueing systems introduced and analysed in this work of thesis, it is assumed that both service time and length of admission period follow phase type distributions, while arrivals are regulated by a Markovian Arrival Process.

Another characteristic of the systems under study is the presence of a single server which offers service to a group of customer simultaneously. Such a type of service has been chosen to model the recently appeared technologies of multi-rate transmission, e.g., multi-rate IEEE802.11 WLAN, where it is assumed simultaneous service of a group of customers, but not just as a usual group, while as one entity with service time defined as the maximum of service times of individual customers. Since the expectation of the maximum of a fixed number of independent random variables is less (and can be much less) than the sum of expectations of these random variables, the average time devoted to the service of an arbitrary customer under the proposed service discipline may be much less than this time under the classical service discipline. Thus, the throughput of the systems under the proposed service discipline is higher and other performance measures of the system may be much better compared to those obtained by the classical discipline.

Such systems were analyzed in communication literature via simulation. No results in terms of queueing theory for such a type of systems were available in literature before.

The thesis is organized as follows.
Chapter 1 is devoted to a brief introduction to the queueing theory. A description of the structure and the main components of a queueing system will be given, and an overview on the models currently present in literature will be done.
In the Chapter 2 a novel admission strategy for a retrial queueing system will be introduced [14]. Server alternates between two states: service providing period and customers admission period. During the admission period, primary customers are accepted to a pool of finite capacity where they wait for the service. If an arriving customer finds the pool full, then he moves to a virtual place called orbit, from which can make repeated attempts to be served after random periods of time. Duration of the admission interval is random and depends on the number of customers in the pool when the admission period begins. However, if the pool becomes full earlier than the duration of admission period expires, this period is terminated and service period begins. Duration of service period depends on the number of customers in the pool at this period beginning epoch. In order to analyse this strategy, the ergodicity condition for the system under study and the main performance indices will be defined. To optimize the performance measures, stationary probability distribution of the states of a multi-dimensional Markov chain describing dynamics of the system will be computed for any fixed set of the system parameters. Advantages of the proposed customer's service discipline will be numerically illustrated, comparing them with the classical discipline and the one in which admission period does not expires when the pool becomes full.
In the Chapter 3 an evolution of the previous strategy will be proposed ad analysed [15]. The main difference is that the access to the pool is not locked during the service period, so the service process can restart immediately if the pool results full at the end of the service period. Some key performance measures will be computed, and the essential advantages of the proposed customer's service discipline will be numerically shown, comparing them also

to the results obtained in the previous chapter.

Retrial is no more considered in the model introduced in the Chapter 4, which is characterised by a batch service discipline [16]. Service to customers is offered in batches of a certain size. If the number of customers in the system at a service completion moment is less than this size, the server does not start next service until the number of customers in the system will reach this size or a random limitation of the idle time of the server will expire, whichever occurs first. Dynamics of such a system is described by a multidimensional Markov chain. Ergodicity condition for this Markov chain will be derived, stationary probability distribution of the states will be computed, formulas for the main performance measures of the system will be attained. Laplace-Stieltjes transform of waiting time will be obtained. Results will numerically illustrate the advantages of this model.

Chapter 1

About Queueing Systems

Queueing theory deals with one of the most unpleasant experiences of life, waiting. Queueing is quite common in many fields, for example, in telephone exchange, in a supermarket, at a petrol station, at computer systems, etc.

It is easy to predict the further course of events. Since it is the random events that are to be blamed and since it is the probability theory that deals with random events, queueing systems must be studied using probabilistic techniques. Thus emerged another branch of the probability theory, the queueing theory, whose founder is believed to be an employee of the Copenhagen Telephone Company, the eminent Danish scientist A.K. Erlang. He was the first who used Markov processes with a discrete (finite or countable) state set to describe queueing systems. Queueing theory became a field of applied probability and many of its results have been used in operative research, computer science, telecommunication, traffic engineering, reliability theory, just to mention some. Later, interest in queueing theory somewhat waned for several reasons. It could be examined one of them, a mathematical one. On the one hand, a typical feature of queueing problems is that almost every queueing system requires its own special research tool and, on the other hand, the great interest in queueing theory has already yielded solutions to many problems that allow simple solutions (especially, in the computational sense). Moreover, analytical methods of studying queueing systems were eventually enhanced with a serious competing tool: simulation modeling. Recently, interest in queueing theory has been revived not only as a result of new applied problems related, particularly, to the development and application of computers, but also due to the advent of new mathematical approaches to their solution. One such approach is the algorithmic one, which is a consequence of the extensive use of computers, especially personal computers, in research and it provides solutions to queueing problems in the form of computer algorithms. The algorithmic approach, though inferior to the traditional analytical methods as regards clarity of results, applicability for optimization, etc., has an indisputable advantage of being oriented towards developing applied software and tables, appreciated in applications much higher than even very elegant formulas.

1.1 Structure of a queueing system

Every queueing system is characterized by these basics components:

- *The input process.* It refers to the arrival of a flow of customers to the system.
- *The service process.* It regards the basic characteristics of the server(s) and the distribution and dependencies of the service times.
- *The system capacity.* It concerns the number of customers that can wait at any given time in a queueing system.
- *The queueing discipline.* It is the rule followed by the server(s) for choosing customers for service.

On the basis of these components, a queueing system can be defined, following the basic classification/notation 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).

1.1.1 Input process

The input process regards the process of arrivals of customers in the queueing system. The access to the system can be various.

The first distinction can be made considering the possibility that customers can try to enter the system singly or in groups. The latter situation is called “bulk arrival”, and is very common in real life. Some examples of such type of arrival are letters arriving at a post office, ships arriving at a port in convoy, people going to a theatre, and so on.
The study of bulk arrival queues may be said to have begun with Erlang's solution of the $M/E_K/1$ queue [13]. The major contributions on bulk arrival queues are made by many researchers, see [42, 43, 62, 63].
After entering the systems in group, customers can be served one by one or they can receive the service in batch, simultaneously. If this last situation also constitutes a characteristic of the system, then it is called “bulk system” because defined by bulk arrival and service.

Moreover, customers arriving to the system can have the same characteristics (so they belong to an only flow) or can be different and separated in several flows (the so called “k flows” if there are k classes of customers). In this case, every flow can be differently handled by the server, e.g. a flow can have the priority, another needs to be processed only after a particular one, etc.

Another distinction can be made on the basis of the customers' behavior when they try the access to the system but the buffer is full. In classical queueing theory it is usually assumed that a customer who can not get service immediately after arrival, either joins the waiting line and then is served according to some queueing discipline, or leaves the system forever. However, in a real

situation, customers which do not find place in queue or impatient customers that decide to leave the system, after a random time return to the system and try to get service again. Systems in which these situations are allowed, are called "retrial queueing systems".
The following are just a few examples which explain this general remark more in detail.

- Telephone systems: everybody knows from his/her own experience that a telephone subscriber who obtains a busy signal repeats the call until the required connection is made. As a result, the flow of calls circulating in a telephone network consists of two parts: the flow of primary calls, which reflects the real wishes of the telephone subscribers, and the flow of repeated calls, which is the consequence of the lack of success of previous attempts.

- Retail shopping queue: in a shop, a customer who finds that a queue is too long may wish to do something else and return later on with the hope that the queue dissolves. Similar behavior may demonstrate some impatient customers who entered the waiting line but then discovered that the residual waiting time was too long.

- Random access protocols in digital communication networks: considering a communication line with slotted time which is shared by several stations, the duration of the slot equals the transmission time of a single packet of data. If two or more stations are transmitting packets simultaneously then a collision takes places, i.e. all packets are destroyed and must be retransmitted. If the stations involved in the conflict would try to retransmit destroyed packets in the nearest slot, then a collision occurs with certainty. To avoid this, each station independently of other ones, transmits the packet with probability p and delays actions until the next slot with probability 1-p, or equivalently, each station introduces a random delay before next attempt to transmit the packet.

In recent years, there has been an increasing interest in the investigation of the retrial phenomenon, especially in communication systems. The operational rules of the random access protocols in computer networks provide a major motivation for the design and control of retransmission policies. In most local area networks (LANs), it is feasible for a station to listen to the channel and, if it is sensed busy, the station reschedules the transmission of the packet to a time later. As a second source of motivation, the cellular networks can be mentioned, where efficient call handling mechanisms greatly improve the quality of service and the network performance. Thus, proper modelling of the mobile cellular network cannot ignore the existence of repeated calls generated by those fresh subscribers who find all the channels of a base station busy. For these reasons, their applications to optical fiber communication networks, circuit-switched networks with hybrid fiber-coax, telephone call centers, fast reservation protocols for asynchronous transfer mode (ATM) networks, etc., are very interesting.

The general structure of a retrial queue is shown in the following figure.

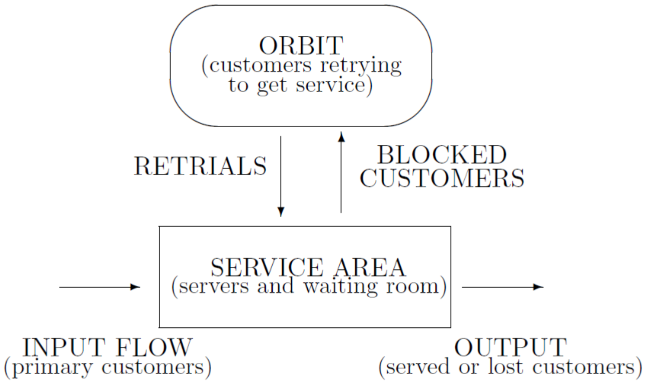

ORBIT
(customers retrying to get service)
RETRIALS
BLOCKED CUSTOMERS
SERVICE AREA
(servers and waiting room)
INPUT FLOW
(primary customers)
OUTPUT
(served or lost customers)

Figure 1.1 General structure of a retrial queue

It is clear from the picture that retrial queues can also be re-

garded as a special type of queueing networks. In their basic form, these networks contain two nodes: the main node where blocking is possible and a delay node for repeated retrials.
In retrial queues, it is usually assumed that if an arriving customer meets at least one idle server, he immediately starts the service. Otherwise, the arriving customer moves to some virtual place called orbit from which it will try to get the service, independently on other customers staying in orbit, later on.
The mean difference between a queue with a buffer and a retrial queueing system is the state of the server: while in the first one the server is permanently busy until the queue becomes empty, in retrial queues the server always stays idle during a random time interval after each service completion moment until a primary customer arrives or a customer retries the access to the system to make an attempt to get a service. Therefore, some period of time is wasted after the service of each customer.
Retrial queueing systems have been introduced and intensively studied in the queueing literature. The early work of Kosten [52] and Cohen [27] shows that retrial queues are suitable mathematical models for the modelling of subscribers' behavior in telephone networks. For other references see, e.g., the books [5, 37] and the bibliographies [3, 38, 70].

1.1.1.1 Input flows

Whatever customers arriving to the queueing system try the access, it is very important to know and describe the distribution and dependencies of the interarrival times (times between successive arrivals). The input flow is distinguished by the random instants of customer arrivals to the system and, for more complicated systems, also by the types of the customers arriving at these instants. It is usually assumed that the input flow is of Poisson nature, although sometimes consideration is given to the recurrent flows or Markov flows. Systems with more general input flows, in particular, with general stationary flows, are in fact unexplored, and at best only partial results on existence of stationary state of operation and convergence to them were obtained.

Random flow

The random flow $\{\tau_k, k \geq 1\}$ over the time interval $[0, \infty)$ is a nondecreasing sequence of random instants $\tau_1, \tau_2, ..., \tau_k, ..., \tau_1 \geq 0$, of advent of certain events. More specifically, it is the flow of customers (lost and/or served) arriving to or departing from the system.

Usually, the random flows are described in probabilistic terms in either of two ways. The first way lies in defining for any $k \geq 1$ the joint distribution function

$$F_{\xi_1,\xi_2,...,\xi_k}(x_1, x_2, ..., x_k) = \mathbf{P}\{\xi_1 < x_1, \xi_2 < x_2, ..., \xi_k < x_k\}$$

of the times $\xi_i = \tau_i - \tau_{i-1}, i \geq 1, \tau_0 = 0$, between the subsequent customer arrivals.

The second way is based on considering arbitrary assemblies of intervals $[0, t_1), [t_1, t_2), ..., [t_{k-1}, t_k), k \geq 1$, and defining the joint distributions

$$G(m_1, m_2, ..., m_k; t_1, t_2, ..., t_k) = \mathbf{P}\{\nu_1 = m_1, ..., \nu_k = m_k\}$$

of the numbers $\nu_i, i = \overline{1, k}$, of customers arriving over the intervals $[t_{i-1}, t_i)$.

Three properties of the random flow can be conveniently highlighted: memoryless, stationarity, and ordinariness:

(i) Memoryless implies independence of the random variables $\nu_1, ...,$ ν_k, that is, the number of customers arriving over some interval $[t_{i-1}, t_i)$ does not affect in any way the number of customers arriving over the rest of the intervals $[t_{j-1}, t_j), i \neq j$;

(ii) Stationarity means that the probabilistic properties of the flow do not vary with time, that is, the numbers of customers arriving over the intervals $[0, t_1), [t_1, t_2), ..., [t_{k-1}, t_k)$, are distributed in the same manner as the numbers of customers arriving over the intervals $[T, t_1 + T), [t_1 + T, t_2 + T), ..., [t_{k-1} + T, t_k + T)$ for any $T > 0$ and $k \geq 1$;

(iii) Ordinariness of the flow implies that customers can arrive only one at a time.

Poisson flow

The random flow with all three before mentioned properties (memoryless, stationarity, and ordinariness) is called the Poisson flow. It can be proved that the number of customers of the Poisson flow arriving over any interval of length t is distributed by the Poisson law with the parameter λt, because the probabilities $p_i(t), i \geq 0$ that exactly i customers will arrive over an interval of length t are

$$p_i(t) = \frac{(\lambda t)^i}{i!} e^{-\lambda t}, \quad i \geq 0.$$

The mean number of customers arriving over time t is λt. Therefore, λ is also the intensity, which is obvious in view of ordinariness of the Poisson flow.

On the strength of the properties of the Poisson flow, it is possible to prove that the times between the consecutive arrivals of customers are independent random variables distributed according to the same - exponential with the parameter λ - law as the time before the arrival of the first customer.

It can be noted that the number $\nu(t)$ of customers arriving over the interval $[0, t)$ constitutes the Poisson process.

Markov flow

Let $\nu(t)$ denote the number of customers arriving over the time interval $[0, t)$ and $\tau_1, \tau_2, ...$ be the instants of their arrivals. It can be assumed the existence of a Markov Process $\{\xi(t), t \geq 0\}$ defined on the finite state set $I = \{1, 2, ..., l\}$. Once it has been assumed $\eta(t) = (\xi(t), \nu(t))$, the process state set $\{\eta(t), t \geq 0\}$ is representable as $\cup_{k=0}^{\infty} I_k$, where $I_k = \{(i, k), i = \overline{1, l}\}, k \geq 0$. Therefore, the process $\{\eta(t), t \geq 0\}$ is in the state $(i, k), i = \overline{1, l}, k \geq 0$, if k customers arrived by 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

$$A = \begin{pmatrix} \Lambda & N & 0 & 0 & \cdots \\ 0 & \Lambda & N & 0 & \cdots \\ 0 & 0 & \Lambda & N & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix}$$

where Λ and N are square matrices of the order l. Obviously, for $i \neq j$ the elements Λ_{ij} 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_{ij} of the matrix N are the transition intensities accompanied by arrivals of customers. These elements are so defined:

$$\begin{aligned} N_{ij} &= \lambda_i p_{ij}, & 1 \leq j \leq l \\ \Lambda_{ij} &= \lambda_i q_{ij}, & 1 \leq j \leq l, i \neq j \\ \Lambda_{ii} &= -\lambda_i, & \end{aligned}$$

with

$$\sum_{\substack{j=1 \\ j \neq i}}^{l} q_{ij} + \sum_{j=1}^{l} p_{ij} = 1.$$

The components p_{ij} and q_{ij} are obtained by considering that the probability of an arrival during an infinitesimal interval of length dt, which defines the transition from the state i to the state j, is given by

$$p_{ij} = N_{ij} dt = N_{ij} \lambda_i^{-1},$$

being $dt = 1/\lambda_i$. Analogously, the probability of the transition from the state i to the state j without any arrival during the time dt, is given by

$$q_{ij} = \Lambda_{ij} dt = \Lambda_{ij} \lambda_i^{-1}.$$

The process of arrivals characterised by a Markov input flow is said "*Markovian Arrival Process*" and denoted by MAP in Kendall notation.

If N is a diagonal matrix, then the Markov flow is called the Markov modulated Poisson process (MMPP). Note that $\Lambda + N$ is the matrix of transition intensities of the Markov process $\{\xi(t),\ t \geq 0\}$.
Understandably, if $l = 1,\ \Lambda_{11} = -\lambda$ and $N_{11} = \lambda$, then the ordinary Poisson flow is got.

Another important case of the Markov flow is the interrupted Poisson process (IPP), which finds wide application for calculating the models of telephone networks by the so-called method of equivalent replacements, and is defined by the matrix N of the order $l = 2$ having only one nonzero and strictly positive diagonal element.

An additional generalisation of the Markovian Arrival Process is the "*Batch Markovian Arrival Process*" (BMAP) in which more than one arrival at a time are allowed. The directing process of the BMAP is obtained considering a 2-dimensional Markov Process $(N(t), J(t))$ on the state space $\{(i,j) : i \geq 0, 1 \leq j \leq m\}$ with an infinitesimal generator Q having the following structure:

$$Q = \begin{pmatrix} D_0 & D_1 & D_2 & D_3 & \cdots \\ 0 & D_0 & D_1 & D_2 & \cdots \\ 0 & 0 & D_0 & D_1 & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix},$$

where $D_k,\ k \geq 0$, are square matrices $m \times m$, D_0 has negative diagonal elements and nonnegative off-diagonal elements, D_k, $k \geq 1$, are non negative and D, defined by

$$D = \sum_{k=0}^{\infty} D_k,$$

is an irreducible infinitesimal generator.
If $N(t)$ represents the number of arrivals up to time t and $J(t)$ an auxiliary state, then the above Markov Process defines a batch

arrival process where transitions from a state (i, j) to state $(i+k, l)$, $k \geq 1, 1 \leq j, l \leq m$, correspond to batch arrivals of size k, and thus batch size can depend on i and j.

Finally, refering again to the matrix A of transition intensities of a Markov process, if the matrix N of the order l is representable as $N = \overrightarrow{\nu}\overrightarrow{\alpha}^T$, where $\overrightarrow{\nu}$ and $\overrightarrow{\alpha}$ are some column vectors of the order l and $\overrightarrow{\alpha}$ is a probabilistic vector, then the corresponding Markov flow is called the "*phase-type*" (PH) flow.

1.1.2 Service Process

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.

A queueing system can be a *single server system* (with a single server) or a *multi-server system* (with more than one server). In the first case, all the enqueued people or objects have to be processed by that server, while in the second case customers enqueued are served by one of the available servers according some parameters, which can be the order of arrival, a priority or their typology.
Moreover, it is possible to distinguish queueing systems also on the base of types of service: when the customer after service departs from the system and never returns, the system is a single-stage one; when the customer must be handled successively by more than one server (go through more than one stage of service), the system is multi-stage.

Server(s) can operate without any interruption until all customers have received their service or can alternate periods of activity and periods in which it does not work. Actually, in many real life situations, particularly where the service facility consists of mechanically operated device, the service gets interrupted due to the occurrence of occasional random failures of the service device. The situations where such kinds of failures are common may be encountered in computer and communication systems, maintenance in production systems etc. In manually operated service facilities the interruptions in service may be on account of strain of service, untimely call by the boss or for similar other reasons. Due to these interruptions in service the various parameters of queueing system are affected. As a result of a breakdown, service facility becomes inoperative and the units demanding service can be served only when it is restored to operative state. In order to model also these kind of situations, queueing systems with the so called "server vacation" have to be considered. For references, classification of vacation models and survey of state of art, see, e.g., [30, 65, 66].

Additionally, service can be characterised by the way to process the customers, e.g. individually or in groups. It has already been said that arrivals can occur in batch, but also the service can be provided not only to an individual customer, but to a group (bulk, batch) of customers. Such queueing systems with "bulk service" have applications in several real-life fields, notably in production and manufacturing, transportation, package delivery, tourism, and amusement parks like Disneyland. Their advantage is to avoid a waste of time and resources when the service process starts for few products or customers. For example:

- In production and manufacturing applications, it may not always be efficient to start the production as soon as the order arrive, but it could be better to wait until a certain number of orders are placed to start the service process;

- In package delivery or people transfer, analogously, it is more convenient to fill the trucks or buses to their capacity to balance the cost / efficiency to the delivery times of packages or travel time of people;

- In the case of amusement parks set up, people queueing up for going through various thrill rides have to be accommodated in groups of varying sizes with restrictions on the minimum and maximum numbers in each ride.

Bailey [7] was the first one to introduce bulk service queueing models, drawing particular attention to the fixed-batch service, what means that the server always serves a fixed number of customers in batch. He derived equilibrium distribution of the queue length by using the embedded-Markov chain technique, in which the discrete time Markov chain is obtained from a non-Markovian process by using regeneration points. The regeneration points may be service completion epoch, customer departure epoch, customer arrival epoch, etc. The stochastic matrix of the obtained Markov chain can be used for finding the performance measures of a queueing model.
It was Neuts [57] to introduce the study of bulk service queueing systems in case the dimension of the batches is not previously fixed. This service rule was called "general bulk service" (GBS) rule, and, according to it, the server can start to provide service only when at least a minimal number of units are present in the system, and they are less then a maximum fixed value. With reference to the literature useful for this work, the book [25] and works by S.R. Chakravarthy should be mentioned, see, e.g., [18, 19, 20, 22, 23, 24]. In these works, it is usually assumed that some threshold, K, is fixed and the service is provided to the group of size not less than K and not more than some integer C. If the number of customers in the buffer at a service completion epoch is less than K, the server waits until this number increases to K.

Among papers dealing with batch service discipline, also the recent results in the paper [8] can be mentioned. There, many examples of real world applications of queues with group service are given and the survey of related research is presented. According to [8], the group service queueing systems can be divided into the following categories:

1. Systems in which the buffer size is (a) finite or (b) infinite;

2. Systems in which the arrivals occur according to a (a) renewal or (b) correlated process;

3. Systems in which the arrivals occur (a) singly or (b) in batches;

4. Systems in which the services are (a) independent of the batch size or (b) dependent on the size of the batch being served;

5. Systems in which the service times are (a) exponential or (b) non-exponential.

It is stressed in [8] that very few papers deal with case 2(b) in combination with 4(b) and 5(b).
For this motivation, in the research activity presented in this work of thesis, a lot of attention has been paid on this combination.

Generally speaking, along with unpredictable random fluctuations, there can be systematic time variations in the customer service times, which can depend on arrival and service of other customers, and so forth. However, it is usually assumed that the times of service of all customers are identically distributed random variables that are independent of each other and all system processes. By analogy with the recurrent flow, service here is said to be recurrent. Then, the time of service of any customer is characterized only by one distribution function denoted by $B(x)$. If customers of more than one type arrive to the queueing system, then the distribution of the service time can be dependent on the type of customer.
The flows of customers in service can be classified in analogy with the above described input flows.

1.1.3 Some distributions of interarrival times and service times

In order to better describe the input flows and service times, some types of distributions have to be introduced.
Let $A(x)$ denote the distribution function of the interarrival time between two consecutive customers, denoted below by ξ, and $B(x)$ the distribution function of time of service.

Regular or determinate flow (D). Customers arrive after a constant time and, consequently

$$A(x) = \begin{cases} 0, & \text{if } x \leq a; \\ 1, & \text{if } x > a. \end{cases}$$

The Laplace-Stiltjes transform $\alpha(s)$ of the distribution $A(x)$ is as follows:

$$\alpha(s) = \int_0^\infty e^{-sx} dA(x) = e^{-sa}.$$

Poisson flow (M). For Poisson flow it results

$$A(x) = 1 - e^{-\lambda x},\ x > 0,$$

that is, the interarrival times are distributed exponentially with the parameter $\lambda, 0 < \lambda < \infty$. The Laplace-Stiltjes transform $\alpha(s)$ of the distribution function $A(x)$ is

$$\alpha(s) = \frac{\lambda}{\lambda + s}.$$

Hyperexponential flow (H). For this flow the distribution function is hyperexponential

$$A(x) = \sum_{i=1}^{l} \alpha_i \left(1 - e^{-\lambda_i x}\right), \ x > 0,$$

with $\alpha_i > 0$, $\sum_{i=1}^{l} \alpha_i = 1$ and $0 < \lambda_i < \infty$, $i = \overline{1, l}$, that is, $A(x)$ is a mix of the exponential distributions with parameters λ_i and weights α_i. Relying on the properties of the exponential distributions, the Laplace-Stiltjes transform $\alpha(s)$ is

$$\alpha(s) = \sum_{i=1}^{l} \frac{\alpha_i \lambda_i}{\lambda_i + s}.$$

Erlangian flow (E). In this case, $A(x)$ is the Erlang distribution function with the parameters l and λ which is usually defined by its density of distribution

$$a(x) = A'(x) = \frac{\lambda^l x^{l-1}}{(l-1)!} e^{-\lambda x}, \ x > 0, \ 0 < \lambda < \infty.$$

The Laplace-Stiltjes transform $\alpha(s)$ of the distribution function $A(x)$ is

$$\alpha(s) = \left(\frac{\lambda}{\lambda + s}\right)^l.$$

Hyper-Erlangian flow (HE). The distribution function $A(x)$ of this flow is hyper-Erlangian:

$$A(x) = \sum_{i=1}^{l} \alpha_i E_{l_i}(x),$$

where $\alpha_i > 0$, $\sum_{i=1}^{l} \alpha_i = 1$ and $E_{l_i}(x)$, $i = \overline{i,l}$, is the Erlang distribution function with the parameters l_i and λ_i. Similar to the hyperexponential distribution, the hyper-Erlangian one is a mix of the Erlang, rather than exponential, distributions with the weights α_i. From the properties of the Erlang distribution function, the Laplace-Stiltjes transform is

$$\alpha(s) = \sum_{i=1}^{l} \alpha_i \left(\frac{\lambda_i}{\lambda_i + s} \right)^{l_i}.$$

Phase-type flow (PH). The phase type distribution can describe both the recurrent arrival flow and the customer service times.
Exponentiality of the involved distributions is a common assumption when modelling the performance of communication systems through queueing models, especially with retrials [1, 4, 69]. The interest of the exponential law is justified by the mathematical tractability of the resulting Markovian models. However, it is known that most real distributions do not follow the exponential law [38]. It has been shown that the Markovian Arrival Process (MAP) and the phase type distribution (PH distribution) provide alternatives to generalize, respectively, the Poisson flow and the exponential distribution, and still keep the tractability for modelling purposes [2, 53]. As recent application of the use of MAP processes and PH distributions to performance evaluation of the IEEE 802.11 medium access protocol and the TCP behavior over cellular radio channels, the papers [39] and [56], respectively, can be seen.

For these reasons, in the queueing systems introduced and analysed in this work of thesis, it is assumed that both service and retrial times follow PH distributions, while arrivals are regulated by a Markovian Arrival Process. In particular, it can be observed that phase type distributions form a versatile family of probability distributions. The exponential, Erlang and hyperexponential distributions belong to this family. Since PH distributions can be

used to approximate general distributions and fit observed data, the resulting models are appropriate for practical purposes.
The PH distribution represents a type of distribution with probabilistic interpretation relying on the notion of fictitious phases. The idea of fictitious phases belongs to A.K. Erlang who used them to "Markovize" the Erlang distribution. Here two examples are shown.
Let $B(x)$ be an Erlang distribution function of the time of servicing an arriving customer with the parameters μ and m, whose expression is such that $b(x) = B'(x) = \frac{\mu^m x^{m-1}}{(m-1)!} e^{-\mu x}$, $x > 0$, $m = 1, 2, ...,$ $0 < \mu < \infty$. Since its Laplace-Stiltjes transform is $\beta(s) = \left(\frac{\mu}{\mu+s}\right)^m$, then it can be verified that this Laplace-Stiltjes transform can be treated as the Laplace-Stiltjes transform of the sum of m independent random variables each of which has the same type of distribution with the parameter μ. Consequently, the corresponding process of service can be decomposed into m components or, as it is the convention, phases which the customer successively goes through one after another. In doing so, the times of going through the phases will be mutually independent and exponentially distributed with the parameter μ.

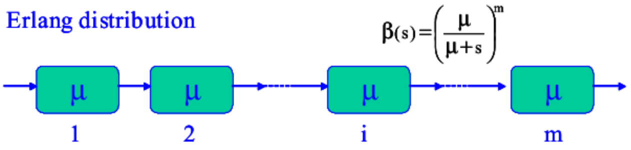
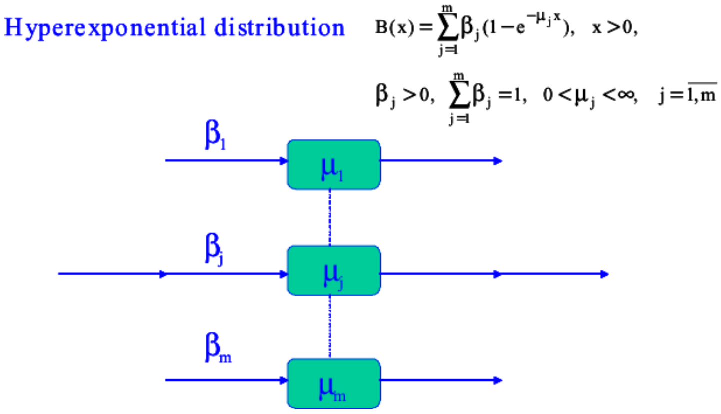


Figure 1.2 Representation of phases when the distribution of service times is Erlangian

The customer arriving to the server must successively go through all the m phases beginning from phase 1. Obviously, at any time instant at most one customer can be served, which means that there is no buffer between the phases. Additionally, the customer cannot be simultaneously in two or more phases of service.

In another case, service time can be considered as hyperexponentially distributed. The expression for the corresponding distribution function, $B(x) = \sum\limits_{j=1}^{m} \beta_j(1 - e^{-\mu_j x}),\ x > 0$, suggests that here one can also extract the phase of service: at the beginning of service, the customer is sent with the probability β_j to the jth phase where it is handled during an exponentially distributed random time with the parameter μ_j, and then the process of service is regarded as completed.



Hyperexponential distribution $B(x)=\sum\limits_{j=1}^{m}\beta_j(1-e^{-\mu_j x}),\ \ x>0,$

$\beta_j>0,\ \sum\limits_{j=1}^{m}\beta_j=1,\ \ 0<\mu_j<\infty,\ \ j=\overline{1,m}$

Figure 1.3 Representation of phases when the distribution of service times is hyperesponential

Each phase of service is arranged in parallel and tagged with the probabilities of customer arrival to the given phase. Since the server can handle at most one customer, this means that at most one phase can be occupied at each time instant.

Therefore, the hyperexponential distribution, as well as the Erlangian one, admits phase interpretation because reflects some process of service with fictitious phases, and in this sense both are phase type distributions.

At this point, one can ask if it is possible to invent a more general distribution function and its corresponding scheme of service with fictitious phases encompassing both successive and parallel service. The answer is positive. This general scheme of service with the fictitious phases is given by the PH distribution proposed by M.F. Neuts.

A brief description of the main notions for the PH distributions is the following: the distribution function $F(x)$ of a non-negative random variable is called phase type distribution or PH distribution if it is representable as

$$F(x) = 1 - \vec{f}^{T} e^{-Gx} \vec{1}, \quad x > 0,$$

where $\vec{f}$ is the m-dimensional vector for which $\sum_{j=1}^{m} f_j \leq 1,\ f_j \geq 0,\ j = \overline{1,m}$ and G is $m \times m$ matrix for which $\sum_{j=1}^{m} G_{ij} \leq 0;\ G_{ij} \geq 0,\ i \neq j;\ G_{ij} < 0,\ i,j = \overline{1,m}$, and at least for one i it results $\sum_{j=1}^{m} G_{ij} < 0$. The pair $\left(\vec{f}, G\right)$ 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, $\nu_i \geq -G_{ii},\ i = \overline{1,m}$, the numbers $\theta_{ij},\ i,j = \overline{1,m}$, 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 $\sum_{j=1}^{m} \theta_{ij} \leq 1,\ \theta_{ij} \geq 0,\ i,j = \overline{1,m}$.

Let now consider 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.

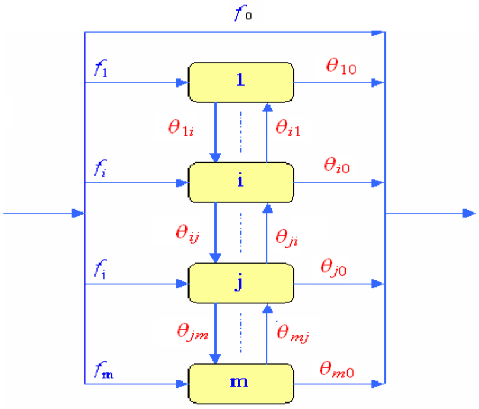


Figure 1.4 Representation of phases (nodes) in an open queueing network

The arriving customer is sent to the node $i, i = \overline{1, m}$, 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 exponentially distributed with the parameter ν_i. Upon leaving the node i, the customer travels to the node $j, j = \overline{1, m}$, with probability θ_{ij} and with the complementary probability $\theta_{i0} = 1 - \sum_{j=1}^{m} \theta_{ij}$ departs from the network.

1.1.4 System capacity

The place destined to waiting is called *buffer*, and it can have finite or infinite capacity. If there is no buffer in the queueing system, then the customer arriving when all servers are busy leaves the system and never returns, as is exemplified by the ordinary telephone systems, and the system is said to be a loss system.

1.1.5 Queueing discipline

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.

A special class of the queueing systems is made by the priority systems where the arriving customers have several priorities, the higher-priority customers having advantage over those with lower priorities, that is, being served before them. Priorities can be nonpreemptive if the higher-priority customers do not interrupt the customers in service and preemptive, otherwise. In the case of preemptive priorities, various modifications are possible: the underserved interrupted customers leave (are pushed out of) the system, are resumed as soon as the higher-priority customers depart from the system, served anew, and so forth.

1.1.6 Performance indexes

In the context of queueing systems, it is relevant to calculate some performance indexes, that is, userdefined characteristics of service which show to what extent the queueing system copes with the tasks incumbent on it. Some examples are:

- The queue length, that represents the number of customers waiting for service;
- The sojourn time, that is the time from the customer's arrival till his departure.
- The waiting time, that is the time from customer's arrival till the beginning of service.

Others performance indexes can be considered, on the basis of the specific system under analysis.

Chapter 2

A MAP/PH/1 Retrial System with Adaptive Pooling Admission Strategy

In this chapter it will be introduced and analysed a novel retrial queueing model where the requests (customers) are served in groups of finite size which are formed during a period of random length that is called admission period. The problem of choosing the optimal length of admission period and optimal size of the groups is faced here under assumptions that the input flow of customers is described by a Markovian Arrival Process, length of admission period and individual service time have phase type distribution.

The aim is to demonstrate that this discipline allows, in general, to essentially reduce the server wasted time (and, thus, to increase system throughput) as well as to provide better quality of service for customers. These can be achieved by means of providing the service to customers not individually, but in groups. Operation time of the server alternates between two states: service providing period and customers admission period. After a service completion moment, customers admission interval starts.

During this interval, primary customers and customers from the orbit are accepted to a pool of customers which will get service after this admission interval. Capacity of the pool is finite and after the moment when the pool becomes full all arriving customers move to the orbit. Admission period can be terminated either because its random duration expires or the pool becomes full. After completion or interruption (due to the pool full) of an admission interval, all customers in the pool are served by the server simultaneously during a random time having distribution depending on the number of customers in the pool.
Such an admission discipline is realistic in some wireless networks, e.g., in multi-rate IEEE802.11 WLAN where a group of requests from users can be processed simultaneously in parallel and processing of the whole group is considered finished if processing of all individual requests belonging to this group is completed. Therefore, the length of the service period of a group has distribution of the maximum of several independent random variables, each of which represents the service time of an individual customer belonging to this group. Since the expectation of the maximum of a fixed number of independent random variables is less (and can be much less) than the sum of expectations of these random variables, the average time devoted to the service of an arbitrary customer under the proposed service discipline may be much less than this time under the classical service discipline. Thus, throughput of the system under the proposed service discipline is higher and other performance measures of the system may be much better compared to the classical admission discipline.

A similar discipline has already been studied in a recent paper [31], but there it is assumed that the admission period can terminate only when its random duration expires, even if the pool is full.
Disadvantage of the cited admission discipline consists of impossibility to start service of a customer immediately after its arrival or making a retrial. The customer has to wait during some time in the pool even if the server is idle and there is no more space in the pool. It was shown in [31] that, under the proper choice of the admission period duration and the pool capacity, the main

performance measures of the system significantly benefit from this discipline.
However, it results very sensitive with respect to duration of admission period. If this duration is chosen in a not optimal way, performance measures of the system may be even worse than those one obtained under the classical admission discipline. Namely, if admission period is too long, it may often occur that the pool becomes full while the server does not start service and wastes time.
In this chapter, a more flexible and adaptive admission strategy is introduced, obtaining an essential improvement of the discipline analysed in [31]. Namely, an admission period can be terminated either because its random duration expires or the pool becomes full. It is intuitively clear that this "adaptive" admission discipline should provide better quality of service in the system because it allows to eliminate situations when the pool is full and customers are ready for service while the server still stays idle.
Sure, if admission period in real life system, which one would like to model, can not be interrupted because, e.g., the server should make some mandatory preparations for the next service period or indeed the server is not idle during the admission period but provides the service to some other flow of customers, only the strategy presented in [31] should be applied. However, if forced termination of an admission period is technically possible, the strategy here presented should be used. So in the numerical section, it will quantitatively be evaluated the profit earned from adaptivity of a new admission strategy under the optimal choice of the characteristic parameters.

It is worth to note that, although discipline of simultaneous service of a whole group of customers, instead of service one-by-one, was already considered in literature, the model under study is considerably different from other works. The main differences comparing to the papers cited in the Chapter 1 are as follows: (i) it is assumed in [18, 19, 20, 22, 23, 24] that distribution of service time of a group does not depend on size of the group while in this work such a dependence is assumed, what is very important from the point of view of modelling multi-rate IEEE802.11 WLAN; (ii) in [18, 19, 20, 22, 23], it is assumed that the system has a buffer, only in [24] retrial model is under study. In [24] service starts when the number of customers in orbit reaches the threshold L.

It is also worth to note that the overwhelming majority of known results for retrial queues are obtained for the systems with a stationary Poisson arrival process. However, such an arrival process is a poor descriptor of information flows in modern telecommunication networks. The Batch Markovian Arrival Process is recommended in literature for description of such flows. First papers where single server retrial queues with the $BMAP$ and arbitrary (more general than phase type) distribution of service time were analyzed are [32] and [34], where the $BMAP/G/1$ and the $BMAP/SM/1$ retrial queues, respectively, were under study. Multi-server retrial queues with the $BMAP$ and PH type distribution of service time were studied in [11, 12, 35, 45, 46, 47, 50, 51]. Actually here it is supposed that the arrival process is described by the Markovian Arrival Process.

Moreover, in some sense the model under study is close to the so called vacation queueing model in which the server interrupts the service, e.g., when the server becomes idle and resumes the service after a random amount of time or after accumulation of a definite number of customers in the system.

So, it results evident that, because the typical feature of many real world wireless communication systems is that the number of customers in orbit is not observable, the investigation of retrial queues with vacations and service resumes when the number of customers in orbit reaches some level has a little sense, while this invention of a pool as a special place for primary and retrial customers accommodation before entering the service definitely makes sense from point of view of potential applications.

2.1 The mathematical model

The system under study is a single server retrial queueing system. The input flow is described by a MAP [1]. Customer's arrival in the MAP is directed by an underlying irreducible continuous time Markov chain ν_t, $t \geq 0$, with the finite state space $\{0, ..., W\}$. Sojourn time of the Markov chain ν_t, $t \geq 0$, in the state ν has exponential distribution with parameter $\lambda_\nu, \nu = \overline{0, W}$. After this sojourn time expires, with probability $p_k(\nu, \nu')$, the process ν_t transits to the state ν', and k customers, $k = 0, 1$, arrive into the system. The intensities of jumps of underlying Markov chain from one state into another, which are accompanied by an arrival of k customers, are combined into the matrices D_k, $k = 0, 1$, of size $(W + 1) \times (W + 1)$. The matrix $D(1) = D_0 + D_1$ is the infinitesimal generator of the process ν_t, $t \geq 0$. The stationary distribution vector $\boldsymbol{\theta}$ of this process is the unique solution to the equations

$$\boldsymbol{\theta} D(1) = \mathbf{0}, \ \boldsymbol{\theta}\mathbf{e} = 1.$$

Here and in the sequel $\mathbf{0}$ is the zero row vector and $\mathbf{e}$ is the column vector of appropriate size consisting of ones. In case the dimensionality of the vector is not clear from context, it is indicated as a lower index, e.g. $\mathbf{e}_{\overline{W}}$ denotes the unit column vector of dimensionality $\overline{W} = W + 1$.

The average intensity λ (fundamental rate) of the MAP is defined as

$$\lambda = \boldsymbol{\theta} D_1 \mathbf{e}.$$

The variance v of intervals between customer arrivals is calculated as

$$v = 2\lambda^{-1}\boldsymbol{\theta}(-D_0)^{-1}\mathbf{e} - \lambda^{-2},$$

the squared coefficient c_{var} of variation is equal to

[1] For more information about the MAP, its special cases, properties and related research see [54] and the survey paper [21]. Usefulness of the MAP in modeling customers flows in telecommunication systems is mentioned in [41, 48]. Among the papers devoted to analysis of the queues with the MAP, [9, 10, 28, 29] can be mentioned.

$$c_{var} = 2\lambda\boldsymbol{\theta}(-D_0)^{-1}\mathbf{e} - 1,$$

while the correlation coefficient c_{cor} of intervals between successive arrivals is given by

$$c_{cor} = (\lambda^{-1}\boldsymbol{\theta}(-D_0)^{-1}D_1(-D_0)^{-1}\mathbf{e} - \lambda^{-2})/v.$$

If an arriving customer meets the server providing the service, it goes to the orbit and repeats attempts to get service in random time intervals whose duration has exponential distribution. Parameter of this distribution is α_i when $i,\ i \geqslant 1$, customers stay in the orbit, $\alpha_0 = 0$. Any dependence of the intensity α_i on i is admitted, such as α_i is monotonically increasing when i becomes large, and tends to infinity when i approaches infinity. Cases $\alpha_i = i\alpha$ and $\alpha_i = i\alpha + \gamma,\ \alpha > 0,\ \ \gamma \geqslant 0$, which satisfy the mentioned conditions, are popular in literature.

Server operates as follows. After a service completion instant, customer's admission interval starts. Duration of this interval is random. It has PH distribution[2] with irreducible representation $(\boldsymbol{\tau}, T)$. Duration of customer's admission interval is governed by the underlying process $\eta_t^{(a)},\ t \geqslant 0$, which is a continuous time Markov chain with state space $\{1, \dots, M^{(a)}\}$. The initial state of the process $\eta_t^{(a)},\ t \geqslant 0$, at the epoch of starting the admission interval is determined by the probabilistic row-vector $\boldsymbol{\tau} = (\tau_1, \dots, \tau_{M^{(a)}})$. The transitions of the process $\eta_t^{(a)},\ t \geqslant 0$, that do not lead to admission interval completion, are defined by the irreducible matrix T of size $M^{(a)} \times M^{(a)}$. The intensities of transitions, which lead to admission interval completion, are given by the column-vector $\mathbf{T}_0 = -T\mathbf{e}$. The distribution function of the admission interval time has the form $T(x) = 1 - \boldsymbol{\tau}e^{Tx}\mathbf{e}$. Its Laplace-Stieltjes transform $\int\limits_0^\infty e^{-sx}dT(x)$ is $\boldsymbol{\tau}(sI - T)^{-1}\mathbf{T}_0$. The

[2] A more detailed description of the PH distribution and its partial cases can be found e.g. in the book [58]. Possibility of the use of PH distribution for approximation (in sense of weak convergence) of an arbitrary distribution is mentioned, e.g., in [6].

average length of admission interval time is given by

$$r_1 = \boldsymbol{\tau}(-T)^{-1}\mathbf{e}.$$

Further $\mu = r_1^{-1}$ will represent the intensity of admission. The matrix $T + \mathbf{T}_0\boldsymbol{\tau}$ is assumed to be irreducible.

During the customer's admission interval, arriving primary customers and customers making attempts from the orbit are accepted to a pool of customers which will get service after the end of admission interval. Capacity of the pool is finite, equal to $N,\ N < \infty$, and the admission period is terminated either if the admission interval expires or the pool becomes full.

After termination of admission interval, all $n,\ n = \overline{0, N}$, customers in the pool are served simultaneously by the server during a time having PH distribution with irreducible representation $(\boldsymbol{\beta}^{(n)}, S^{(n)})$. Underlying process of this distribution is $\eta_t^{(s,n)},\ t \geqslant 0$, with a finite state space $\{1, \dots, M^{(s,n)}\}$. It can be denoted $\mathbf{S}_0^{(n)} = -S^{(n)}\mathbf{e}$, $n = \overline{0, N}$. The average service time of a group of n customers is defined by formula

$$b_1^{(n)} = \boldsymbol{\beta}^{(n)}(-S^{(n)})^{-1}\mathbf{e},\ n = \overline{0, N}.$$

For simplification of the notation, let it assume that the customer may provide the service even if the number of customers in the pool after completion of admission interval is equal to 0. This service may be interpreted, e.g. as maintenance or vacation or sleep period of the server. If one would like to exclude service provisioning to a group of size 0 from the model, he or she may set in the algorithms $M^{(s,0)} = 1,\ \boldsymbol{\beta}^{(s,0)} = 1,$ and $S^{(0)}$ be equal to a very large in modulus negative number. In this case, the server takes another customer's admission interval if the pool is empty at a given moment of customer's admission interval completion.

From the point of view of mathematical generality, in further derivations it will not strictly be specified the way of choosing the irreducible representations $(\boldsymbol{\beta}^{(n)}, S^{(n)}),\ n = \overline{0, N}$. Only the following two assumptions will be made:

(i) Let the service time of an individual customer have PH distribution with irreducible representation $(\boldsymbol{\beta}, S)$ and average value

$b_1 = \boldsymbol{\beta}(-S)^{-1}\mathbf{e}$. Size of the vector $\boldsymbol{\beta}$ is assumed to be equal to some integer $M,\ M \geqslant 1$.

(ii) The following inequalities

$$b_1^{(1)} \leqslant b_1^{(2)} \leqslant \ldots \leqslant b_1^{(N)} < N b_1^{(1)}$$

are fulfilled.

In numerical results section, a concrete form of the irreducible representations $(\boldsymbol{\beta}^{(n)}, S^{(n)}),\ \ n = \overline{1, N},$ will be chosen as representations of distribution of maximum of n independent random variables having PH distribution with irreducible representation $(\boldsymbol{\beta}, S)$.

The main goal is to compute the average number of customers in the system (in the orbit and in the pool or in the service) at arbitrary time moment and the probability that an arbitrary customer avoids visiting of orbit and, then, to find the optimal mean value of customer's admission interval under different values of the pool capacity N and different correlation in the arrival process. Towards this end, it is necessary to get stability condition of the system and compute the stationary distribution of the system states as well as to derive formulas for computation of the main performance measures of the system for any fixed set of its parameters.

2.2 The process of the system states

Let

- i_t be the number of customers in the orbit, $i_t \geqslant 0$,
- r_t be the current state of the server: $r_t = 0$ if admission period is in a progress and $r_t = 1$ if server provides the service,
- n_t be equal to the number of customers in the pool (if $r_t = 0$), $n_t = \overline{0, N-1}$, or in a group of customers receiving service (if $r_t = 1$), $n_t = \overline{0, N}$,

- ν_t be the state of the underlying process of the MAP, $\nu_t = \overline{0, W}$,

- η_t be the state of the underlying process of the PH process of customers admission (if $r_t = 0$) or service (if $r_t = 1$), $\eta_t = \overline{1, M^{(a)}}$, if $r_t = 0$, $\eta_t = \overline{1, M^{(s,n)}}$, if $r_t = 1$ and $n_t = n$, $n = \overline{0, N}$,

at the epoch t, $t \geq 0$.

It is easy to see that the five-dimensional process

$$\xi_t = \{i_t,\ r_t,\ n_t,\ \nu_t,\ \eta_t\},\ \ t \geq 0,$$

is an irreducible continuous time Markov chain with one component (i_t) having infinite state space and four finite components.

To analyse behavior and properties of this Markov chain, the infinitesimal generator of the chain has to be computed. Let this generator be denoted as $\mathbf{Q}$. The diagonal entries $\mathbf{Q}_{(i,r,n,\nu,\eta),(i,r,n,\nu,\eta)}$ are negative. Modulus of each diagonal entry defines intensity of departure of the Markov chain from the corresponding state of the Markov chain. The non-diagonal entry $\mathbf{Q}_{(i,r,n,\nu,\eta),(i',r',n',\nu',\eta')}$ is non-negative and defines intensity of transition of the Markov chain from the state (i, r, n, ν, η) to the state $(i', r', n', \nu', \eta')$.

To simplify the structure of generator $\mathbf{Q}$ and following traditional methodology of analysis of multi-dimensional Markov chains, let the states of the Markov chain ξ_t be enumerated in the lexicographic order. All the states of the chain having value (i, r, n) as the first three components will be composed to a *level* (i, r, n). The level $(i, 0, n)$, $n = \overline{0, N-1}$, contains $\overline{W}M^{(a)}$ states and the level $(i, 1, n)$, $n = \overline{0, N}$, contains $\overline{W}M^{(s,n)}$ states. Analogously, the levels (i, r, n) will be composed to *macro-level* (i, r), and then a *super-level* i will be formed as a composition of macro-levels (i, r), $r = 0, 1$, $i \geqslant 0$.

It follows that it is possible to represent the generator of the chain in the way described below, where this notation has been used: diag$\{\ \}$ denotes the diagonal matrix with diagonal entries listed in the brackets, $\mathrm{diag}^{+}\{\ \}$ indicates the matrix having the up-diagonal entries listed in the brackets and all other entries equal to 0, I_k is the identity matrix of order k, $E_N^+ = \mathrm{diag}^+\{1,\dots,1\}$ is the square matrix of size N, $\otimes$ is the symbol of Kronecker product of matrices and $\oplus$ is the symbol of Kronecker sum of matrices.

Lemma 1. *Generator* $\mathbf{Q}$ *has three block diagonal structure:*

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_{0,0} & \mathbf{Q}_{0,1} & O & O & \dots \\ \mathbf{Q}_{1,0} & \mathbf{Q}_{1,1} & \mathbf{Q}_{1,2} & O & \dots \\ O & \mathbf{Q}_{2,1} & \mathbf{Q}_{2,2} & \mathbf{Q}_{2,3} & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \\ & & & & \end{pmatrix}$$

where non-zero blocks $\mathbf{Q}_{i,j}$ *defining intensities of transitions from super-level* i *to super-level* j, $j = \max\{0, i-1\}, i, i+1$, *are defined as follows:*

- $$\mathbf{Q}_{i,i} = \begin{pmatrix} \mathbf{Q}_{(i,0),(i,0)} & \mathbf{Q}_{(i,0),(i,1)} \\ \mathbf{Q}_{(i,1),(i,0)} & \mathbf{Q}_{(i,1),(i,1)} \end{pmatrix}$$

  where the square matrix $\mathbf{Q}_{(i,0),(i,0)}$ *is defined by*

  $$\mathbf{Q}_{(i,0),(i,0)} = I_N \otimes (D_0 \oplus T) - \alpha_i I_N \otimes I_{\overline{W}M^{(a)}} + E_N^+ \otimes D_1 \otimes I_{M^{(a)}};$$

  the matrix $\mathbf{Q}_{(i,0),(i,1)}$ *is given by*

  $$\mathbf{Q}_{(i,0),(i,1)} = \begin{pmatrix} B_0 & O & \dots & O & O \\ O & B_1 & \dots & O & O \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & \dots & B_{(N-1)} & D_1 \otimes \mathbf{e}_{M^{(a)}} \otimes \boldsymbol{\beta}^{(N)} \end{pmatrix},$$

  where $B_i = I_{\overline{W}} \otimes \mathbf{T}_0 \otimes \boldsymbol{\beta}^{(i)}$*;*

  $\mathbf{Q}_{(i,1),(i,0)}$ *is the non-square matrix of the form*

  $$\mathbf{Q}_{(i,1),(i,0)} = \begin{pmatrix} I_{\overline{W}} \otimes \mathbf{S}_0^{(0)} \otimes \boldsymbol{\tau} & O & \dots & O \\ \vdots & \vdots & \ddots & \vdots \\ I_{\overline{W}} \otimes \mathbf{S}_0^{(N)} \otimes \boldsymbol{\tau} & O & \dots & O \\ & & & \end{pmatrix},$$

having $(N+1)$ block rows and N block columns;

$$\mathbf{Q}_{(i,1),(i,1)} = \mathrm{diag}\{D_0 \oplus S^{(0)}, \dots, D_0 \oplus S^{(N)}\};$$

- $$\mathbf{Q}_{i,i-1} = \begin{pmatrix} \mathbf{Q}_{(i,0),(i-1,0)} & \mathbf{Q}_{(i,0),(i-1,1)} \\ O & O \\ & \end{pmatrix},$$

  where

  $$\mathbf{Q}_{(i,0),(i-1,0)} = \alpha_i \mathrm{diag}^{+}\{1, \dots, 1, 0\} \otimes I_{\overline{W}M^{(a)}},$$

  $$\mathbf{Q}_{(i,0),(i-1,1)} = \begin{pmatrix} O & \dots & O & O \\ \vdots & \ddots & \vdots & \vdots \\ O & \dots & O & \alpha_i I_{\overline{W}} \otimes \mathbf{e}_{M^{(a)}} \otimes \boldsymbol{\beta}^{(N)} \end{pmatrix};$$

- $$\mathbf{Q}_{i,i+1} = \mathrm{diag}\{O, \mathbf{Q}_{(i,1),(i+1,1)}\},$$

  where

  $$\mathbf{Q}_{(i,1),(i+1,1)} = \mathrm{diag}\{D_1 \otimes I_{M^{(s,0)}}, \dots, D_1 \otimes I_{M^{(s,N)}}\}.$$

Proof of the lemma consists in analysing the Markov chain $\xi_t,\ t \geq 0$, transitions during the infinitesimal interval of time and further combining corresponding transition intensities into the matrix blocks.

It follows from the previous lemma that the Markov chain ξ_t is level-dependent Quasi-Birth-and-Death process.
Quasi-Birth-and-Death processes are the Markov chains, extensively studied by M.F. Neuts, having two properties named as "skip-free to the left" and "skipfree to the right". Chains having the "skip-free to the left" property are studied as M/G/1 type Markov chains (called multi-dimensional quasi-Toeplitz Markov chains in [33]), while chains having the "skip-free to the right" property are referred to as G/M/1 type Markov chains.

More specifically, this Markov chain belongs to the class of Asymptotically Quasi-Toeplitz Markov Chains (AQTMCs) introduced and analysed in paper [49].
Here its definition is reported:

Definition 2. *Given $A = (A_{i,l})_{i,l \geq 0}$ the generator of the chain in block form, where $A_{i,l}$ is the matrix formed by the intensities $a_{(i,\mathbf{r})(l,\nu)}$ of transition from the state $(i,\mathbf{r})$ to (l,ν), and said $T_i, i \geq 0$ the diagonal matrix with $-a_{(i,\mathbf{r})(i,\mathbf{r})}$ as its diagonal entries, an irreducible continuous Markov chain ξ_t, $t \geq 0$, is called Asymptotically Quasi-Toeplitz Markov chain if*
(i) $A_{i,l} = 0$ for $l < i-1$, $i > 0$,
(ii) There exist matrices Y_k, $k \geq 0$, such that

$$\begin{aligned} Y_k &= \lim_{i\to\infty} T_i^{-1} A_{i,i+k-1}, \quad k = 0, 2, 3, ..., \\ Y_1 &= \lim_{i\to\infty} T_i^{-1} A_{i,i} + I, \end{aligned}$$

(iii) The matrix $\sum_{k=0}^{\infty} Y_k$ is stochastic.

Taking into account this definition, it can be proved that the considered Markov chain is Asymptotically Quasi-Toeplitz.
Let $\mathcal{K}^{(a)}$ be the diagonal matrix with the diagonal entries given by the moduli of the diagonal entries of the matrix $D_0 \oplus T$ and $\mathcal{K}_n^{(s)}$ be the diagonal matrix with the diagonal entries equal to the moduli of the diagonal entries of the matrix $D_0 \oplus S^{(n)}$, $n = \overline{0, N}$.

Let $\mathbf{R}_i$ be the diagonal matrix with the diagonal entries given by the moduli of the diagonal entries of the matrix $\mathbf{Q}_{i,i}$. It can be verified that the matrix $\mathbf{R}_i$ is defined by formula

$$\mathbf{R}_i = \text{diag}\{I_N \otimes \mathcal{K}^{(a)} + \alpha_i I_N \otimes I_{\overline{W}M^{(a)}}, \text{diag}\{\mathcal{K}_0^{(s)}, \dots, \mathcal{K}_N^{(s)}\}\}.$$

The following lemma can be stated.

Lemma 3. *The following limits exist*

$$\mathbf{Y}_0 = \lim_{i\to\infty} \mathbf{R}_i^{-1} \mathbf{Q}_{i,i-1}, \ \mathbf{Y}_1 = \lim_{i\to\infty} \mathbf{R}_i^{-1} \mathbf{Q}_{i,i} + I, \ \mathbf{Y}_2 = \lim_{i\to\infty} \mathbf{R}_i^{-1} \mathbf{Q}_{i,i+1},$$

and are defined by formulas

$$\mathbf{Y}_0 = \begin{pmatrix} E_N^{+} \otimes I_{\overline{W}M^{(a)}} & \mathbf{Y}_0^{(0,1)} \\ O & O \end{pmatrix}$$

where $\mathbf{Y}_0^{(0,1)}$ *is the matrix of size* $N \times (N+1)$ *that is obtained by supplementing the square zero matrix from the right with the block column having entries* $\left(O, ..., O, I_{\overline{W}} \otimes \mathbf{e}_{M^{(a)}} \otimes \boldsymbol{\beta}^{(N)}\right)$*:*

$$\mathbf{Y}_0^{(0,1)} = \begin{pmatrix} O & & & O \\ & \ddots & & \vdots \\ & & \ddots & O \\ & & O & I_{\overline{W}} \otimes \mathbf{e}_{M^{(a)}} \otimes \boldsymbol{\beta}^{(N)} \end{pmatrix};$$

$$\mathbf{Y}_1 = \begin{pmatrix} O & O \\ \mathbf{Y}_1^{(1,0)} & \mathbf{Y}_1^{(1,1)} \end{pmatrix},$$

where $\mathbf{Y}_1^{(1,0)}$ *is the non-square matrix of size* $(N+1) \times N$ *:*

$$\mathbf{Y}_1^{(1,0)} = \begin{pmatrix} (\mathcal{K}_0^{(s)})^{-1}(I_{\overline{W}} \otimes \mathbf{S}_0^{(0)} \otimes \boldsymbol{\tau}) & O & \dots & O \\ \vdots & \vdots & \ddots & \vdots \\ (\mathcal{K}_N^{(s)})^{-1}(I_{\overline{W}} \otimes \mathbf{S}_0^{(N)} \otimes \boldsymbol{\tau}) & O & \dots & O \end{pmatrix},$$

$$\mathbf{Y}_1^{(1,1)} = \mathrm{diag}\{I_{\overline{W}M^{(s,n)}} + (\mathcal{K}_n^{(s)})^{-1}(D \oplus S^{(n)}),\ n = \overline{0,N}\},$$

$$D = D_0 + D_1;$$

$$\mathbf{Y}_2 = \mathrm{diag}\{O, \mathbf{Y}_2^{(1,1)}\},$$

$$\mathbf{Y}_2^{(1,1)} = \mathrm{diag}\{(\mathcal{K}_n^{(s)})^{-1}(D_1 \otimes I_{M^{(s,n)}}),\ n = \overline{0,N}\}.$$

This lemma can be easily proved by means of direct calculation of the limits. One can summarize that

(i) the limits $\mathbf{Y}_0$, $\mathbf{Y}_1$, and $\mathbf{Y}_2$ exist,

(ii) the matrix $\mathbf{Y} = \mathbf{Y}_0 + \mathbf{Y}_1 + \mathbf{Y}_2$ is stochastic,

(iii) all blocks of generator $\mathbf{Q}$ below the sub-diagonal are zero matrices.

All conditions of definition of $AQTMC$ given in [49] are fulfilled and, thus, the Markov chain ξ_t under study belongs to the class of Asymptotically Quasi-Toeplitz Markov Chains. It gives the opportunity to derive ergodicity and non-ergodicity conditions for this Markov chain and compute its steady-state distribution.

2.3 Ergodicity Condition

Theorem 4. *The Markov chain ξ_t is ergodic if the inequality*

$$\lambda b_1^{(N)} < N \tag{2.3.1}$$

is fulfilled and is non-ergodic if

$$\lambda b_1^{(N)} > N.$$

Here $b_1^{(N)} = \boldsymbol{\beta}^{(N)}(-S^{(N)})^{-1}\mathbf{e}$ is the average duration of service of a group consisting of N customers and λ is the fundamental rate of the MAP.

Proof. It can be easily verified that the matrix $\mathbf{Y}$ introduced above is irreducible. Thus, it follows from [49] that the Markov chain ξ_t is ergodic if the inequality

$$\mathbf{y}\mathbf{Y}_0\mathbf{e} > \mathbf{y}\mathbf{Y}_2\mathbf{e}$$

holds true, where the vector $\mathbf{y}$ is the unique solution of the system of linear algebraic equations

$$\mathbf{y}\mathbf{Y} = \mathbf{y}, \ \ \mathbf{y}\mathbf{e} = 1. \tag{2.3.2}$$

The Markov chain ξ_t is non-ergodic if the inequality

$$\mathbf{y}\mathbf{Y}_0\mathbf{e} < \mathbf{y}\mathbf{Y}_2\mathbf{e}$$

holds true.

Let the solution to the system (2.3.2) be found. This should allow to get ergodicity condition in a nice analytic form. Obviously, the vector $\mathbf{y}$ has the following structure:

$$\mathbf{y} = (\mathbf{y}_0^{(0)}, \mathbf{y}_0^{(1)}, \dots, \mathbf{y}_0^{(N-1)}, \mathbf{y}_1^{(0)}, \mathbf{y}_1^{(1)}, \dots, \mathbf{y}_1^{(N)})$$

where the sub-vector $\mathbf{y}_r^{(n)}$ corresponds to the state r, $r = 0, 1$, of component r_t of the Markov chain ξ_t and the state n of component n_t, where $n = \overline{0, N-1}$ if $r_t = 0$ and $n = \overline{0, N}$ if $r_t = 1$.

By substituting this form of the vector $\mathbf{y}$ to equation $\mathbf{yY} = \mathbf{y}$ of system (2.3.2), the following system of linear algebraic equations for the components $\mathbf{y}_r^{(n)}$ is got:

$$\mathbf{y}_0^{(0)} = \sum_{n=0}^{N} \mathbf{y}_1^{(n)} (\mathcal{K}_n^{(s)})^{-1} (I_{\overline{W}} \otimes \mathbf{S}_0^{(n)} \otimes \boldsymbol{\tau}) \tag{2.3.3}$$

$$\mathbf{y}_0^{(n)} = \mathbf{y}_0^{(n-1)}, \ n = \overline{1, N-1}, \tag{2.3.4}$$

$$\mathbf{y}_1^{(n)} = \mathbf{y}_1^{(n)} (I + (\mathcal{K}_n^{(s)})^{-1} (D \oplus S^{(n)})), \ n = \overline{0, N-1}, \tag{2.3.5}$$

$$\mathbf{y}_1^{(N)} = \mathbf{y}_0^{(N-1)} (I_{\overline{W}} \otimes \mathbf{e}_{M^{(a)}} \otimes \boldsymbol{\beta}^{(N)}) + \mathbf{y}_1^{(N)} (I + (\mathcal{K}_N^{(s)})^{-1} (D \oplus S^{(N)})). \tag{2.3.6}$$

Let this system be solved.
Equations (2.3.5) can be evidently rewritten as

$$\mathbf{y}_1^{(n)} (\mathcal{K}_n^{(s)})^{-1} (D \oplus S^{(n)}) = \mathbf{0}, \ n = \overline{0, N-1}.$$

Because the matrix $D \oplus S^{(n)}$ is an irreducible sub-generator with strong domination of a diagonal entry at least in one row, by theorem of O. Tausski [3] this matrix is non-singular. Therefore, these equations have only trivial solution, i.e.,

$$\mathbf{y}_1^{(n)} = \mathbf{0}, \ n = \overline{0, N-1},$$

[3] *Diagonal Dominance Theorem.* Let A be a complex $n \times n$ matrix and let A_i be the sum of the absolute values of the non-diagonal elements in the i-th row. If A is diagonally dominant ($|a_{ii}| > A_i, i = 1, ..., n$) then $\det A \neq 0$.

and from (2.3.3) and (2.3.4) it follows that

$$\mathbf{y}_0^{(n)} = \mathbf{y}_1^{(N)}(\mathcal{K}_N^{(s)})^{-1}(I_{\overline{W}} \otimes \mathbf{S}_0^{(N)} \otimes \boldsymbol{\tau}),\ n = \overline{0, N-1}. \tag{2.3.7}$$

Thus, at this point there is only one unknown vector: $\mathbf{y}_1^{(N)}$. From (2.3.6), the relation

$$\mathbf{y}_0^{(N-1)}(I_{\overline{W}} \otimes \mathbf{e}_{M^{(a)}} \otimes \boldsymbol{\beta}^{(N)}) + \mathbf{y}_1^{(N)}(\mathcal{K}_N^{(s)})^{-1}(D \oplus S^{(N)}) = \mathbf{0}, \tag{2.3.8}$$

can be obtained. This equation, due to (2.3.7), is equivalent to

$$\mathbf{y}_1^{(N)}(\mathcal{K}_N^{(s)})^{-1}\left[I_{\overline{W}} \otimes \mathbf{S}_0^{(N)}\boldsymbol{\beta}^{(N)} + (D \oplus S^{(N)})\right] = \mathbf{0}, \tag{2.3.9}$$

being

$$\begin{aligned}(I_{\overline{W}} \otimes \mathbf{S}_0^{(N)} \otimes \boldsymbol{\tau})(I_{\overline{W}} \otimes \mathbf{e}_{M^{(a)}} \otimes \boldsymbol{\beta}^{(N)}) &= \\ = I_{\overline{W}} \otimes \left[\left(\mathbf{S}_0^{(N)} \otimes \boldsymbol{\tau}\right)\left(\mathbf{e}_{M^{(a)}} \otimes \boldsymbol{\beta}^{(N)}\right)\right] &= \\ = I_{\overline{W}} \otimes \mathbf{S}_0^{(N)} \otimes \boldsymbol{\beta}^{(N)} = I_{\overline{W}} \otimes \mathbf{S}_0^{(N)}\boldsymbol{\beta}^{(N)}.\end{aligned}$$

From normalisation condition ($\mathbf{ye} = 1$) this relation is got:

$$\left[N\mathbf{y}_1^{(N)}(\mathcal{K}_N^{(s)})^{-1}(I_{\overline{W}} \otimes \mathbf{S}_0^{(N)} \otimes \boldsymbol{\tau}) + \mathbf{y}_1^{(N)}\right] = 1.$$

By introducing notation $\mathbf{z}_1 = \mathbf{y}_1^{(N)}(\mathcal{K}_N^{(s)})^{-1}$, system (2.3.9) can be rewritten in the form

$$\mathbf{z}_1\left[I_{\overline{W}} \otimes \mathbf{S}_0^{(N)}\boldsymbol{\beta}^{(N)} + (D \oplus S^{(N)})\right] = \mathbf{0}. \tag{2.3.10}$$

By direct substitution, it can be verified that solution to the system (2.3.10) has the following simple form:

$$\mathbf{z}_1 = \boldsymbol{\theta} \otimes (\boldsymbol{\beta}^{(N)}(-S^{(N)})^{-1}) \tag{2.3.11}$$

where $\boldsymbol{\theta}$ is the vector of stationary distribution of the underlying process of arrivals.
In turn, the inequality $\mathbf{yY}_0\mathbf{e} > \mathbf{yY}_2\mathbf{e}$ can be rewritten as

$$\mathbf{z}_1(D_1 \otimes I_{M^{(s,N)}})\mathbf{e} < (N-1)\,\mathbf{z}_1(I_{\overline{W}} \otimes \mathbf{S}_0^{(N)} \otimes \boldsymbol{\tau})\mathbf{e} - \mathbf{z}_1(D \oplus S^{(N)})\mathbf{e}. \tag{2.3.12}$$

This relation is derived from the products

$$yY_2e = \mathbf{y}_1^{(N)} \left(\mathcal{K}_N^{(s)}\right) (D_1 \otimes I_{M^{(s,N)}})\, e$$

and

$$\begin{aligned} yY_0e &= (N-1)\,\mathbf{y}_0^{(0)}e + \mathbf{y}_0^{(0)}(I_{\overline{W}} \otimes \mathbf{e}_{M^{(a)}} \otimes \boldsymbol{\beta}^{(N)}) = \\ &= (N-1)\,\mathbf{y}_1^{(N)} \left(\mathcal{K}_N^{(s)}\right)^{-1} (I_{\overline{W}} \otimes \mathbf{S}_0^{(N)} \otimes \boldsymbol{\tau})e + \\ &\quad + \mathbf{y}_1^{(N)} \left(\mathcal{K}_N^{(s)}\right)^{-1} \left(I_{\overline{W}} \otimes \mathbf{S}_0^{(N)}\boldsymbol{\beta}^{(N)}\right) = \\ &= (N-1)\,\mathbf{y}_1^{(N)} \left(\mathcal{K}_N^{(s)}\right)^{-1} (I_{\overline{W}} \otimes \mathbf{S}_0^{(N)} \otimes \boldsymbol{\tau})e + \\ &\quad + \mathbf{y}_1^{(N)} \left(\mathcal{K}_N^{(s)}\right)^{-1} \left(D \otimes I + I \otimes S^{(N)}\right) e. \end{aligned}$$

Substituting expression (2.3.11) to system (2.3.12), the inequality (2.3.1) is got. Condition of non-ergodicity is easily proven by analogy. Theorem is proven.

Remark 5. *Inequality (2.3.1) is intuitively clear. Ergodicity (stability) condition for any queueing system is defined by its ability to reduce the customers number in the system in situation when this number is huge. Evidently, for the system under study in such a situation customers number in the group receiving the service will be exactly N and average service time will be equal to $b_1^{(N)}$. The average number of customers arriving during the service time will be equal to $\lambda b_1^{(N)}$ while the number of customers departing from the system at service completion moment is given by N. Thus, an intuitively clear condition of the system ergodicity should be of form $\lambda b_1^{(N)} < N$ what coincides with strictly proven condition (2.3.1). The throughput of the system (the maximal intensity of customers flow that can be successfully processed by the system), which is one of the main performance measures of the system, is equal to $\frac{N}{b_1^{(N)}}$.*

Remark 6. *For the strategy considered in [31], ergodicity condition of the system has the form*

$$\lambda\left(r_1 + b_1^{(N)}\right) < N$$

where r_1 is the mean length of admission period. The throughput of the system is equal to $\frac{N}{r_1+b_1^{(N)}}$. Therefore, the adaptive strategy proposed in this work provides higher value of the throughput than the strategy considered in [31]. This is intuitively obvious because the adaptive strategy excludes situations when the pool is full while the service does not start.

It is worth to remind that, in turn, the throughput of the system with the strategy used in [31] can be much higher than the throughput of the system with the classical admission discipline suggesting the service of customers one by one.

2.4 Key performance indices of the system

Further let the inequality (2.3.1) be assumed fulfilled. Then the stationary distribution of the Markov chain ξ_t exists.

Let the stationary state probabilities of the chain be denoted as

$$\boldsymbol{\pi}(i,r,n,\nu,\eta) = \lim_{t\to\infty} P\{i_t = i,\ r_t = r,\ n_t = n,\ \nu_t = \nu,\ \eta_t = \eta\},$$

$i \geq 0,\ r = 0,1,\ \nu = \overline{0,W},\ n = \overline{0,N-1},\ \eta = \overline{1,M^{(a)}}$, if $r = 0$ and $\eta = \overline{1,M^{(s,n)}}, n = \overline{0,N}$, if $r = 1$.

Let $\boldsymbol{\pi}(i,r,n)$, $\boldsymbol{\pi}(i,r)$, $\boldsymbol{\pi}_i$ be the row vectors of probabilities of the states belonging, respectively, to the level (i,r,n), the macro-level (i,r) and the super-level $i,\ i \geq 0$.

Computation of the vectors $\boldsymbol{\pi}_i$, $i \geqslant 0$, for $AQTMC$ ξ_t can be performed based on numerically stable algorithm presented in [49]. Because the generator of the Markov chain ξ_t has three-diagonal block structure, while the algorithm in [49] assumes more general structure, the algorithm from [49] can be rewritten as follows.

Step 1. Compute the matrix G as the minimal non-negative solution to the matrix equation

$$G = Y_0 + Y_1 G + Y_2 G^2.$$

This equation is M. Neuts' equation, see, e.g., [59], for the discrete-time Markov chain having $Y(z) = Y_0 + Y_1 z + Y_2 z^2$, $|z| < 1$, as the generating function of its transition probability matrices. There is a lot of various algorithms for solving such a type of equations.

Step 2. Calculate the matrices $G_{i_0-1}, G_{i_0-2}, \dots, G_0$ using the equation of the backward recursion

$$G_i = (-\mathbf{Q}_{i+1,i+1} - \mathbf{Q}_{i+1,i+2} G_{i+1})^{-1} \mathbf{Q}_{i+1,i},$$

$i = i_0 - 1, i_0 - 2, \dots, 0$, with boundary condition $G_i = G$, $i \geq i_0$, where i_0 is an integer defined in such a way that, for a preassigned small positive number ϵ (the accuracy of the calculations), the inequality $\|G_{i_0} - G\| < \epsilon$ holds.

Step 3. Compute the matrices F_i by formulas

$$F_0 = I,\ F_i = \prod_{l=1}^{i} \mathbf{Q}_{l-1,l}[-(\mathbf{Q}_{l,l} + \mathbf{Q}_{l,l+1} G_l)]^{-1},\ i \geq 1.$$

Step 4. Compute the vector $\boldsymbol{\pi}_0$ as the unique solution to the system

$$\boldsymbol{\pi}_0(\mathbf{Q}_{0,0} + \mathbf{Q}_{0,1} G_0) = \mathbf{0},\ \boldsymbol{\pi}_0 \sum_{l=0}^{\infty} F_l \mathbf{e} = 1.$$

Step 5. Calculate the vectors $\boldsymbol{\pi}_i$ using

$$\boldsymbol{\pi}_i = \boldsymbol{\pi}_0 F_i,\ i \geq 0.$$

The described algorithm is numerically stable because it works with only non-negative matrices.

As soon as the vectors $\boldsymbol{\pi}_i,\ i \geq 0$, have been computed, various performance measures of the system can be calculated. Note that formulas for the main performance measures contain infinite sums. However, this does not create any essential difficulty in computer implementation. It is well known that if the ergodicity condition holds true, the stationary probability vectors $\boldsymbol{\pi}_i$ converge in norm to zero vector when i approaches infinity. So, computation of an infinite sum may be stopped if the norm of the term of sum becomes less than some preassigned value ϵ (e.g., $\epsilon = 10^{-15}$).

Below the main performance indices are reported:

- Average number of customers in the orbit

$$L_o = \sum_{i=1}^{\infty} i\boldsymbol{\pi}_i\mathbf{e}.$$

- Fraction of time when server has admission period

$$F^{(a)} = \sum_{i=0}^{\infty} \boldsymbol{\pi}(i,0)\mathbf{e}.$$

- Fraction of time when server has service period

$$F^{(s)} = \sum_{i=0}^{\infty} \boldsymbol{\pi}(i,1)\mathbf{e}.$$

- Average number of customers in the pool at an arbitrary moment conditional that admission period is in a progress

$$N_p = (F^{(a)})^{-1}\sum_{i=0}^{\infty}\sum_{n=1}^{N-1} n\boldsymbol{\pi}(i,0,n)\mathbf{e}.$$

- Average number of customers in the service at an arbitrary moment conditional that service period is in a progress

$$N_s = (F^{(s)})^{-1}\sum_{i=0}^{\infty}\sum_{n=1}^{N} n\boldsymbol{\pi}(i,1,n)\mathbf{e}.$$

- Average number of customers in the system at an arbitrary moment

$$L = L_o + \sum_{i=0}^{\infty}\left(\sum_{n=1}^{N-1} n\boldsymbol{\pi}(i,0,n)\mathbf{e} + \sum_{n=1}^{N} n\boldsymbol{\pi}(i,1,n)\mathbf{e}\right).$$

- Probability that an arbitrary customer will visit the orbit

$$P_{orbit} = \lambda^{-1}\left[\sum_{i=0}^{\infty}\boldsymbol{\pi}(i,1)\text{diag}\{(D_1 \otimes I_{M^{(s,n)}}), n = \overline{0,N}\}\mathbf{e}\right].$$

- Probability that an arbitrary customer will not visit the orbit

$$P_{pool} = 1 - P_{orbit} = \lambda^{-1}\sum_{i=0}^{\infty}\sum_{n=0}^{N-1}\boldsymbol{\pi}(i,0,n)(D_1 \otimes I_{M^{(a)}})\mathbf{e}.$$

- Probability that an arbitrary customer enters the service immediately upon arrival

$$P_{imm} = \lambda^{-1}\sum_{i=0}^{\infty}\boldsymbol{\pi}(i,0,N-1)(D_1 \otimes I_{M^{(a)}})\mathbf{e}.$$

- Probability that at an arbitrary time the server provides the service to the empty pool (e.g., it has maintenance or vacation) conditional that the service period is in a progress

$$P_{empty} = (F^{(s)})^{-1}\sum_{i=0}^{\infty}\boldsymbol{\pi}(i,1,0)\mathbf{e}.$$

- Probability that at an arbitrary time the server provides the service to the full pool

$$P_{full} = (F^{(s)})^{-1}\sum_{i=0}^{\infty}\boldsymbol{\pi}(i,1,N)\mathbf{e}.$$

- Conditional Laplace-Stieltjes transform of the time spent in the pool by an arbitrary customer conditional it enters the pool upon arrival

$$v_{pool}(s) = \lambda(P_{pool} - P_{imm})^{-1} \sum_{i=0}^{\infty} \sum_{n=0}^{N-2} \boldsymbol{\pi}(i,0,n)$$
$$*(D_1\mathbf{e} \otimes I_{M^{(a)}})(sI - T)^{-1}\mathbf{T}_0,\ Re\ s > 0.$$

- Average time spent in the pool by an arbitrary customer conditional it enters the pool upon arrival

$$V_{pool} = \lambda(P_{pool}-P_{imm})^{-1} \sum_{i=0}^{\infty} \sum_{n=0}^{N-2} \boldsymbol{\pi}(i,0,n)(D_1\mathbf{e}\otimes I_{M^{(a)}})(-T)^{-1}\mathbf{e}.$$

- Average time spent by an arbitrary customer in service conditional it does not visit the orbit upon arrival

$$V_{service} =$$
$$(\lambda P_{pool})^{-1}\Bigg[b_1^{(N)} \sum_{i=0}^{\infty} \boldsymbol{\pi}(i,0,N-1)(D_1 \otimes I_{M^{(a)}})\mathbf{e}$$
$$+\sum_{i=0}^{\infty}\sum_{n=0}^{N-2} \boldsymbol{\pi}(i,0,n)(D_1 \otimes I_{M^{(a)}})\Bigg[\sum_{k=0}^{N-2-n} b_1^{(n+k+1)}$$
$$\int_0^{\infty} \mathbf{P}(k,t)\mathbf{e} \otimes e^{Tt}\mathbf{T}_0 dt + b_1^{(N)} \sum_{k=N-1-n}^{\infty} \int_0^{\infty} \mathbf{P}(k,t)\mathbf{e} \otimes e^{Tt}\mathbf{T}_0 dt\Bigg]\Bigg]$$

where the matrices $\mathbf{P}(k,t)$, $k \geq 0$, are defined as coefficients in matrix expansion

$$\sum_{k=0}^{\infty} \mathbf{P}(k,t)z^k = e^{(D_0+D_1 z)t},\ |z| < 1.$$

MAP/PH/1 retrial system with a classical admission strategy

In order to make comparisons between the presented admission strategy and the classical one, here the results for the $MAP/PH/1$ retrial system with a classical admission strategy are briefly presented.

The behavior of the $MAP/PH/1$ retrial system with a classical admission strategy is described by the four-dimensional Markov chain $\hat{\xi}_t = \{i_t, r_t, \nu_t, \eta_t\},\ t \geqslant 0$, where i_t is the number of customers in the orbit, $i_t \geqslant 0,\ r_t$ is the current state of the server: $r_t = 0$ if the server is idle and $r_t = 1$ if server provides the service, ν_t is the state of the underlying process of the MAP, $\nu_t = \overline{0, W}$, η_t is the state of the underlying process of the PH service process.

It can be shown that the non-zero blocks $\hat{\mathbf{Q}}_{i,j}$ of the generator $\hat{\mathbf{Q}}$ of this Markov chain are defined as follows:

$$\hat{\mathbf{Q}}_{i,i} = \begin{pmatrix} -\alpha_i I + (D_0 \otimes I) & D_1 \otimes I \\ I \otimes (\mathbf{S}_0 \boldsymbol{\beta}) & D_0 \oplus S \end{pmatrix},$$

$$\hat{\mathbf{Q}}_{i,i+1} = \begin{pmatrix} O & O \\ O & D_1 \otimes I \end{pmatrix}, \quad \hat{\mathbf{Q}}_{i,i-1} = \begin{pmatrix} O & \alpha_i I \\ O & O \end{pmatrix}$$

and the matrices $Y_k,\ k = 0, 1, 2$, are given by formulas

$$Y_1 = \begin{pmatrix} O & O \\ \mathcal{R}^{-1}(I \otimes (\mathbf{S}_0 \boldsymbol{\beta})) & I + \mathcal{R}^{-1}(D_0 \oplus S) \end{pmatrix},$$

$$Y_0 = \begin{pmatrix} O & I \\ O & O \end{pmatrix}, Y_2 = \begin{pmatrix} O & I \\ O & \mathcal{R}^{-1}(D_1 \otimes I) \end{pmatrix}$$

where $\mathcal{R}$ is the diagonal matrix with the diagonal entries defined by moduli of the corresponding diagonal entries of the matrix $D_0 \oplus S$.

Stationary probability vectors $\boldsymbol{\pi}_i,\ i \geqslant 0$, of the Markov chain $\hat{\xi}_t$ are computed by means of the algorithm presented before. Once these vectors are computed, the average number of customers in the system at an arbitrary moment is computed by formula

$$L = \sum_{i=1}^{\infty} i \boldsymbol{\pi}_i \mathbf{e} + \sum_{i=0}^{\infty} \boldsymbol{\pi}(i, 1)\mathbf{e},$$

and the probability that an arbitrary customer will not visit the orbit is given by

$$P_{imm} = \lambda^{-1} \sum_{i=0}^{\infty} \boldsymbol{\pi}(i,0)(D_1 \otimes I)\mathbf{e}.$$

2.5 Numerical results

The goal of this section is to show that the proposed mechanism of customers access to the simultaneous service via the pool results in better performance of the system comparing to the classical mechanism and strategy used in [31]. It has already been noted above that the proposed discipline provides higher throughput of the system. Here the aim is to numerically show that, under the proper choice of the pool capacity and the admission period distribution, the adopted strategy provides smaller average number of customers in the system and higher probability that an arbitrary customer will get service without visiting the orbit in comparison to a classical admission strategy and strategy used in [31] for the $MAP/PH/1$ retrial system with the same arrival process, service process of a single customer and retrial intensity.

Now, the way to choose the service time distribution of a group consisting of an arbitrary number of customers is briefly discussed. It has been assumed above that the service time of an individual customer has PH distribution with irreducible representation $(\boldsymbol{\beta}, S)$ and the size of the vector $\boldsymbol{\beta}$ is M. Evidently, it is reasonable to set $(\boldsymbol{\beta}^{(1)}, S^{(1)}) = (\boldsymbol{\beta}, S)$. For the groups of size n, $n = \overline{2, N}$, the most favorable assumption for the discipline under study is that $(\boldsymbol{\beta}^{(n)}, S^{(n)}) = (\boldsymbol{\beta}^{(1)}, S^{(1)})$ for all $n,\ n = \overline{1, N}$, i.e., service time of any group is the same as the service time of an individual customer. Such an assumption is quite realistic, e.g., in applications to some transportation and manufacturing systems. Admission period corresponds to passenger's (items) loading to a vehicle, e.g., bus, (oven) and service time corresponding to passengers delivering or sightseeing providing (thermal processing). Definitely, in such situations service time is the same for any number of passengers (items).

Thinking about possible applications in telecommunication area, it can be assumed that the server has N parallel identical lines and the service of a whole group is defined as the maximum of service times of customers belonging to this group.
Therefore, in this numerical study it is assumed that $(\boldsymbol{\beta}^{(n)}, S^{(n)})$ is defined as irreducible representation of the maximum of n independent identically variables having PH distribution with irreducible representation $(\boldsymbol{\beta}, S)$, $n = \overline{1, N}$. The irreducible representation $(\boldsymbol{\beta}^{(n)}, S^{(n)})$ can be recursively constructed as follows:

$$\boldsymbol{\beta}^{(n)} = \begin{pmatrix} \boldsymbol{\beta} \otimes \boldsymbol{\beta}^{(n-1)} & | & \mathbf{0}_{M_{n-1}} & | & \mathbf{0}_M \end{pmatrix},$$

$$S^{(n)} = \begin{pmatrix} S \oplus S^{(n-1)} & | & \mathbf{S}_0 \otimes I_{M_{n-1}} & | & I_M \otimes \mathbf{S}_0^{(n-1)} \\ ------- & | & ---- & | & ---- \\ O & | & S^{(n-1)} & | & O \\ ------- & | & ---- & | & ---- \\ O & | & O & | & S \\ \end{pmatrix}$$

where dimension M_n of vector $\boldsymbol{\beta}^{(n)}$ is defined by $M_n = (M+1)^n - 1$, $n \geqslant 2$.

Below, results of three numerical experiments are presented, aiming to illustrate advantage of the proposed strategy of customers access comparing to the classical strategy and strategy introduced in [31] for different load of the system and capacity of the pool.

In experiments, three different $MAPs$ having the same fundamental rate $\lambda = 0.6$ but different correlation of successive inter-arrival times will be considered.

The first MAP is the stationary Poisson process. It is defined by $D_0 = -0.6$, $D_1 = 0.6$. The coefficient of variation of inter-arrival times is equal to 1. The coefficient of correlation of successive inter-arrival times is equal to zero, so we will code this process as MAP_0.

The second MAP coded as $MAP_{0.2}$ has coefficient of correlation $c_{cor} = 0.2$ and the squared coefficient of variation 12.34. It is defined by the matrices

$$D_0 = diag\{-0.81156, -0.026346\}, \quad D_1 = \begin{pmatrix} 0.80616 & 0.0054 \\ 0.014676 & 0.01167 \end{pmatrix}.$$

The third MAP coded as $MAP_{0.38}$ has coefficient of correlation $c_{cor} = 0.38$ and coefficient of variation $c_{var}^2 = 12.39$. It is defined by the matrices

$$D_0 = \begin{pmatrix} -2.016 & 0 \\ 0.0006 & -0.0654 \end{pmatrix}, \; D_1 = \begin{pmatrix} 1.995 & 0.021 \\ 0.0072 & 0.0576 \end{pmatrix}.$$

As the main performance measures of the system under study, the probability P_{pool} of service of an arbitrary customer without visiting the orbit, and the average number L of customers in the system at an arbitrary time instant are considered.

In the first experiment, $\alpha = 0.6$ is fixed as individual intensity of retrials, μ as intensity of exponential distribution of admission period, $0 < \mu \leqslant 25$, distribution of service time of individual customer is Erlangian of order 2 with mean value equal to 1. For fair comparison of the used admission discipline with the classical one, it is assumed that service time of the empty pool is exponentially distributed with huge rate (the rate is set equal to 10000). The dependencies of P_{pool} and L on intensity μ for $N = 3$ are compared respect to different admission strategies: classical strategy, strategy proposed in [31] (marked on figures as 'Old strategy') and strategy adopted in this paper (marked on figures as 'New strategy') with different correlation in arrival process. Let it stress again that all these MAP have the same fundamental rate $\lambda = 0.6$.
Figures 2.1-2.2 show dependencies of P_{pool} and L on μ for MAP_0.

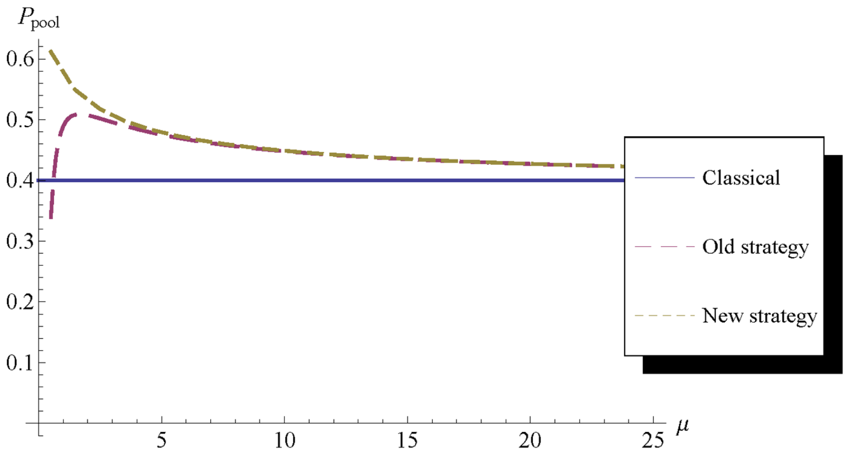


Figure 2.1 P_{pool} for different values of admission rate μ, different strategies of customers admission and correlation 0 in arrival process

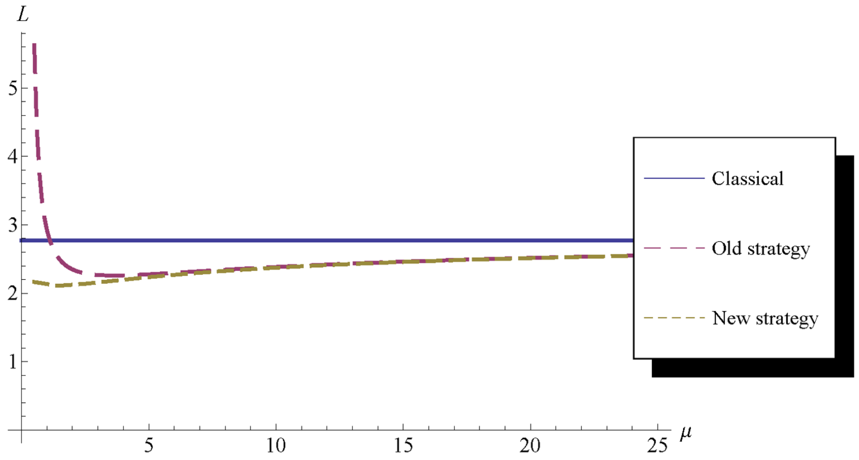


Figure 2.2 L for different values of admission rate μ, different strategies of customers admission and correlation 0 in arrival process

Figures 2.3-2.4 report the behaviour of P_{pool} and L with respect to μ for $MAP_{0.2}$.

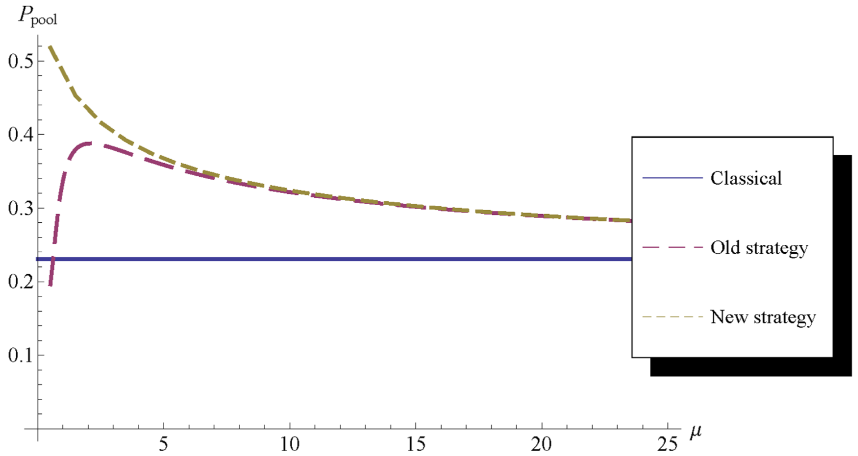


Figure 2.3 P_{pool} for different values of admission rate μ, different strategies of customers admission and correlation 0.2 in arrival process

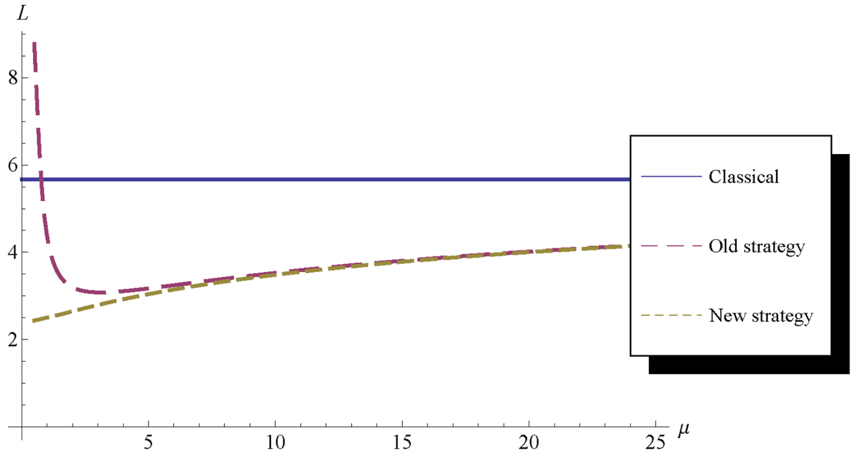


Figure 2.4 L for different values of admission rate μ, different strategies of customers admission and correlation 0.2 in arrival process

Figures 2.5-2.6 depict P_{pool} and L as function of μ for $MAP_{0.38}$.

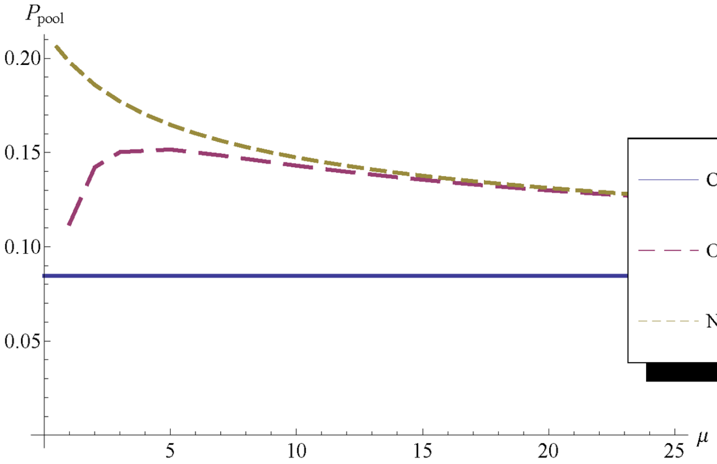


Figure 2.5 P_{pool} for different values of admission rate μ, different strategies of customers admission and correlation 0.38 in arrival process

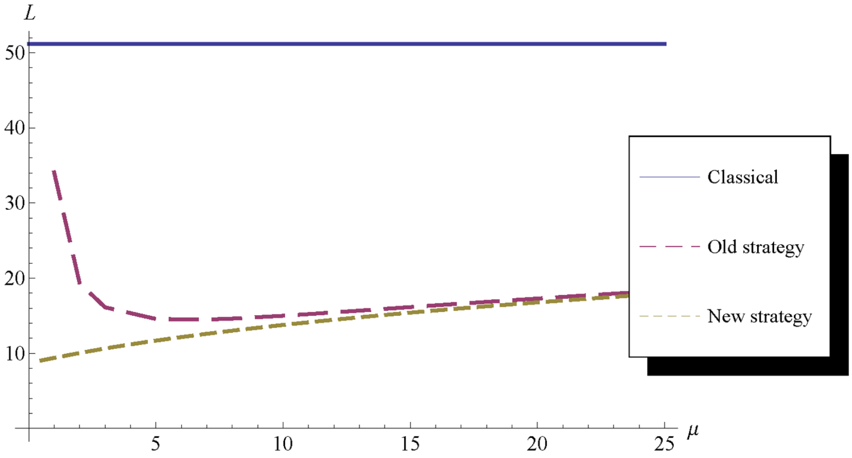


Figure 2.6 L for different values of admission rate μ, different strategies of customers admission and correlation 0.38 in arrival process

Based on these figures, the following conclusions can be made:

- Both strategies of access, offered here and in [31] (conditional on proper choice of capacity of the pool) give much better value of the chosen performance indices, than the classical one, for all considered values of correlation in arrival process.

- Positive correlation in arrival process leads to degradation of performance indices of the system for all strategies of arrival.

- Under unsuccessful choice of duration of admission period (too small μ or too long admission period), old strategy may give performance measures of the system even worth than the classical one. New strategy always gives better values of performance measures than the classical one.

- For large values of μ, results for the new and old strategies converge. This is natural because for large μ (short admission period) admission period always finishes due to the end of admission interval. More early termination of admission period (due to full pool) occurs very rare.

- For the new strategy, the highest value of probability P_{pool} is achieved for very long admission period. For old strategy, the highest value of probability P_{pool} is achieved for some finite value of μ. E.g., in Figure 2.1, the highest value of probability $P_{pool} = 0.5087$ is achieved for $\mu = 0.7$. So, if the main performance measure of the system is probability P_{pool} of service without visiting the orbit, advantage of the new strategy is avoidance of optimal choice of duration of admission period.

- If the main performance measure of the system is the mean number L of customers in orbit, the optimal values of L are achieved at different values of μ. E.g., in Figure 2.2, the smallest value of $L = 2.2594$ in case of the old strategy is achieved when $\mu = 3.6$. The smallest value of $L = 2.1096$ in case of the new strategy is achieved when $\mu = 1.4$. Classical strategy provides value $L = 2.7747$ for all μ. It is clear from Figures 2.3-2.4 and 2.5-2.6 that the profit of strategies of customers access via the pool becomes huge. E.g., in Figure 2.6, $L = 51,165$ under the classical strategy and L is less than 18 for the new strategy of customers access via the pool.

In the second experiment, $MAP_{0.2}$ is fixed as arrival flow scaled in such a way that the mean arrival rate is 0.9, not 0.6 as in the previous example and $\alpha = 0.9$ as individual intensity of retrials.

Figures 2.7-2.8 illustrate the effect of varying the intensity μ in this example.

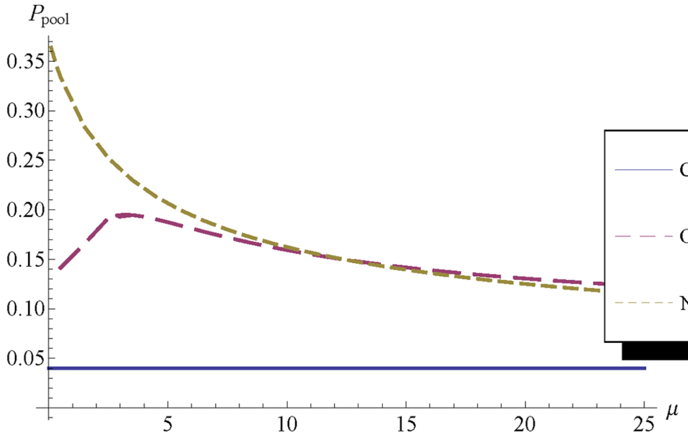


Figure 2.7 P_{pool} for different values of admission rate μ, and different strategies of customers admission for $\lambda = 0.9$

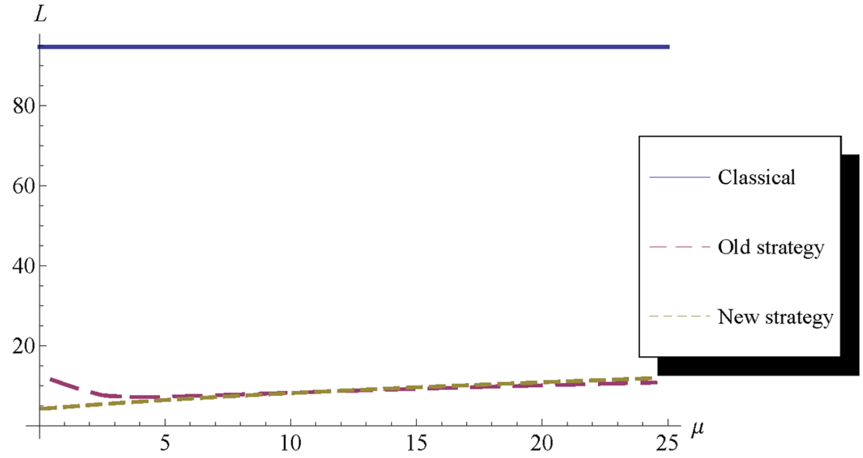


Figure 2.8 L for different values of admission rate μ, and different strategies of customers admission for $\lambda = 0.9$

Comparing these figures with Figures 2.3-2.4, correspondingly, it can be seen that in case of high load of the system the profit gained by application of the proposed strategy of customers admission becomes more essential.

In the third experiment, again $MAP_{0.2}$ is fixed as arrival flow with the mean arrival rate 0.6, individual retrial rate is set equal to 0.6 and only admission strategy analyzed in this paper is considered.

Figures 2.9-2.10 illustrate the effect of varying the intensity μ for various values of capacity N of the pool.

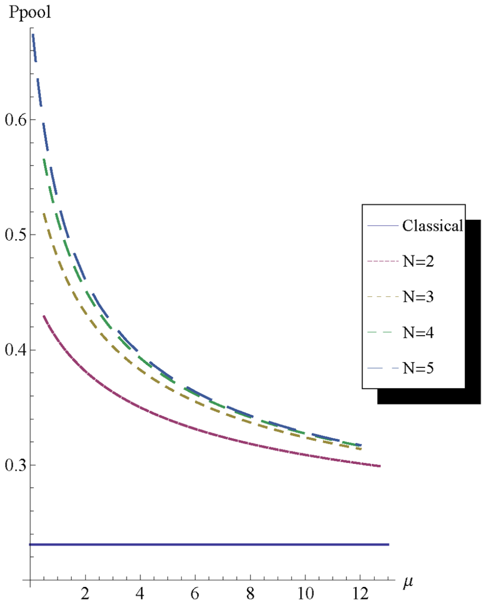


Figure 2.9 P_{pool} for different values of admission rate μ, and different N

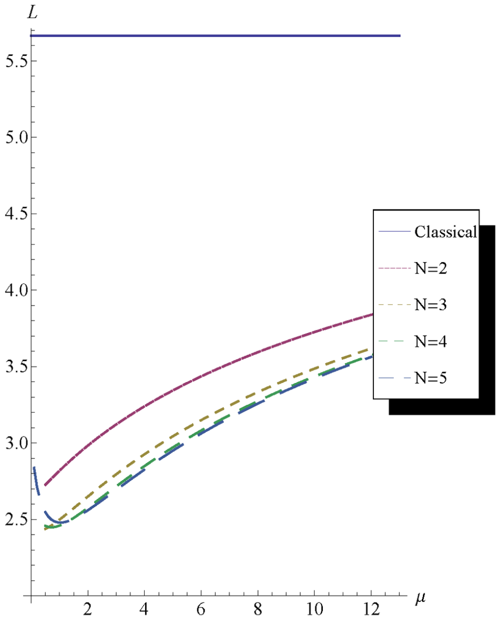


Figure 2.10 L for different values of admission rate μ, and different N

Based on these figures, one can conclude that probability P_{pool} increases when N grows while the optimal (with respect to μ) average number L of customers in the system may decrease when N grows. For $N = 2$ the minimal value of L is achieved when $\mu = 0.5$ and is equal to 2.72588. For $N = 3$ the minimal value of L is achieved when $\mu = 0.5$ and is equal to 2.43734. For $N = 4$ the minimal value of L is attained when $\mu = 0.7$ and is equal to 2.44807. For $N = 5$ the minimal value of L is obtained when $\mu = 1$ and is equal to 2.47889. Therefore, the optimal, with respect of L, value of N in this example is equal to 3.

It follows from the figures that the strategy of customers admission via the pool gives significant improvement of the system performance measures comparing to the classical strategy. However, careful choice of parameters N and μ, which characterize the introduced strategy of customers admission, is necessary to achieve the best performance for any fixed set of the system parameters. This choice may be different depending on which performance measure P_{pool} or L (or, may be, some other) is the most important in concrete application of the model.

Chapter 3

A MAP/PH/1 Retrial System with Permanent Pooling Admission Strategy

In this chapter it will be introduced and analysed a novel discipline of customers admission for a single server retrial queue, starting from the model described in the previous chapter.

Comparing to the admission discipline considered in Chapter 2 and in [31], the following significant improvements are proposed:

- It is supposed in [31] that service is resumed only after that the admission period expires (even if the pool is already full). This assumption holds good for many real world systems, in which the admission period cannot be terminated ahead of the schedule. E.g., in polling systems, in which the server indeed provides service to many queues and during the admission period it provides service to other queues or in systems, in which a certain sequence of technological operations should be implemented during the admission period and no one of these operations can be cancelled.
  Here and in the model described in Chapter 2 it is assumed

that such an admission period interruption is possible and service is resumed also if the admission period does not expire but the pool becomes full.

- In the models introduced in Chapter 2 and in [31], it is assumed that admission period starts when the service period finishes and customers are not admitted and go to the orbit during the service period. Here it is assumed that admission of customers continues during service period as well. So, if the pool becomes full during a service period, a new service period starts immediately.
  This assumption makes the current model more realistic in context of analysis of IEEE802.11 WLAN multi-rate protocols because no special admission period after each service period is supposed in these protocols. Mobile stations can prepare information for transmission during the next frame within the current frame.

- Because the pool is always empty at service completion moment in [31] and in the model described in Chapter 2, durations of successive admission periods are independent identically distributed random variables. In the model under study, the pool may be not empty at service completion moment and it is assumed that admission period is skipped if the pool is already full and the distribution of duration of admission period depends on the number of customers in the pool at beginning of the admission period, otherwise. Because the considered model describes operation of the existing networks with IEEE802.11 WLAN multi-rate protocols, while it was not analysed in queueing literature before, the results obtained by this novel strategy are very important for practice. Even small improvements of existing protocols can lead to essential increase of the bandwidth of the system. The obtained results can help to optimally adjust the parameters of the protocol to parameters of the arrival flows and transmission rates. It is worth to note that the standard versions of multi-rate protocols suggest immediate

> start of service of the next group after service completion of the previous group even if the size of the group is less than the recommended one. Therefore, the evident advantages of simultaneous transmission of a large group are exploited not in the full extent.

Such an admission discipline is realistic not only in some wireless networks, e.g., in multi-rate IEEE802.11 WLAN, but also in multi-rate protocols, several mobile stations share the same physical channel, see, e.g., [67]. Under the use of such protocols, a group of requests from users can be processed simultaneously in parallel and the processing of the whole group is considered finished if processing of all individual requests belonging to the group is completed.

Intuitive reasonability of the proposed discipline can be also deduced by real models different from telecommunication systems. Let the service of customers consist of sightseeing from the bus (or aircraft). The server is the bus driver (or pilot). Admission period consists of time during which passengers should occupy free places in the bus (boarding time). Discipline in [31] suggests that duration of admission period is randomly defined and even the bus becomes full, the bus will not start sightseeing tour before admission period finishes. Here it is assumed that, when the bus becomes full earlier than admission period ends, the bus starts the tour. To reduce waiting time of potential customers and potentially increase the profit, one can rent not one, but two buses. When one bus starts the tour, the newly arriving passengers are allowed to start boarding to the second bus (this was not allowed by the discipline analysed in the previous chapter). If the first bus returns from the tour and the second bus is already full, the driver moves to the second bus and starts the tour while the first bus is ready for boarding. If the second bus is not full, it makes sense to wait for a while for arrival of more passengers. This waiting time has to be longer if the bus was almost empty and can be quite short if the bus was almost full.

It has already been said that discipline of simultaneous service of a whole group of customers, instead of service one-by-one, was already considered in literature as bulk service discipline. Several works have been cited in the previous chapters. Essential difference of the current model compared to the cited papers is the following. Here it is allowed to start the service after certain random time, called as admission period, even if the number of customers ready for service does not reach the threshold value. This makes the admission strategy more flexible providing better quality of service to customers.

Last, but not the least, practically all papers in the field of bulk queues regard systems with a finite or infinite buffer. Here, like in the previous introduced strategy, customers retrials are considered. The Markov chain describing behavior of the considered system is not space homogeneous and the well known powerful tool for analysis of multi-dimensional Markov chains, namely, theory of Quasi-Birth-and-Death Processes is not applicable here. Note that the queue with bulk service and queueing system introduced in this chapter are in some sense very close to so called queues with assembly-like service, see [61]. In those queues, service can be started only when all buffers of the system are not empty. So, the set of places at the head of each queue can be interpreted as a pool.

The analysed model can be also interpreted as a vacation model, see, e.g. the book [66] and recent papers [40, 60, 64], in which admission period can be seen as vacation time. Service discipline is a bit similar to the N-limited gated discipline.

3.1 The mathematical model

The system under study in this chapter is a single server retrial queueing system, in which the input flow is described by a Markovian Arrival Process. Customer's arrival in the MAP is directed by an underlying irreducible continuous time Markov chain ν_t, $t \geq 0$, with the finite state space $\{0, ..., W\}$. Sojourn time of the Markov chain ν_t, $t \geq 0$, in the state ν has exponential distribution with parameter $\lambda_\nu, \nu = \overline{0, W}$. After this sojourn time expires, with probability $p_k(\nu, \nu')$, the process ν_t jumps to the state ν', and k customers, $k = 0, 1$, arrive into the system. The intensities of jumps of underlying Markov chain from one state into another with generation of k customers are combined into the matrices $D_k,\ k = 0, 1$, of size $(W+1) \times (W+1)$. The matrix $D(1) = D_0 + D_1$ is the infinitesimal generator of the process $\nu_t,\ t \geq 0$. The invariant probability vector (vector of stationary distribution) $\boldsymbol{\theta}$ of this process is computed as the unique solution to the equations

$$\boldsymbol{\theta} D(1) = \mathbf{0},\ \boldsymbol{\theta}\mathbf{e} = 1.$$

The average intensity λ (fundamental rate) of the MAP is defined as $\lambda = \boldsymbol{\theta} D_1 \mathbf{e}$ and gives the expected number of arrivals per unit of time in the stationary mode. The variance v of intervals between customer arrivals is calculated as $v = 2\lambda^{-1}\boldsymbol{\theta}(-D_0)^{-1}\mathbf{e} - \lambda^{-2}$, the squared coefficient c_{var} of variation is equal to $c_{var} = 2\lambda\boldsymbol{\theta}(-D_0)^{-1}\mathbf{e} - 1$, while the coefficient c_{cor} of correlation of successive intervals between arrivals is given by

$$c_{cor} = (\lambda^{-1}\boldsymbol{\theta}(-D_0)^{-1} D_1 (-D_0)^{-1}\mathbf{e} - \lambda^{-2})/v.$$

In the system under study, customers can enter the service only after joining a pool. The pool is some virtual place where the admitted customer should wait for starting the service. Capacity of the pool is assumed to be equal to an integer number N, $1 \leqslant N < \infty$. If the pool is full at the customer arrival instant, this customer moves to a virtual place having an infinite capacity, the so called orbit, from which tries to get access to the service after random intervals of time. All customers admitted to the pool are processed simultaneously. The pool becomes empty at the moment of service beginning. After emptying the pool, accumulation of customers starts again.

The customer staying in the orbit repeats, independently of other customers, the attempts to get service in random time intervals whose duration has exponential distribution. Therefore, under the fixed number of customers in the orbit, the total flow of retrials from the orbit is characterized by inter-arrival times having exponential distribution. Parameter of this distribution is α_i when

$i,\ i \geqslant 0$, customers stay in the orbit. Any dependence of the intensity α_i on i is admitted such as α_i is monotonically increasing when i becomes large and tends to infinity when i approaches infinity. Cases $\alpha_i = i\alpha$ and $\alpha_i = i\alpha + \gamma,\ i \geqslant 1,\ \alpha > 0,\ \ \gamma \geqslant 0,\ \alpha_0 = 0$, which satisfy the mentioned conditions, are popular in literature. If, during the repeated attempt, the pool is not full, the customer moves from the orbit to the pool. If the pool is full, the customer returns to the orbit and repeats its attempts until it succeeds to enter the pool.

Server operates as follows. The pool in not locked during the service period. Primary customers or customers from the orbit can enter the system during both the admission and service periods, provided that the pool is not full. The admission period is terminated either if the admission interval expires or the pool becomes full. As it was already stated above, when service period starts, all customers from the pool get service simultaneously, as one group. The number of customers in a group can vary from 1 to N. Service time of a group consisting of n customers has PH type distribution with irreducible representation $(\boldsymbol{\beta}^{(n)}, S^{(n)}), n = \overline{0, N}$. If, at the service completion instant, the pool is full, another service period starts immediately. Duration of this period has PH type distribution with irreducible representation $(\boldsymbol{\beta}^{(N)}, S^{(N)})$. If there are only $n,\ 0 \leqslant n \leqslant N-1$, customers in the pool, admission interval starts. Duration of this interval is random. It has PH type distribution with irreducible representation $(\boldsymbol{\tau}^{(m)}, T^{(m)})$, where $m = \overline{0, n}$ is the number of customers in the pool at admission period beginning.

It means the following. Duration of customer's admission interval is governed by the underlying process $\eta_t^{(a,m)},\ t \geqslant 0$, which is a continuous time Markov chain with state space $\{1, \ldots, M^{(a,m)}\}$. The initial state of the process $\eta_t^{(a,m)},\ t \geqslant 0$, at the epoch of starting the admission interval is determined by the probabilistic row-vector $\boldsymbol{\tau}^{(m)} = (\tau_1, \ldots, \tau_{M^{(a,m)}})$. The intensities of transitions of the process $\eta_t^{(a,m)},\ t \geqslant 0$, that do not lead to admission interval completion, are defined by the irreducible matrix $T^{(m)}$ of size $M^{(a,m)} \times M^{(a,m)}$. The intensities of transitions, which lead to admission interval completion, are given by the column-vector $\mathbf{T}_0^{(m)} = -T^{(m)}\mathbf{e}$. The admission interval time distribution function has the form $T^{(m)}(x) = 1 - \boldsymbol{\tau}^{(m)} e^{T^{(m)}x}\mathbf{e}$. Laplace-Stieltjes transform $\int\limits_0^\infty e^{-sx} dT^{(m)}(x)$ of this distribution function is $\boldsymbol{\tau}^{(m)}(sI - T^{(m)})^{-1}\mathbf{T}_0^{(m)}$. The average length of admission interval time is given by

$$r_1^{(m)} = \boldsymbol{\tau}^{(m)} \left(-T^{(m)}\right)^{-1} \mathbf{e}.$$

Intuitively in this model, the larger is m, the shorter the average length of admission period should be.
Further the value $\mu^{(m)} = \left(r_1^{(m)}\right)^{-1}$ will represent the intensity of admission. The matrix $T^{(m)} + \mathbf{T}_0^{(m)}\boldsymbol{\tau}^{(m)}$ is assumed to be irreducible.
Service time of a group consisting of n customers has PH type distribution with irreducible representation $(\boldsymbol{\beta}^{(n)}, S^{(n)})$. Underlying process of this distribution is $\eta_t^{(s,n)}$, $t \geqslant 0$, with a finite state space $\{1, \dots, M^{(s,n)}\}$. The average service time of a group of n customers is defined by formula

$$b_1^{(n)} = \boldsymbol{\beta}^{(n)}(-S^{(n)})^{-1}\mathbf{e}, \ n = \overline{0, N}.$$

For simplification of notation, it can be assumed that the customer may provide the service even if the number of customers in the pool after completion of admission interval is equal to 0. This service may be interpreted, e.g. as maintenance or vacation or sleep period of the server. If one would like to exclude service provisioning to a group of size 0 from the model, he or she may set in our algorithms $M^{(s,0)} = 1$, $\boldsymbol{\beta}^{(s,0)} = 1$, and $S^{(0)}$ be equal to a very large in modulus negative number. In this case, the server takes another customer's admission interval if the pool is empty at a given moment of customer's admission interval completion.

From the point of view of mathematical generality, in further derivations the way of choosing the irreducible representations $(\boldsymbol{\beta}^{(n)}, S^{(n)})$, $n = \overline{0, N}$ will not strictly be specified. Only the following two assumptions, the same as in [31] and in the model in the previous chapter, are made:

(i) Let the service time of an individual customer have PH type distribution with irreducible representation $(\boldsymbol{\beta}, S)$ and average value $b_1 = \boldsymbol{\beta}(-S)^{-1}\mathbf{e}$. Size of the vector $\boldsymbol{\beta}$ is assumed to be equal to some integer $M,\ M \geqslant 1$.

(ii) The following inequalities

$$b_1^{(1)} \leqslant b_1^{(2)} \leqslant \ldots \leqslant b_1^{(N)} < N b_1^{(1)}$$

are fulfilled.

In numerical results section, concrete form of the irreducible representations $(\boldsymbol{\beta}^{(n)}, S^{(n)}),\ \ n = \overline{1, N},$ will be chosen as representations of distribution of maximum of n independent random variables having PH type distribution with irreducible representation $(\boldsymbol{\beta}, S)$.

The main goals are the same of the analysis of the previous model: to compute the average number of customers in the system (in the orbit and in the pool and in the service) at arbitrary time moment and the probability that an arbitrary customer avoids visiting of orbit and, then, to find the optimal mean value of customer's admission interval under different values of the pool capacity N and different correlation in the arrival process. To this end, it is necessary to get stability condition of the system and compute the stationary distribution of the system states as well as to derive formulas for computation of the main performance measures of the system for any fixed set of its parameters.

3.2 The process of the system states

Let

- i_t be the number of customers in the orbit, $i_t \geqslant 0$,
- r_t be the current state of the server: $r_t = 0$ if admission period is in a progress and $r_t = 1$ if server provides the service,

- n_t be equal to the number of customers in the pool, $n_t = \overline{0, N-1}$ if $r_t = 0$ or $n_t = \overline{0, N}$ if $r_t = 1$,

- m_t be equal to the number of customers in the pool at admission period beginning, $m_t = \overline{0, n_t}$ if $r_t = 0$, or the number of customers in service, $m_t = \overline{0, N}$ if $r_t = 1$,

- ν_t be the state of the underlying process of the MAP, $\nu_t = \overline{0, W}$,

- η_t be the state of the underlying process of the PH process of customers admission, $\eta_t = \overline{1, M^{(a,m_t)}}$, $m_t = \overline{0, n_t}$, if $r_t = 0$ or the state of the underlying process of the PH process of customers service, $\eta_t = \overline{1, M^{(s,m_t)}}$, $m_t = \overline{0, N}$, if $r_t = 1$,

at the epoch $t,\ t \geq 0$.

The six-dimensional process $\xi_t = \{i_t,\ r_t,\ n_t,\ m_t,\ \nu_t,\ \eta_t\},\ \ t \geq 0$, is an irreducible continuous time Markov chain with one component (i_t) having infinite state space and five finite components.

In order to analyse behavior and properties of the Markov chain ξ_t, the infinitesimal generator of the chain has to be computed. Let this generator be denoted as $\mathbf{Q}$. The diagonal entries $\mathbf{Q}_{(i,r,n,m,\nu,\eta),(i,r,n,m,\nu,\eta)}$ are negative. Modulus of each diagonal entry defines intensity of departure of the Markov chain from the corresponding state of the Markov chain. The non-diagonal entry $\mathbf{Q}_{(i,r,n,m,\nu,\eta),(i',r',n',m',\nu',\eta')}$ is non-negative and defines intensity of transition of the Markov chain from the state (i, r, n, m, ν, η) to the state $(i', r', n', m', \nu', \eta')$.

To simplify the structure of generator $\mathbf{Q}$ and following traditional methodology of analysis of multi-dimensional Markov chains, let the states of the Markov chain ξ_t be enumerated in the lexicographic order and let all the states of the chain having value (i, r, n, m) as the first four components be composed to a *sub-level* (i, r, n, m). Then, sub-levels (i, r, n, m) are composed to the *level*

(i,r,n). The level $(i,0,n),\ n=\overline{0,N-1}$, contains

$$K^{(a,n)}=\overline{W}\sum_{m=0}^{n}M^{(a,m)}$$

states and the level $(i,1,n),\ n=\overline{0,N}$, contains

$$K^{(s)}=\overline{W}\sum_{m=0}^{N}M^{(s,m)}$$

states. Analogously, the levels (i,r,n) will be composed to *macro-level* (i,r), and then a *super-level* i is formed as a composition of macro-levels $(i,r),\ r=0,1,\ \ i\geq 0$.

Lemma 7. *Generator* $\mathbf{Q}$ *has three block diagonal structure:*

$$\mathbf{Q}=\begin{pmatrix} \mathbf{Q}_{0,0} & \mathbf{Q}_{0,1} & O & O & \dots \\ \mathbf{Q}_{1,0} & \mathbf{Q}_{1,1} & \mathbf{Q}_{1,2} & O & \dots \\ O & \mathbf{Q}_{2,1} & \mathbf{Q}_{2,2} & \mathbf{Q}_{2,3} & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

where non-zero blocks $\mathbf{Q}_{i,j}$ *define intensities of transitions from super-level* i *to super-level* $j,\ \ j=\max\{0,i-1\},i,i+1$.

The matrix $\mathbf{Q}_{i,i}$ *is defined as follows:*

$$\mathbf{Q}_{i,i}=\begin{pmatrix} \mathbf{Q}_{(i,0),(i,0)} & \mathbf{Q}_{(i,0),(i,1)} \\ \mathbf{Q}_{(i,1),(i,0)} & \mathbf{Q}_{(i,1),(i,1)} \end{pmatrix}$$

where

- $\mathbf{Q}_{(i,0),(i,0)}$ *is a two block diagonal matrix with the diagonal blocks defined by*

$$\mathbf{Q}_{(i,0,n),(i,0,n)}=$$
$$diag\left\{D_0\otimes I_{M^{(a,m)}}-\alpha_i I_{\overline{W}M^{(a,m)}}+I_{\overline{W}}\otimes\mathbf{T}^{(m)},m=\overline{0,n}\right\},$$
$$n=\overline{0,N-1},$$

and the up-diagonal blocks defined by the matrices obtained by supplementing the square matrix diag $\left\{D_1 \otimes I_{M^{(a,m)}}, m = \overline{0,n}\right\}$, $n = \overline{0, N-2}$, *with the zero block column from the right:*

$$\mathbf{Q}_{(i,0,n),(i,0,n+1)} = \begin{pmatrix} D_1 \otimes I_{M^{(a,0)}} & & & O \\ & \ddots & & \vdots \\ & & \ddots & \vdots \\ & & D_1 \otimes I_{M^{(a,n)}} & O \end{pmatrix}.$$

- $\mathbf{Q}_{(i,0),(i,1)}$ *is the non-square matrix with* N *block rows and* $N+1$ *block columns having the form:*

$$\mathbf{Q}_{(i,0),(i,1)} = \begin{pmatrix} \mathbf{Q}_{(i,0,0),(i,1,0)} & O & \dots & \dots & O \\ \vdots & \vdots & \ddots & & \vdots \\ \mathbf{Q}_{(i,0,N-2),(i,1,0)} & \vdots & & \ddots & \vdots \\ \mathbf{Q}_{(i,0,N-1),(i,1,0)} & O & \dots & \dots & O \end{pmatrix},$$

where $\mathbf{Q}_{(i,0,n),(i,1,0)}, n = \overline{0, N-2}$, *is the non-square matrix having* $n+1$ *block rows and* $N+1$ *block columns which consists of all zero blocks except the nth block column defined by:*

$$\mathbf{Q}_{(i,0,n),(i,1,0)} = \begin{pmatrix} O & \dots & O & I_{\overline{W}} \otimes \mathbf{T}_0^{(0)} \otimes \boldsymbol{\beta}^{(n)} & O & \dots & O \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ O & \dots & O & I_{\overline{W}} \otimes \mathbf{T}_0^{(n)} \otimes \boldsymbol{\beta}^{(n)} & O & \dots & O \end{pmatrix},$$

and $\mathbf{Q}_{(i,0,N-1),(i,1,0)}$ *is the non-square matrix having* N *block rows and* $N+1$ *block columns defined by:*

$$\mathbf{Q}_{(i,0,N-1),(i,1,0)} =$$
$$\begin{pmatrix} O & \dots & O & I_{\overline{W}} \otimes \mathbf{T}_0^{(0)} \otimes \boldsymbol{\beta}^{(N-1)} & D_1 \otimes \mathbf{e}_{M^{(a,0)}} \otimes \boldsymbol{\beta}^{(N)} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ O & \dots & O & I_{\overline{W}} \otimes \mathbf{T}_0^{(N-1)} \otimes \boldsymbol{\beta}^{(N-1)} & D_1 \otimes \mathbf{e}_{M^{(a,N-1)}} \otimes \boldsymbol{\beta}^{(N)} \end{pmatrix};$$

- $\mathbf{Q}_{(i,1),(i,0)}$ *is the non-square matrix of the form:*

$$\mathbf{Q}_{(i,1),(i,0)} = \begin{pmatrix} \mathbf{Q}_{(i,1,0),(i,0,0)} & & \\ & \ddots & \\ & & \mathbf{Q}_{(i,1,N-1),(i,0,N-1)} \\ O & \dots & O \end{pmatrix}$$

where $\mathbf{Q}_{(i,1,n),(i,0,n)}, n = \overline{0, N-1}$, *is the non-square matrix having* $N+1$ *block rows and* $n+1$ *block columns of the form:*

$$\mathbf{Q}_{(i,1,n),(i,0,n)} = \begin{pmatrix} O & \dots & O & I_{\overline{W}} \otimes \mathbf{S}_0^{(0)} \otimes \boldsymbol{\tau}^{(n)} \\ \vdots & \ddots & \vdots & \vdots \\ O & \dots & O & I_{\overline{W}} \otimes \mathbf{S}_0^{(N)} \otimes \boldsymbol{\tau}^{(n)} \end{pmatrix},$$

here $\mathbf{S}_0^{(n)} = S^{(n)}\mathbf{e},\ n = \overline{0, N}$;

- $\mathbf{Q}_{(i,1),(i,1)}$ *is the square matrix with* $N+1$ *block rows and* $N+1$ *block columns having the structure presented below where all blocks are zero matrices except the blocks marked by symbol* $*$*:*

$$\mathbf{Q}_{(i,1),(i,1)} = \begin{pmatrix} * & * & & & \\ & * & * & & \\ & & * & & \\ & & & \ddots & \\ & & & & * \\ * & & & & * \end{pmatrix}.$$

Here the diagonal blocks are defined by

$$\mathbf{Q}_{(i,1,n),(i,1,n)} = \\ diag\left\{D_0 \otimes I_{M^{(s,m)}} - \alpha_i\left(1 - \delta_{n,N}\right) I_{\overline{W}M^{(s,m)}} \right. \\ \left. + I_{\overline{W}} \otimes \mathbf{S}^{(m)}, m = \overline{0, N}\right\}, n = \overline{0, N},$$

the up-diagonal blocks are given by

$$\mathbf{Q}_{(i,1,n),(i,1,n+1)} = diag\left\{D_1 \otimes I_{M^{(s,m)}}, m = \overline{0, N}\right\}, n = \overline{0, N-1},$$

and the non-zero block at left-low corner equal to

$$\mathbf{Q}_{(i,1,N),(i,1,0)} = \begin{pmatrix} O & \dots & O & I_{\overline{W}} \otimes \mathbf{S}_0^{(0)} \otimes \boldsymbol{\beta}^{(N)} \\ \vdots & \ddots & \vdots & \vdots \\ O & \dots & O & I_{\overline{W}} \otimes \mathbf{S}_0^{(N)} \otimes \boldsymbol{\beta}^{(N)} \end{pmatrix}.$$

The matrix $\mathbf{Q}_{i,i-1}$ *has this form:*

$$\mathbf{Q}_{i,i-1} = \begin{pmatrix} \mathbf{Q}_{(i,0),(i-1,0)} & \mathbf{Q}_{(i,0),(i-1,1)} \\ O & \mathbf{Q}_{(i,1),(i-1,1)} \end{pmatrix}$$

where

- $\mathbf{Q}_{(i,0),(i-1,0)}$ *is the square matrix of the form:*

  $$\mathbf{Q}_{(i,0),(i-1,0)} = \mathrm{diag}^{+}\{\mathbf{Q}_{(i,0,n),(i-1,0,n+1)}, n = \overline{0, N-2}\},$$

  and $\mathbf{Q}_{(i,0,n),(i-1,0,n+1)}, n = \overline{0, N-2}$, *is the non-square matrix of the form:*

  $$\mathbf{Q}_{(i,0,n),(i-1,0,n+1)} = \alpha_i \begin{pmatrix} I_{\overline{W}M^{(a,0)}} & & & O \\ & \ddots & & \vdots \\ & & I_{\overline{W}M^{(a,n)}} & O \end{pmatrix};$$

- $\mathbf{Q}_{(i,0),(i-1,1)}$ *is the non-square matrix having* N *block rows and* $N+1$ *block columns of the form:*

  $$\mathbf{Q}_{(i,0),(i-1,1)} = \begin{pmatrix} O & O & \dots & \dots & O \\ \vdots & \vdots & \ddots & & \vdots \\ O & \vdots & & \ddots & \vdots \\ \mathbf{Q}_{(i,0,N-1),(i-1,1,0)} & O & \dots & \dots & O \end{pmatrix}$$

  with $\mathbf{Q}_{(i,0,N-1),(i-1,1,0)}$ *the non-square matrix having* N *block rows and* $N+1$ *block columns of the form:*

  $$\mathbf{Q}_{(i,0,N-1),(i-1,1,0)} = \alpha_i \begin{pmatrix} O & \dots & O & I_{\overline{W}} \otimes \mathbf{e}_{M^{(a,0)}} \otimes \boldsymbol{\beta}^{(N)} \\ \vdots & \ddots & \vdots & \vdots \\ O & \dots & O & I_{\overline{W}} \otimes \mathbf{e}_{M^{(a,N-1)}} \otimes \boldsymbol{\beta}^{(N)} \end{pmatrix};$$

- $\mathbf{Q}_{(i,1),(i-1,1)}$ *is the square matrix having* $N+1$ *block rows and block columns defined by*

  $$\mathbf{Q}_{(i,1),(i-1,1)} = \mathrm{diag}^{+}\{\mathbf{Q}_{(i,1,n),(i-1,1,n+1)}, n = \overline{0, N-1}\},$$

  with

  $$\mathbf{Q}_{(i,1,n),(i-1,1,n+1)} = \alpha_i \mathrm{diag}\{I_{\overline{W}} \otimes I_{M^{(s,m)}}, m = \overline{0, N}\}.$$

The matrix $\mathbf{Q}_{i,i+1}$ *is the following:*

$$\mathbf{Q}_{i,i+1} = \begin{pmatrix} O & O \\ O & \mathbf{Q}_{(i,1),(i+1,1)} \end{pmatrix}$$

where

- $\mathbf{Q}_{(i,1),(i+1,1)}$ *is the matrix of the form:*

  $$\mathbf{Q}_{(i,1),(i+1,1)} = \mathrm{diag}\{O, \mathbf{Q}_{(i,1,N),(i+1,1,N)}\},$$

  with

  $$\mathbf{Q}_{(i,1,N),(i+1,1,N)} = \mathrm{diag}\{D_1 \otimes I_{M^{(s,0)}}, \dots, D_1 \otimes I_{M^{(s,N)}}\}.$$

Proof of the lemma consists of analysis of the Markov chain $\xi_t,\ t \geq 0$, transitions during the infinitesimal interval of time and further combining corresponding transition intensities into the matrix blocks. Expressions for blocks $\mathbf{Q}_{i,j}$ here are more cumbersome than in [31] and in Chapter 2, due to the noted above existence of six, not five, components of the Markov chain ξ_t. Symbols of Kronecker product and sum of matrices are very useful here for a compact description of joint transition of several independent Markov processes.

Because the form of the blocks $\mathbf{Q}_{i,j}$ does not depend only on the difference $j-i$, but depends on i and j separately, the Markov chain ξ_t does not belong to well known class of Quasi-Birth-and-Death processes, see [58], for which the stationary distribution of the states has the matrix-geometric form. The Markov chain ξ_t is

level-dependent Quasi-Birth-and-Death process.
Fortunately, this Markov chain belongs to the class of Asymptotically Quasi-Toeplitz Markov Chains ($AQTMCs$) introduced and analysed in paper [49]. Let this fact be proved.

Let the matrix $\mathcal{K}_n^{(a)}$ be defined as follows

$$\mathcal{K}_n^{(a)} = \text{diag}\left\{\Sigma_n^{(a,0)}, ..., \Sigma_n^{(a,n)}\right\},$$

where $\Sigma_n^{(a,m)}$ is the diagonal matrix with the diagonal entries given by the moduli of the diagonal entries of the matrices $D_0 \oplus T^{(m)}$, $m = \overline{0, n}$.

Let the matrix $\mathcal{K}_n^{(s)}$ be defined as follows

$$\mathcal{K}_n^{(s)} = \text{diag}\left\{\Sigma_n^{(s,0)}, ..., \Sigma_n^{(s,N)}\right\},$$

where $\Sigma_n^{(s,m)}$ is the diagonal matrix with the diagonal entries equal to the moduli of the diagonal entries of the matrices $D_0 \oplus S^{(m)}$, $m = \overline{0, N}$.

Let $\mathbf{R}_i$ be the diagonal matrix with the diagonal entries given by the moduli of the diagonal entries of the matrix $\mathbf{Q}_{i,i}$. It can be verified that the matrix $\mathbf{R}_i$ is defined by the formula

$$\mathbf{R}_i = \text{diag}\left\{\text{diag}\left\{\mathcal{K}_n^{(a)} + \alpha_i I_{K^{(a,n)}}, n = \overline{0, N-1}\right\},\right.$$
$$\left.\text{diag}\left\{\text{diag}\left\{\mathcal{K}_n^{(s)} + \alpha_i I_{K^{(s)}}, n = \overline{0, N-1}\right\}, \mathcal{K}_N^{(s)}\right\}\right\}.$$

Lemma 8. *The following limits exist*

$$\mathbf{Y}_0 = \lim_{i\to\infty} \mathbf{R}_i^{-1}\mathbf{Q}_{i,i-1},\ \mathbf{Y}_1 = \lim_{i\to\infty} \mathbf{R}_i^{-1}\mathbf{Q}_{i,i} + I,\ \mathbf{Y}_2 = \lim_{i\to\infty} \mathbf{R}_i^{-1}\mathbf{Q}_{i,i+1},$$

and are defined by:

- $$\mathbf{Y}_0 = \begin{pmatrix} E_N^+ \otimes I_{\overline{W}M^{(a)}} & \mathbf{Y}_0^{(0,1)} \\ O & E_{N+1}^+ \otimes I_{\overline{W}M^{(s)}} \end{pmatrix}$$

where $\mathbf{Y}_0^{(0,1)}$ *is the matrix of the form*

$$\mathbf{Y}_0^{(0,1)} = \begin{pmatrix} O & O & \cdots & \cdots & O \\ \vdots & \vdots & \ddots & & \vdots \\ O & O & \cdots & \cdots & O \\ \mathbf{A} & O & \cdots & \cdots & O \end{pmatrix}$$

with a non-square matrix $\mathbf{A}$ *having* N *block rows and* $N+1$ *block columns of the form*

$$\mathbf{A} = \begin{pmatrix} O & \dots & O & I_{\overline{W}} \otimes \mathbf{e}_{M^{(a,0)}} \otimes \boldsymbol{\beta}^{(N)} \\ \vdots & \ddots & \vdots & \vdots \\ O & \dots & O & I_{\overline{W}} \otimes \mathbf{e}_{M^{(a,N-1)}} \otimes \boldsymbol{\beta}^{(N)} \end{pmatrix};$$

- $$\mathbf{Y}_1 = \begin{pmatrix} O & O \\ O & \mathbf{Y}_1^{(1,1)} \end{pmatrix}$$

where $\mathbf{Y}_1^{(1,1)}$ *is the square matrix of the form:*

$$\mathbf{Y}_1^{(1,1)} = \begin{pmatrix} O & \cdots & \cdots & O & O \\ \vdots & \ddots & & \vdots & \vdots \\ \vdots & & \ddots & \vdots & \vdots \\ O & \cdots & \cdots & O & O \\ \mathbf{B} & O & \cdots & O & \mathbf{C} \end{pmatrix},$$

with

$$\mathbf{B} = \begin{pmatrix} O & \dots & O & \left(\Sigma_N^{(s,0)}\right)^{-1}\left(I_{\overline{W}} \otimes \mathbf{S}_0^{(0)} \otimes \boldsymbol{\beta}^{(N)}\right) \\ \vdots & \ddots & \vdots & \vdots \\ O & \dots & O & \left(\Sigma_N^{(s,N)}\right)^{-1}\left(I_{\overline{W}} \otimes \mathbf{S}_0^{(N)} \otimes \boldsymbol{\beta}^{(N)}\right) \end{pmatrix},$$

$$\mathbf{C} = \mathrm{diag}\left\{ I_{\overline{W}M^{(s,n)}} + \left(\Sigma_N^{(s,n)}\right)^{-1}(D_0 \oplus S^{(n)}),\ n = \overline{0,N} \right\};$$

- $$\mathbf{Y}_2 = \begin{pmatrix} O & O \\ O & \mathbf{Y}_2^{(1,1)} \end{pmatrix}$$

  where $\mathbf{Y}_2^{(1,1)}$ *has the form*

  $$\mathbf{Y}_2^{(1,1)} = \mathrm{diag}\left\{O, \mathrm{diag}\left\{\left(\Sigma_N^{(s,n)}\right)^{-1}(D_1 \otimes I_{M^{(s,n)}}),\ n = \overline{0, N}\right\}\right\}.$$

Proof of this lemma is easily implemented by means of direct calculation of the limits.

It can be summarized that

(i) the limits $\mathbf{Y}_0,\ \mathbf{Y}_1$, and $\mathbf{Y}_2$ exist,

(ii) the matrix $\mathbf{Y} = \mathbf{Y}_0 + \mathbf{Y}_1 + \mathbf{Y}_2$ is stochastic,

(iii) all blocks of generator $\mathbf{Q}$ below the sub-diagonal are zero matrices.

All conditions of definition of $AQTMC$ given in [49] are fulfilled and, thus, the Markov chain ξ_t under study belongs to the class of Asymptotically Quasi-Toeplitz Markov Chains. This gives the opportunity to derive ergodicity and non-ergodicity conditions for this Markov chain and compute its steady-state distribution.

3.3 Ergodicity Condition

Theorem 9. *The Markov chain* ξ_t *is ergodic if the inequality*

$$\lambda b_1^{(N)} < N$$

is fulfilled and is non-ergodic if

$$\lambda b_1^{(N)} > N.$$

Here $b_1^{(N)} = \boldsymbol{\beta}^{(N)}(-S^{(N)})^{-1}\mathbf{e}$ is the average duration of service of a group consisting of N customers and λ is the fundamental rate of the MAP .

Proof. It follows from [49] that the Markov chain ξ_t is ergodic if the inequality

$$\mathbf{y}\mathbf{Y}_0\mathbf{e} > \mathbf{y}\mathbf{Y}_2\mathbf{e}$$

holds true, where the vector $\mathbf{y}$ is the unique solution of the system of linear algebraic equations

$$\mathbf{y}\mathbf{Y} = \mathbf{y}, \;\; \mathbf{y}\mathbf{e} = 1 \qquad (3.3.1)$$

and the matrices $\mathbf{Y}_k,\ k = 0, 1, 2$, and $\mathbf{Y}$ were introduced in the previous lemma.
The Markov chain ξ_t is non-ergodic if the opposite inequality holds true.

It can be easily verified that the matrix $\mathbf{Y}$ has the following structure:

$$\mathbf{Y} = \left(\begin{array}{cccc|cccc}
O & I & & & O & O & & \\
& \ddots & \ddots & & \vdots & & \ddots & \\
& & \ddots & I & O & & & \ddots \\
& & & O & \mathbf{A} & & & O \\ \hline
O & & & & O & I & & \\
& \ddots & & & & \ddots & \ddots & \\
& & \ddots & & & & O & I \\
& & & O & \mathbf{B} & & & I+\mathbf{C}'
\end{array}\right)$$

where

$$\mathbf{C}' = \mathrm{diag}\left\{\left(\Sigma_N^{(s,n)}\right)^{-1}(D_0 \oplus S^{(n)} + D_1 \otimes I_{M^{(s,n)}}),\ n = \overline{0, N}\right\}.$$

The system (3.3.1) consists of $\sum\limits_{n=0}^{N-1} K^{(a,n)} + (N+1)K^{(s)} + 1$ equations. In the simplest case, when the arrival flow is described by the stationary Poisson arrival process and admission and service times have an exponential distribution, the number of equations is equal to $2(N+1)$. If admission and service times have phase type distribution with at least two phases, this number is much larger. So, in general, the finite system of linear algebraic equations (3.3.1) has to be solved with help of computer. After computing the vector $\mathbf{y}$ as solution of this system, the vector should be substituted into inequality $\mathbf{y}\mathbf{Y}_0\mathbf{e} > \mathbf{y}\mathbf{Y}_2\mathbf{e}$ to check whether it is fulfilled or not.

However, the matrix $\mathbf{Y}$ is sparse. This helps to solve the system (3.3.1) analytically and get the ergodicity condition in a nice analytic form.

The unknown vector $\mathbf{y}$ can be structured as follows:

$$\mathbf{y} = (\mathbf{y}_0^{(0)}, \mathbf{y}_0^{(1)}, \ldots, \mathbf{y}_0^{(N-1)}, \mathbf{y}_1^{(0)}, \mathbf{y}_1^{(1)}, \ldots, \mathbf{y}_1^{(N)})$$

where the sub-vector $\mathbf{y}_r^{(n)}$ corresponds to the state r, $r = 0, 1$, of the component r_t of the Markov chain ξ_t and the state n of the component n_t, where $n = \overline{0, N-1}$ if $r_t = 0$ and $n = \overline{0, N}$ if $r_t = 1$. These sub-vectors, in turn, are structured as follows: $\mathbf{y}_0^{(n)} = \left(\mathbf{y}_0^{(n,0)}, \mathbf{y}_0^{(n,1)}, \ldots, \mathbf{y}_0^{(n,m)}\right), m = \overline{0, n},\ n = \overline{0, N-1}$, and $\mathbf{y}_1^{(n)} = \left(\mathbf{y}_1^{(n,0)}, \mathbf{y}_1^{(n,1)}, \ldots, \mathbf{y}_1^{(n,N)}\right),\ n = \overline{0, N}$.

By substituting this form of the vector $\mathbf{y}$ to equation $\mathbf{y}\mathbf{Y} = \mathbf{y}$ of system (3.3.1), the following system of linear algebraic equations for the components $\mathbf{y}_r^{(n)}$, $r = 0, 1$, $n = \overline{0, N-1}$ if $r_t = 0$ and $n = \overline{0, N}$ if $r_t = 1$ is got:

$$\mathbf{y}_0^{(0)} = \mathbf{0},\ \mathbf{y}_0^{(k+1)} = \mathbf{y}_0^{(k)},\ k = \overline{0, N-2}, \tag{3.3.2}$$

$$\mathbf{y}_1^{(0)} = \mathbf{y}_0^{(N-1)}\mathbf{A} + \mathbf{y}_1^{(N)}\mathbf{B}, \tag{3.3.3}$$

$$\mathbf{y}_1^{(k+1)} = \mathbf{y}_1^{(k)},\ k = \overline{0, N-2}, \tag{3.3.4}$$

$$\mathbf{y}_1^{(N)} = \mathbf{y}_1^{(N-1)}I + \mathbf{y}_1^{(N)}I + \mathbf{y}_1^{(N)}\mathbf{C}'. \tag{3.3.5}$$

It evidently follows from system (3.3.2) that $\mathbf{y}_0^{(n)} = 0$, $n = \overline{0, N-1}$. From the equation (3.3.3), taking into account sparse structure of the matrix **B**, it can be obtained

$$\mathbf{y}_1^{(0,n)} = \mathbf{0},\ n = \overline{0, N-1},$$
$$\mathbf{y}_1^{(0,N)} = \sum_{m=0}^{N} \mathbf{y}_1^{(N,m)} \left(\Sigma_N^{(s,m)}\right)^{-1} \left(I_{\overline{W}} \otimes \mathbf{S}_0^{(m)} \otimes \boldsymbol{\beta}^{(N)}\right).$$

So, considering the equations (3.3.3), (3.3.4), (3.3.5), it can be attained

$$\mathbf{y}_1^{(0,m)} = \mathbf{y}_1^{(1,m)} = ... = \mathbf{y}_1^{(N-1,m)} = \mathbf{0},\ m = \overline{0, N-1},$$

$$\mathbf{y}_1^{(0,N)} = \mathbf{y}_1^{(1,N)} = ... = \mathbf{y}_1^{(N-1,N)} =$$

$$= \sum_{m=0}^{N} \mathbf{y}_1^{(N,m)} \left(\Sigma_N^{(s,m)}\right)^{-1} \left(I_{\overline{W}} \otimes \mathbf{S}_0^{(m)} \otimes \boldsymbol{\beta}^{(N)}\right). \quad (3.3.6)$$

Taking this into account, the equation (3.3.5) can be rewritten as follows:

$$\mathbf{y}_1^{(N,k)} \left(\Sigma_N^{(s,k)}\right)^{-1} \left(D_0 \oplus S^{(k)} + D_1 \otimes I_{M^{(s,k)}}\right) = \mathbf{0},\ k = \overline{0, N-1}, \quad (3.3.7)$$
$$\mathbf{y}_1^{(N-1,N)} + \mathbf{y}_1^{(N,N)} \left(\Sigma_N^{(s,N)}\right)^{-1} \left(D_0 \oplus S^{(N)} + D_1 \otimes I_{M^{(s,N)}}\right) = \mathbf{0}. \quad (3.3.8)$$

By denoting $D = D_0 + D_1$, the first equation from the system (3.3.7) can be rewritten as:

$$\mathbf{y}_1^{(N,0)} \left(\Sigma_N^{(s,0)}\right)^{-1} \left(D \otimes I + I \otimes S^{(0)}\right) = \mathbf{0}.$$

It can be observed that $D \otimes I + I \otimes S^{(0)}$ is an irreducible subgenerator with strong domination of a diagonal entry at least in one row. So, by theorem of O. Tausski this matrix is non-singular, and therefore these equations have only trivial solution, i.e.,

$$\mathbf{y}_1^{(N,0)} = \mathbf{0}.$$

Analogously these relations are got:

$$\mathbf{y}_1^{(N,m)} = \mathbf{0},\ m = \overline{1, N-1}.$$

From (3.3.8) and (3.3.6) the following equation is obtained:

$$\mathbf{y}_1^{(N,N)} \left(\Sigma_N^{(s,N)}\right)^{-1} \left[\left(I_{\overline{W}} \otimes \mathbf{S}_0^{(N)} \otimes \boldsymbol{\beta}^{(N)}\right) + D \oplus S^{(N)}\right] = \mathbf{0}, \tag{3.3.9}$$

which, introducing denotation $\mathbf{z}_1 = \mathbf{y}_1^{(N,N)} \left(\Sigma_N^{(s,N)}\right)^{-1}$, can be rewritten in the form

$$\mathbf{z}_1 \left[\left(I_{\overline{W}} \otimes \mathbf{S}_0^{(N)} \otimes \boldsymbol{\beta}^{(N)}\right) + D \oplus S^{(N)}\right] = \mathbf{0}. \tag{3.3.10}$$

By direct substitution, it can be verified that solution to the system (3.3.10) has the following simple form:

$$\mathbf{z}_1 = \boldsymbol{\theta} \otimes (\boldsymbol{\beta}^{(N)}(-S^{(N)})^{-1}) \tag{3.3.11}$$

where $\boldsymbol{\theta}$ is the vector of stationary distribution of the underlying process of arrivals.
In turn, ergodicity condition $\mathbf{y}\mathbf{Y}_0\mathbf{e} > \mathbf{y}\mathbf{Y}_2\mathbf{e}$ is transformed to the form

$$\mathbf{z}_1 \left(D_1 \otimes I_{M^{(S,n)}}\right) \mathbf{e} < -N\mathbf{z}_1 D \oplus S^{(N)}\mathbf{e}. \tag{3.3.12}$$

Substituting expression (3.3.11) to system (3.3.12) and using so called mix product rule for Kronecker products of matrices, easily inequality representing ergodicity condition is got, i.e.,

$$\lambda b_1^{(N)} < N.$$

Condition of non-ergodicity is easily proven by analogy.

Remark 10. *For the system under study, when it is overloaded, the average number of customers arriving during the service time is equal to $\lambda b_1^{(N)}$ (here $b_1^{(N)}$ is the average service time of a group of exactly N customers) while the number of customers departing from the system at service completion moment is given by N.*
Thus, an intuitively clear condition of the system ergodicity should be of form $\lambda b_1^{(N)} < N$ what coincides with the strictly already proven condition.

3.4 Key performance indices of the system

Further let the inequality representing the ergodicity condition be fulfilled. Then the stationary distribution of the Markov chain ξ_t exists. Denote the stationary state probabilities of the chain as

$$\boldsymbol{\pi}(i,r,n,m,\nu,\eta) =$$
$$\lim_{t\to\infty} P\{i_t = i,\ r_t = r,\ n_t = n,\ m_t = m,\ \nu_t = \nu,\ \eta_t = \eta\},$$
$$i \geq 0,\ r = 0,1,\ \nu = \overline{0,W},$$

$\eta = \overline{1, M^{(a,m)}}$, $m = \overline{0,n}$, $n = \overline{0,N-1}$, if $r = 0$ and $\eta = \overline{1, M^{(s,m)}}$, $m = \overline{0,N}$, $n = \overline{0,N}$, if $r = 1$.

Let $\boldsymbol{\pi}(i,r,n,m)$ be the row vector of probabilities of the states belonging to the sub-level (i,r,n,m), $\boldsymbol{\pi}(i,r,n)$ be the row vector of probabilities of the states belonging to the level (i,r,n), $\boldsymbol{\pi}(i,r)$ be the row vector of probabilities of the states belonging to the macro-level (i,r) and $\boldsymbol{\pi}_i$ be the row vector of probabilities of the states belonging to the super-level i, $i \geq 0$.

Computation of the vectors $\boldsymbol{\pi}_i$, $i \geqslant 0$, for $AQTMC$ ξ_t can be performed based on numerically stable algorithm presented for the model introduced in the previous chapter.

As soon as the vectors $\boldsymbol{\pi}_i$, $i \geq 0$, have been computed, various performance measures of the system can be calculated:

- Average number of customers in the orbit

$$L_o = \sum_{i=1}^{\infty} i\boldsymbol{\pi}_i\mathbf{e}.$$

- Fraction of time when server has admission period

$$F^{(a)} = \sum_{i=0}^{\infty} \boldsymbol{\pi}(i,0)\mathbf{e}.$$

- Fraction of time when server has service period

$$F^{(s)} = \sum_{i=0}^{\infty} \boldsymbol{\pi}(i,1)\mathbf{e}.$$

- Average number of customers in the pool at an arbitrary moment

$$N_p = \sum_{i=0}^{\infty} \left(\sum_{n=1}^{N-1} n\boldsymbol{\pi}(i,0,n)\mathbf{e} + \sum_{n=1}^{N} n\boldsymbol{\pi}(i,1,n)\mathbf{e} \right).$$

- Average number of customers in the pool at an arbitrary moment conditional that admission period is in a progress

$$N_p^{(a)} = \left(F^{(a)}\right)^{-1} \sum_{i=0}^{\infty} \sum_{n=1}^{N-1} n\boldsymbol{\pi}(i,0,n)\mathbf{e}.$$

- Average number of customers in the pool at an arbitrary moment conditional that service period is in a progress

$$N_p^{(s)} = \left(F^{(s)}\right)^{-1} \sum_{i=0}^{\infty} \sum_{n=1}^{N} n\boldsymbol{\pi}(i,1,n)\mathbf{e}.$$

- The average number of customers in service at an arbitrary moment

$$N_{service} = \sum_{i=0}^{\infty} \sum_{n=0}^{N} \sum_{m=0}^{N} m\boldsymbol{\pi}(i,1,n,m)\mathbf{e}.$$

- Average number of customers in service at an arbitrary moment conditional that service period is in a progress

$$N_s = \left(F^{(s)}\right)^{-1} \sum_{i=0}^{\infty} \sum_{n=0}^{N} \sum_{m=0}^{N} m\boldsymbol{\pi}(i,1,n,m)\mathbf{e}.$$

- The probability that n customers are in the pool at service period completion instant

$$P_{serv\_compl}(n)=\sum_{i=0}^{\infty}\boldsymbol{\pi}(i,1,n)\ diag\left\{I_{\overline{W}}\otimes\mathbf{S}_0^{(m)},m=\overline{0,N}\right\}\mathbf{e}$$

$$\times\left[\sum_{i=0}^{\infty}\sum_{k=0}^{N}\boldsymbol{\pi}(i,1,k)\ diag\left\{I_{\overline{W}}\otimes\mathbf{S}_0^{(m)},m=\overline{0,N}\right\}\mathbf{e}\right]^{-1},$$

$$n=\overline{0,N}.$$

- The probability that n customers are in the pool at admission period beginning instant

$$P_{adm\_begin}(n)=P_{serv\_compl}(n)\left(\sum_{k=0}^{N-1}P_{serv\_compl}(k)\right)^{-1},$$

$$n=\overline{0,N-1}.$$

  Mention that the probabilities that n customers are in the pool at service period completion instant and n customers are staying in the pool at admission period beginning instant do not coincide because sometimes the system can have several service period in turn without intermediate admission periods.

- Probability that n customers are staying in the pool at admission period beginning instant

$$P_{adm\_begin}(n)=P_{serv\_compl}(n)\left(\sum_{k=0}^{N-1}P_{serv\_compl}(k)\right)^{-1}.$$

- Average number of customers in the pool at the moment of starting admission period

$$N_{start}^{(a)}=\sum_{n=0}^{N-1}n\ P_{adm\_begin}(n),\ n=\overline{0,N-1}.$$

- Average number of customers in the pool at service period completion instant

$$L_{serv\_compl} = \sum_{n=1}^{N} n\ P_{serv\_compl}(n).$$

- Probability that an arbitrary admission period finishes due to the end of the phase type distributed admission time

$$P^{PH}_{adm\_end} = \sum_{i=0}^{\infty}\sum_{n=0}^{N-1} \boldsymbol{\pi}(i,0,n)\ diag\left\{I_{\overline{W}} \otimes \mathbf{T}_0^{(m)}, m=\overline{0,n}\right\}\mathbf{e}\times$$

$$\left[\sum_{i=0}^{\infty}\sum_{n=0}^{N-1} \boldsymbol{\pi}(i,0,n)\ diag\left\{I_{\overline{W}} \otimes \mathbf{T}_0^{(m)},\ m=\overline{0,n}\right\}\mathbf{e}+\right.$$

$$\sum_{i=0}^{\infty} \boldsymbol{\pi}(i,0,N-1) diag\left\{D_1 \otimes I_{M^{(a,m)}}, m=\overline{0,N-1}\right\}\mathbf{e}+$$

$$\left.\sum_{i=0}^{\infty} \alpha_i \boldsymbol{\pi}(i,0,N-1) diag\left\{I_{\overline{W}} \otimes I_{M^{(a,m)}}, m=\overline{0,N-1}\right\}\mathbf{e}\right]^{-1}.$$

- The probability that an arbitrary admission period finishes because the pool becomes full:

$$P^{(full)}_{adm\_end} = 1 - P^{PH}_{adm\_end} =$$

$$= \sum_{i=0}^{\infty} \boldsymbol{\pi}(i,0,N-1)\ \Big[\ diag\left\{D_1 \otimes I_{M^{(a,m)}}, m=\overline{0,N-1}\right\}\mathbf{e}+$$

$$\alpha_i\ diag\left\{I_{\overline{W}M^{(a,m)}}, m=\overline{0,N-1}\right\}\mathbf{e}\Big]\times$$

$$\left[\sum_{i=0}^{\infty}\sum_{n=0}^{N-1} \boldsymbol{\pi}(i,0,n)\ diag\left\{I_{\overline{W}} \otimes \mathbf{T}_0^{(m)},\ m=\overline{0,n}\right\}\mathbf{e}+\right.$$

$$\sum_{i=0}^{\infty} \boldsymbol{\pi}(i,0,N-1) diag\left\{D_1 \otimes I_{M^{(a,m)}}, m=\overline{0,N-1}\right\}\mathbf{e}+$$

$$\left.\sum_{i=0}^{\infty}\alpha_i\boldsymbol{\pi}(i,0,N-1)diag\left\{I_{\overline{W}}\otimes I_{M^{(a,m)}}, m=\overline{0,N-1}\right\}\mathbf{e}\right]^{-1}.$$

This probability can be split into the probability that an arbitrary admission period finishes by the pool fulling out due to a primary customer arrival and the probability that an arbitrary admission period finishes by the pool fulling out due to a customer arrival from the orbit. Formula for the first probability has the form of a fraction with the same denominator as in formula for $P_{adm\_end}^{(full)}$ and the numerator $\sum\limits_{i=0}^{\infty}\boldsymbol{\pi}(i,0,N-1)\ diag\left\{D_1\otimes I_{M^{(a,m)}}, m=\overline{0,N-1}\right\}\mathbf{e}$. Formula for the second probability has a similar form with numerator $\sum\limits_{i=0}^{\infty}\alpha_i\boldsymbol{\pi}(i,0,N-1)\mathbf{e}_{K^{(a,N-1)}}$.

- Probability that an arbitrary customer arrives during the admission period and is placed to the pool:

$$P_{pool}^{(a)}=\lambda^{-1}\sum_{i=0}^{\infty}\sum_{n=0}^{N-2}\boldsymbol{\pi}(i,0,n)diag\{D_1\otimes I_{M^{(a,m)}}, m=\overline{0,n}\}\mathbf{e}.$$

- Probability that an arbitrary customer immediately starts the service

$$P_{imm}^{(a)}=\lambda^{-1}\sum_{i=0}^{\infty}\boldsymbol{\pi}(i,0,N-1)diag\{D_1\otimes I_{M^{(a,m)}}, m=\overline{0,N-1}\}\mathbf{e}.$$

- Probability that an arbitrary customer arrives during the service period and is placed to the pool:

$$P_{pool}^{(s)}=\lambda^{-1}\sum_{i=0}^{\infty}\sum_{n=0}^{N-1}\boldsymbol{\pi}(i,1,n)diag\{D_1\otimes I_{M^{(s,m)}}, m=\overline{0,N}\}\mathbf{e}.$$

- Probability that an arbitrary customer arrives during the service period and moves to the orbit:

$$P_{orbit}^{(s)}=\lambda^{-1}\sum_{i=0}^{\infty}\boldsymbol{\pi}(i,1,N)diag\{D_1\otimes I_{M^{(s,m)}}, m=\overline{0,N}\}\mathbf{e}.$$

- Probability that an arbitrary customer will not visit the orbit:
$$P_{pool} = 1 - P^{(s)}_{orbit}.$$
- Average number of customers in the system:
$$L = L_0 + \sum_{i=0}^{\infty} \left[\sum_{n=0}^{N-1} n\boldsymbol{\pi}(i,0,n)\mathbf{e} + \sum_{n=0}^{N} \sum_{m=0}^{N} (m+n)\boldsymbol{\pi}(i,1,n,m)\mathbf{e} \right].$$
- Second row moment of the distribution of the number of customers in the system (in the orbit, pool and service):
$$L^{(2)} = \sum_{i=0}^{\infty} \left[\sum_{n=0}^{N-1} (i+n)^2 \boldsymbol{\pi}(i,0,n)\mathbf{e} + \sum_{n=0}^{N} \sum_{m=0}^{N} (i+m+n)^2 \boldsymbol{\pi}(i,1,n,m)\mathbf{e} \right].$$
- The variance of the number of customers in the system:
$$V = L^{(2)} - (L)^2.$$
- Average number of customers in the pool at arbitrary moment:
$$N_p = \sum_{i=0}^{\infty} \left(\sum_{n=1}^{N-1} n\boldsymbol{\pi}(i,0,n)\mathbf{e} + \sum_{n=1}^{N} n\boldsymbol{\pi}(i,1,n)\mathbf{e} \right).$$

3.5 Numerical results

It has already been noted above that the proposed discipline provides higher throughput of the system. In this section the aim is to numerically show that the current strategy provides smaller average number of customers in the system and higher probability that an arbitrary customer will get service without visiting the orbit, in comparison to a classical admission strategy, strategy presented in [31], and strategy analysed before for the $MAP/PH/1$ retrial system with the same arrival process, service process of a single customer and retrial intensity.

In order to do this, the results of three numerical experiments are presented below, for different load of the system and capacity of the pool.
The way to choose the service time distribution of a group consisting of a fixed number of customers has already been discussed in Chapter 2. Importance of correlation in arrival process also will be demonstrated via consideration of two $MAPs$ having the same fundamental rate λ but different correlation of successive inter-arrival times (MAP_0 and $MAP_{0.2}$, defined in Chapter 2, will be considered).
As the main performance measures of the system under study, the following are considered:

- Probability P_{pool} of providing service to an arbitrary customer without visiting the orbit,

- Average number L of customers in the system at an arbitrary time instant.

In the first experiment, $\alpha = 0.6$ has been fixed as individual intensity of retrials, μ as the parameter of exponential distribution of admission period, $0 < \mu \leqslant 25,$ distribution of service time of individual customer as Erlangian distribution of order 2 with mean value equal to 1. For fair comparison of the proposed admission discipline with the classical one, it is assumed that service time of the empty pool is exponentially distributed with huge rate (the rate is set equal to 10 000).
The dependencies of P_{pool} and L are compared on intensity μ for $N = 3$ and different admission strategies: classical strategy, strategy proposed in [31] (marked on figures as 'Pooling'), strategy analysed in the previous chapter (marked on figures as 'Adaptive

Pooling') and strategy proposed in this chapter (marked on figures as 'Permanent Pooling') with different correlations in arrival process.

Figure 3.1 and 3.2 show dependencies of P_{pool} and L on μ for MAP_0 and $MAP_{0.2}$, correspondingly.

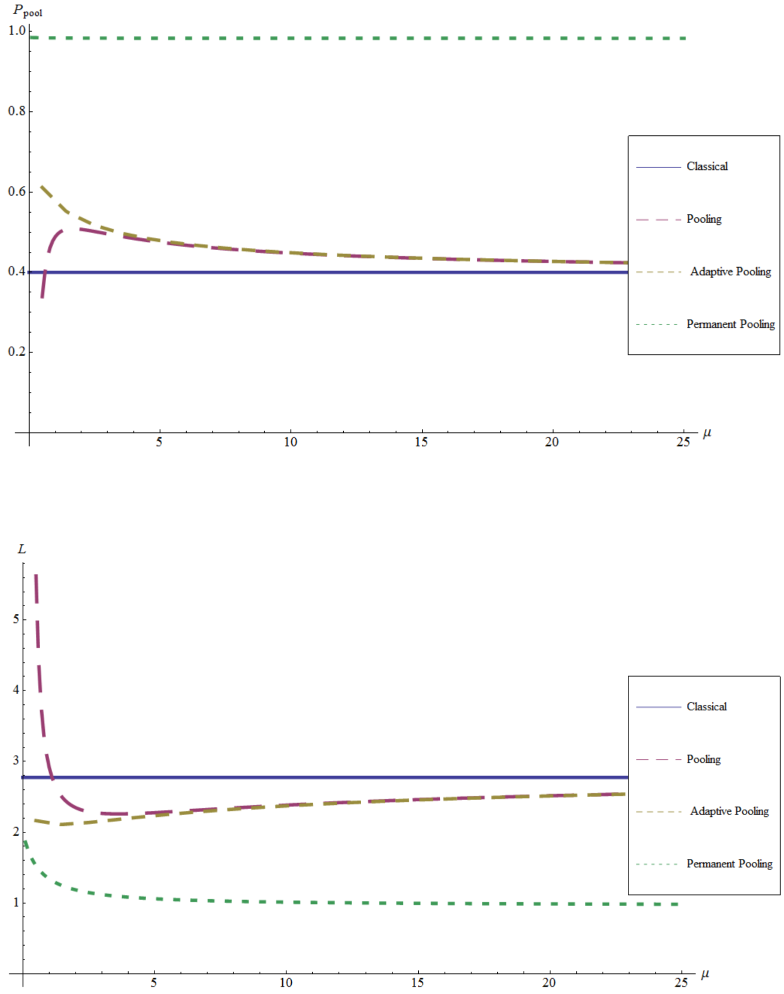


Figure 3.1 P_{pool} and L for different values of admission rate μ, different strategies of customers admission and correlation 0 in arrival process

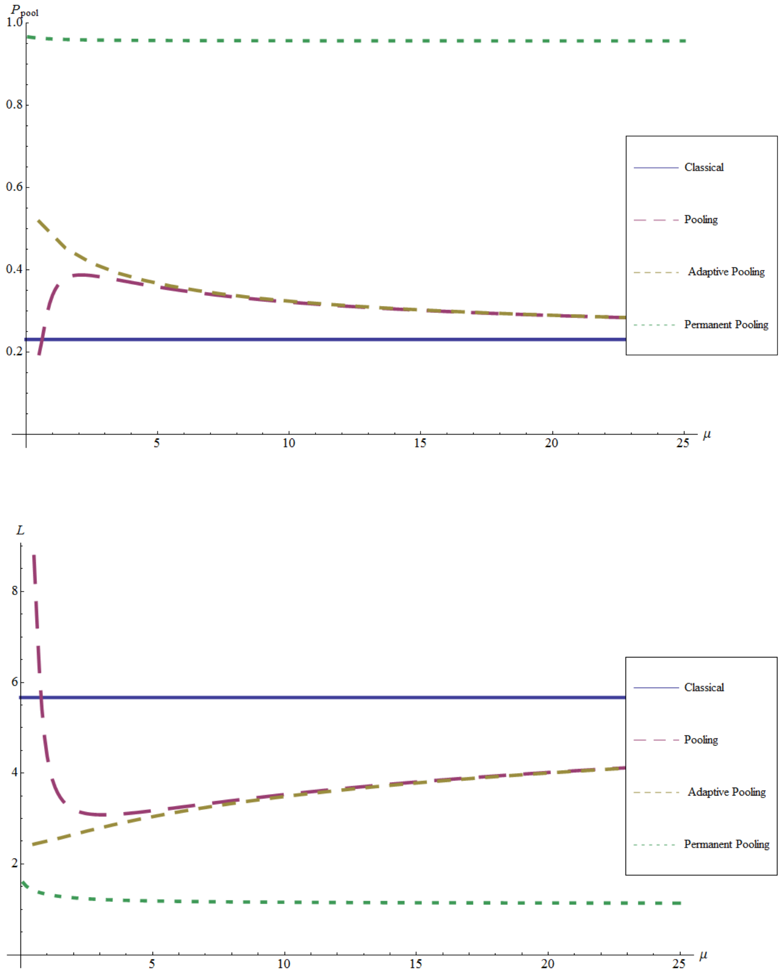


Figure 3.2 P_{pool} and L for different values of admission rate μ, different strategies of customers admission and correlation 0.2 in arrival process

Based on these figures, the following conclusions can be made:

- Although both the Adaptive Pooling strategy and the Pooling strategy of access give pretty good values of the chosen performance indices compared to those obtained with the classical strategy, these figures reveal that the new strategy of access with a Permanent Pooling, introduced in this chapter, gives much better value of the chosen performance indices for all considered values of correlation in arrival process.

- It is seen from Figure 3.1 that for the Permanent Pooling, the highest value of probability P_{pool} is achieved for very long admission period (small values of μ) and then it slightly decreases with growth of μ. For $\mu = 0.1$, $P_{pool} = 0.985235$ and for $\mu = 25$, $P_{pool} = 0.983295$. For classical strategy, $P_{pool} = 0.4$. For Pooling strategy, the highest value of probability P_{pool} is achieved for a finite value of $\mu, \mu = 1.7$, and its value is $P_{pool} = 0.5087$, while for Adaptive Pooling strategy, the highest value of probability P_{pool} is achieved for a finite value of $\mu, \mu = 0.1$, and its value is $P_{pool} = 0.6311$.
  Advantage of the strategy here proposed consists of small sensitivity of achieved performance measures of the system with respect to duration of admission period. This allows to almost avoid making the optimal choice of duration of admission period. The optimal values of L are achieved at different values of μ. The smallest value of $L = 2.2594$ in case of Pooling strategy is achieved when $\mu = 3.6$. For the Adaptive Pooling strategy the smallest value of $L = 2.110234$ is attained in correspondence of $\mu = 1.5$. In case of the Permanent Pooling, the value of L decreases with increasing of μ and $L = 0.981812$ when $\mu = 25$. Classical strategy provides value $L = 2.7747$ for all μ.

- Figure 3.2 reveals that for the Permanent Pooling the highest value of probability P_{pool} again is achieved for very long admission period and is equal to 0.966948 for $\mu = 0.1$. Also for Adaptive Pooling strategy the highest value of P_{pool} is achieved for $\mu = 0.1$ and is equal to 0.5577. For Pooling strategy, the highest value of probability P_{pool} is achieved for a finite value of $\mu, \mu = 2.1$, and its value is $P_{pool} = 0.3878$. For classical strategy, $P_{pool} = 0.23075$. The optimal values of L are achieved at different values of μ. The smallest value of $L = 3.0827$ in case of the Pooling strategy is achieved when $\mu = 3.1$. For the Adaptive Pooling strategy, the value of L is lower when $\mu = 0.1$ and it is equal to 2.40541. In case of the Permanent Pooling, the value of L decreases with increasing of μ. For $\mu = 25$, $L = 1.13708$. Classical strategy provides value $L = 5.665$ for all μ.

- Comparing Figures 3.1 and 3.2, one can evidently see that the key performance measures of the system as well as the optimal values of the duration of admission periods significantly depend on correlation in the arrival process. This explains the necessity of consideration of Markovian Arrival Process (that may have coefficient of correlation from -1 till 1) instead of the stationary Poisson arrival process (having zero correlation), if the considered queueing system is applied for analysis of real world system in which arrival process exhibits correlation.

In the second experiment, it is fixed as arrival flow the $MAP_{0.2}$ scaled in such a way that the mean arrival rate is 0.9, not 0.6 as it was in the previous example. This is easily achieved by multiplying all entries of the matrices D_0 and D_1 by the factor 1.5.

Figure 3.3 illustrates the effect of varying the intensity μ. The line corresponding to the classical strategy on the picture for L is not presented because L is very large ($L = 94.75$) for all values of μ.

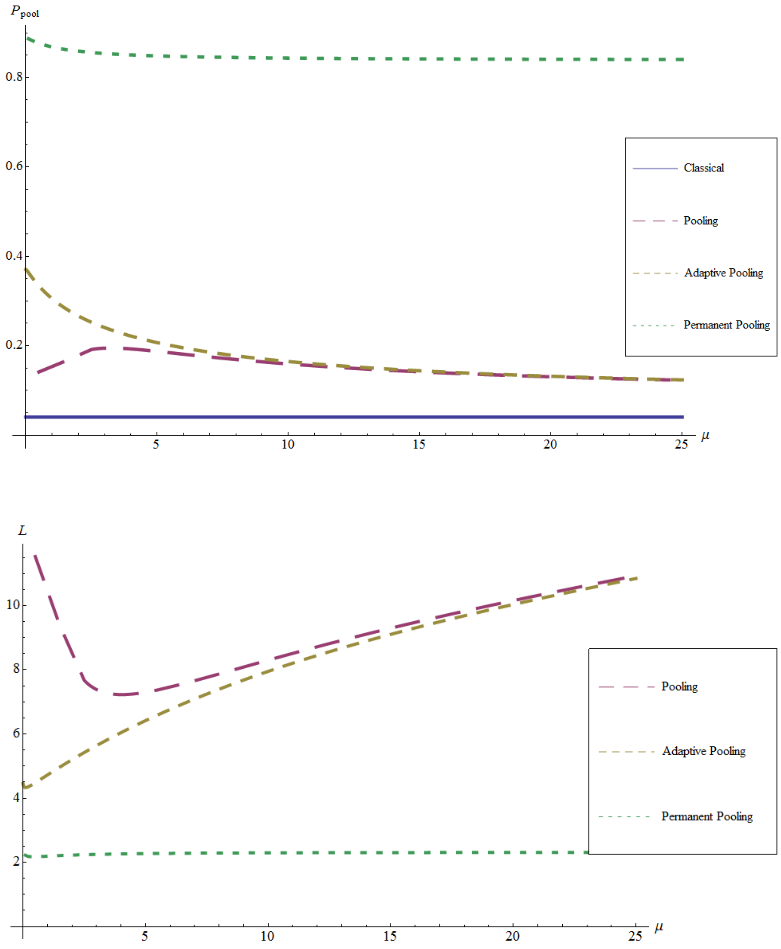


Figure 3.3 P_{pool} and L for different values of admission rate μ, and different strategies of customers admission for $\lambda = 0.9$

Comparing this figure with Figure 3.2, it can be seen that in case of higher load of the system the profit gained by application of the proposed strategy of customers admission becomes more essential.

In the third experiment, again $MAP_{0.2}$ is fixed as arrival flow with the mean arrival rate 0.6 and only admission strategy analyzed in this chapter is valuated for capacity of the pool $N = 2$, $N = 3$ and $N = 4$. The respective values for the classical strategy (which corresponds to $N = 1$) do not depend on the value of μ and are equal to $P_{pool} = 0.23$ and $L = 5.665$ what is greatly worse then in the case of Permanent Pooling strategy. Figure 3.5 illustrates the effect of varying the intensity μ for various values of capacity N of the pool.

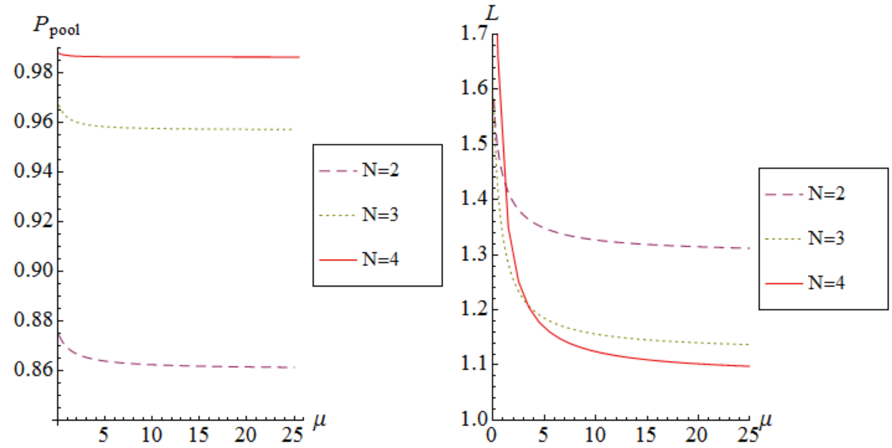


Figure 3.4 P_{pool} and L for different values of admission rate μ, and different N

It is clear from this figure that the increase of capacity N of the pool implies better performance of the system under the properly chosen value of μ. This can be easily explained by the numerical results in [31] which show that the increase of the number of simultaneously served customers implies, especially in case of not very large variance of service times, much smaller value of average time of service per one customer. However, when μ is chosen too small (admission period is pretty long), the value of the average number of customers in the system L can be larger for larger pool capacity N.

Chapter 4

A MAP/PH/1 System with Flexible Group Service

A novel customers batch service discipline for a single server queue will be introduced and analysed in this chapter. Service to customers is offered in batches of a certain size, formed during both the admission and the service period in the so called pool, and enqueued in a buffer. Customers arrive according to a Markovian Arrival Process, the individual customer's service time has phase type distribution and the service time of a batch is defined as the maximum of individual service times of customers which form a batch.

The overwhelming majority of queueing literature is devoted to queues where service to customers is provided one by one. However, queues with batch (bulk, group) service also got their portion of attention. In such queues, service is provided not to an individual customer, but to groups of customers. Usually, the minimal and the maximal size of a group are predefined. Some examples of such type of systems and their applications have been shown in the first chapter.

It has also been stressed that, among all the group service disciplines studied until now, rarely the researchers have put their attention on systems in which the arrivals occur according to correlated process, with the services dependent on the size of the batch being served, and the service times non-exponential. For this reason, the model that has been considered in this work has these characteristics.

In particular, the considered model has two distinguishing features:

- It is usually assumed in analysis of the group service queueing systems that some integer threshold, say N, is fixed and the service does not start if the number of customers in the queue is less than N. Here, instead, it is assumed that the idle time of the server is limited and if this time expires then the service starts, even if the number of customers in the queue is less than N. This assumption better suits to certain real world situations. E.g., in modelling operation of passengers delivering system in airport, the shuttle has to start travel even if it is not completely loaded because: (i) the passenger can be late to his/her flight due to long waiting and (ii) there is some schedule and the next shuttle should arrive for loading.

- Although the majority of the obtained analytical results are valid for an arbitrary dependence of a batch service time on the number of customers in the batch being served, in the numerical examples, which will be shown in the next sections, the service time of a batch is defined as the maximum of individual service times of customers which form a batch. This assumption comes from the model of so called multi-rate information transmission that is supposed, e.g., in IEEE802.11 WLAN.

4.1 The mathematical model

A single server queueing system, in which the input flow is described by a Markovian Arrival Process, is considered. Customer's arrival in the MAP is directed by an underlying irreducible continuous time Markov chain $\nu_t,\ t \geq 0$, with the finite state space $\{0, ..., W\}$. Sojourn time of the Markov chain ν_t, $t \geq 0$, in the state ν has exponential distribution with parameter $\lambda_\nu, \nu \in \{0, ..., W\}$. After this sojourn time expires, with probability $p_k(\nu, \nu')$, the process ν_t jumps to the state ν', and k customers, $k = 0, 1$, arrive into the system. The intensities of jumps of underlying Markov chain from one state into another with generation of k customers are combined into the matrices $D_k,\ k = 0, 1$, of size $(W+1)\times(W+1)$. The matrix $D(1) = D_0 + D_1$ is the infinitesimal generator of the process $\nu_t,\ t \geq 0$. The invariant probability vector (vector of stationary distribution) $\boldsymbol{\theta}$ of this process is computed as the unique solution to the equations

$$\boldsymbol{\theta} D(1) = \mathbf{0},\ \boldsymbol{\theta}\mathbf{e} = 1.$$

The average intensity λ (fundamental rate) of the MAP is defined as $\lambda = \boldsymbol{\theta} D_1 \mathbf{e}$ and gives the expected number of arrivals per unit of time in the stationary mode. The variance v of intervals between customer arrivals is calculated as

$$v = 2\lambda^{-1}\boldsymbol{\theta}(-D_0)^{-1}\mathbf{e} - \lambda^{-2},$$

the squared coefficient c_{var} of variation is equal to

$$2\lambda\boldsymbol{\theta}(-D_0)^{-1}\mathbf{e} - 1,$$

while the coefficient c_{cor} of correlation of successive intervals between arrivals is given by

$$c_{cor} = (\lambda^{-1}\boldsymbol{\theta}(-D_0)^{-1}D_1(-D_0)^{-1}\mathbf{e} - \lambda^{-2})/v.$$

It is assumed that basically the customers have to receive the service in batches of size N, where N is a certain fixed in advance integer. Presented below consideration assumes by default that $N \geqslant 2$. However, results for $N = 1$ are easily obtained from the given formulas as well. Note that $N = 1$ corresponds to the usual service of the customers one by one.

Due to the batch service, an arriving customer has a chance to start service immediately upon arrival only if it arrives at the moment when the server is idle and there are $N-1$ customers in the queue. Arrival of such a customer triggers the start of service of a batch containing N customers. If the customer arrives when the server is busy or it is idle and the number of customers in the queue is less than $N-1$, then the customer is placed to the buffer. Capacity of the buffer is infinite. Customers in the buffer are placed in the order of their arrival. The discipline of customers selection from the buffer at the service completion moment is defined as follows. If during this moment at least N customers are staying in the buffer, the batch consisting of exactly N customers starts service. Such a batch is called as a *block*. If the number of customers in the buffer at the service completion moment is less than N, the set of such customers is called as a pool. At this moment, the so called admission period starts. The server resumes the service when the number of the customers in the pool reaches the value N (in this case customers in the pool form a block and the pool becomes empty) or the admission period expires. In the latter case, if the queue is not empty all customers from the pool start service simultaneously. If the queue is empty, a new admission period starts. Therefore, the server can simultaneously provide service to a batch of N customers (block) if the block was present in the buffer at the service completion epoch or is accumulated there during the admission period, or to the batch of n customers, $n=\overline{1,N-1}$, if the admission period expired when the number of customers in the pool was equal to n.

The structure of the queueing system under study is presented in Figure 4.1.

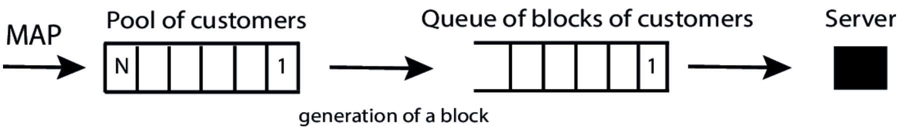


Figure 4.1 Structure of the queueing system.

Duration of admission period has PH type distribution with irreducible representation $(\boldsymbol{\tau}, T)$. This means that it is governed by the underlying process $\eta_t^{(0)},\ t \geqslant 0$, which is a continuous time Markov chain with state space $\{1, \dots, M^{(0)}, M^{(0)} + 1\}$. The initial state of the process $\eta_t^{(0)},\ t \geqslant 0$, at the epoch of starting the admission period is determined in the set $\{1, \dots, M^{(0)}\}$ of transient states by the probabilistic row-vector $\boldsymbol{\tau} = (\tau_1, \dots, \tau_{M^{(0)}})$. The transitions of the process $\eta_t^{(0)},\ t \geqslant 0$, within the set $\{1, \dots, M^{(0)}\}$ do not lead to admission period completion and their intensities are defined by the sub-generator T of size $M^{(0)} \times M^{(0)}$. The intensities of transition to the absorbing state $M^{(0)} + 1$, which lead to admission period completion, are defined by the vector $\mathbf{T}_0 = -T\mathbf{e}$. The admission period time distribution function has the form $T(x) = 1 - \boldsymbol{\tau} e^{Tx}\mathbf{e}$. Laplace-Stieltjes transform $\int\limits_0^\infty e^{-sx} dT(x)$ of this distribution function is $\boldsymbol{\tau}(sI - T)^{-1}\mathbf{T}_0$. The average length of admission period is given by

$$r_1 = \boldsymbol{\tau}(-T)^{-1}\mathbf{e}.$$

Further the value $\mu = r_1^{-1}$ will be called as the intensity of admission. The matrix $T + \mathbf{T}_0\boldsymbol{\tau}$ is assumed to be irreducible.

Duration of simultaneous service of $n,\ n = \overline{1, N}$, customers has PH type distribution with irreducible representation $(\boldsymbol{\beta}^{(n)}, S^{(n)})$. Underlying process of this distribution is $\eta_t^{(n)},\ t \geqslant 0$, with a finite state space of transient states $\{1, \dots, M^{(n)}\}$. The average service time of a group of n customers is defined by

$$b_1^{(n)} = \boldsymbol{\beta}^{(n)}(-S^{(n)})^{-1}\mathbf{e},\ n = \overline{1, N}.$$

The problem of fitting the measurements of arrival and service processes in real world systems with a Markovian arrival process and a PH distribution can be solved by analogy with [17, 55].

The aim of further analysis is to evaluate impact of the value of the threshold N and intensity of admission μ on the system performance.

4.2 The process of the system states

It can be seen that the dynamics of the system under study are completely described by the multi-dimensional process

$$\xi_t = \{i_t, m_t, r_t, \eta_t^{(r_t)}, \nu_t\},\ t \geq 0,$$

where:

- i_t is the number of batches of customers in the system, $i_t \geq 0$. The number i_t includes one batch in service, if any, and $i_t - 1$ blocks in the queue, if any;
- m_t is the number of customers in the pool, $m_t = \overline{0, N-1}$;
- r_t is the number of customers in service: $r_t = 0$ if $i_t = 0$ and, consequently, admission period is in a progress, and $r_t = \overline{1, N}$ if $i_t \geq 1$;
- η_t is the state of the underlying process of the PH process of customers admission, $\eta_t^{(0)} = \overline{1, M^{(0)}}$, or the state of the underlying process of the PH process of customers service, $\eta_t^{(r_t)} = \overline{1, M^{(r_t)}}$, $r_t = \overline{1, N}$;
- ν_t is the state of the underlying process of the MAP, $\nu_t = \overline{0, W}$.

Note that the total number of customers in the system at an arbitrary moment t is equal to $r_t + (i_t - 1)N + m_t$ if $i_t \geqslant 1$ or to m_t if $i_t = 0$.

Given all above assumptions, the five-dimensional process ξ_t is an irreducible continuous time Markov chain with one component (i_t) having infinite state space and four finite components. Its state space is defined by

$$\left\{ \left(0, m, 0, \eta^{(0)}, \nu\right), \eta^{(0)} = \overline{1, M^{(0)}} \right\} \bigcup$$
$$\left\{ \left(i, m, r, \eta^{(r)}, \nu\right), i \geq 1, r = \overline{1, N}, \eta^{(r)} = \overline{1, M^{(r)}} \right\},$$
$$m = \overline{0, N-1}, \nu = \overline{0, W}.$$

To analyse behavior and properties of the Markov chain ξ_t, the infinitesimal generator of the chain has to be computed. Let this generator be denoted as $\mathbf{Q}$. The diagonal entries $\mathbf{Q}_{(i,m,r,\eta,\nu),(i,m,r,\eta,\nu)}$ are negative. Modulus of each diagonal entry defines the intensity of departure of the Markov chain from the corresponding state of the Markov chain. The non-diagonal entry $\mathbf{Q}_{(i,m,r,\eta,\nu),(i',m',r',\eta',\nu')}$ is non-negative and defines the intensity of transition of the Markov chain from the state (i,m,r,η,ν) to the state (i',m',r',η',ν').

To simplify the structure of generator $\mathbf{Q}$ and follow traditional methodology of analysis of multi-dimensional Markov chains, it is convenient to make a lexicographic enumeration of the states of the Markov chain ξ_t and compose all the states of the chain having value (i,m,r) as the first three components to the *macro-states* $(0,m,0), m=\overline{0,N-1}$, and $(i,m,r), m=\overline{0,N-1}, r=\overline{1,N}$. The macro-state $(0,m,0)$, $m=\overline{0,N-1}$, contains

$$K^{(0)} = NM^{(0)}\,(W+1)$$

states and the macro-state $(i,m,r), m=\overline{0,N-1}, r=\overline{1,N}$, consists of

$$K^{(r)} = NM^{(r)}\,(W+1)$$

states. Analogously, the macro-states (i,m,r) will be composed to *extra-states* $(0,m)\equiv(0,m,0)\,,(i,m)\equiv((i,m,1)\,,...,(i,m,N))$, $i\geqslant 1$, and then *super-states* 0 will be formed as a composition of extra-states $(0,m)$, $m=\overline{0,N-1}$, and i as a composition of extra-states (i,m), $m=\overline{0,N-1}$, $i\geq 1$.

The following lemma holds.

Lemma 11. *Generator* $\mathbf{Q}=(\mathbf{Q}_{i,j})\,,\ i\geq 0,\ \max\{0,i-1\}\leq j\leq i+1$, *has a three block diagonal structure:*

$$\mathbf{Q}=\begin{pmatrix} \mathbf{Q}_{0,0} & \mathbf{Q}_{0,1} & O & O & \cdots \\ \mathbf{Q}_{1,0} & \mathbf{Q}_{1,1} & \mathbf{Q}_{1,2} & O & \cdots \\ O & \mathbf{Q}_{2,1} & \mathbf{Q}_{2,2} & \mathbf{Q}_{2,3} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \\ \end{pmatrix}$$

where non-zero blocks $\mathbf{Q}_{i,j}$ *defining the intensities of transitions from super-state* i *to super-state* j, $j = \max\{0, i-1\}, i, i+1,$ *are defined as follows:*

- $\mathbf{Q}_{0,0}$ *is a two block diagonal matrix defined by*

$$\mathbf{Q}_{0,0} =$$
$$I_N \otimes (\mathbf{T} \oplus D_0) + \widehat{I}_N \otimes (\mathbf{T}_0 \tau \otimes I_{\overline{W}}) + \mathbf{E}_N^{+} \otimes (I_{M^{(0)}} \otimes D_1)\,;$$

- $\mathbf{Q}_{0,1}$ *is the block matrix having the structure presented below:*

$$\mathbf{Q}_{0,1} = \begin{pmatrix} O & O & \cdots & O \\ (\mathbf{Q}_{0,1})_{1,0} & & & \\ (\mathbf{Q}_{0,1})_{2,0} & \vdots & \ddots & \vdots \\ \vdots & & & \\ (\mathbf{Q}_{0,1})_{N-1,0} & O & \cdots & O \\ & & & \end{pmatrix},$$

where

$$(\mathbf{Q}_{0,1})_{m,0} = \left(\underbrace{O, \ldots, O}_{m-1}, \mathbf{T}_0 \otimes \boldsymbol{\beta}^{(m)} \otimes I_{\overline{W}}, \underbrace{O, \ldots, O}_{N-m} \right),$$
$$m = \overline{1, N-2},$$

and

$$(\mathbf{Q}_{0,1})_{N-1,0} = \left(\underbrace{O, \ldots, O}_{N-2}, \mathbf{T}_0 \otimes \boldsymbol{\beta}^{(N-1)} \otimes I_{\overline{W}}, \mathbf{e}_{M^{(0)}} \otimes \boldsymbol{\beta}^{(N)} \otimes D_1 \right);$$

- $\mathbf{Q}_{i,i}, i \geq 1$ *is a two block diagonal matrix with the diagonal blocks defined by*

$$(\mathbf{Q}_{i,i})_{m,m} = diag\left\{\mathbf{S}^{(r)} \oplus D_0,\ r = \overline{1, N}\right\}, m = \overline{0, N-1},$$

and the up-diagonal blocks defined by

$$(\mathbf{Q}_{i,i})_{m,m+1} = diag\left\{I_{M^{(r)}} \otimes D_1,\ r = \overline{1, N}\right\}, m = \overline{0, N-2};$$

- $\mathbf{Q}_{i,i+1}$ *is the matrix defined by*

$$\mathbf{Q}_{i,i+1} = \begin{pmatrix} O & O & \dots & \dots & O \\ \vdots & \vdots & \ddots & & \vdots \\ O & \vdots & & \ddots & \vdots \\ (\mathbf{Q}_{i,i+1})_{N-1,0} & O & \dots & \dots & O \end{pmatrix}$$

where $(\mathbf{Q}_{i,i+1})_{N-1,0}$ *is the matrix of the form:*

$$(\mathbf{Q}_{i,i+1})_{N-1,0} = diag\left\{I_{M^{(r)}} \otimes D_1,\ r = \overline{1,N}\right\};$$

- $\mathbf{Q}_{i,i-1}$ *is the matrix defined by*

$$\mathbf{Q}_{i,i-1} = diag\{(\mathbf{Q}_{i,i-1})_{m,m}, m = \overline{0,N-1}\},$$

with

$$(\mathbf{Q}_{i,i-1})_{m,m} = \begin{pmatrix} O & \dots & O & \mathbf{S}_0^{(1)} \otimes \boldsymbol{\beta}^{(N)} \otimes I_{\overline{W}} \\ \vdots & \ddots & \vdots & \vdots \\ O & \dots & O & \mathbf{S}_0^{(N)} \otimes \boldsymbol{\beta}^{(N)} \otimes I_{\overline{W}} \end{pmatrix};$$

- $\mathbf{Q}_{1,0}$ *is the square matrix having N block rows and block columns of the form:*

$$\mathbf{Q}_{1,0} = diag\{(\mathbf{Q}_{1,0})_{m,m}, m = \overline{0,N-1}\},$$

with

$$(\mathbf{Q}_{1,0})_{m,m} = \begin{pmatrix} \mathbf{S}_0^{(1)} \otimes \boldsymbol{\tau} \otimes I_{\overline{W}} \\ \vdots \\ \mathbf{S}_0^{(N)} \otimes \boldsymbol{\tau} \otimes I_{\overline{W}} \end{pmatrix}.$$

It is easy to see that for $i \geq 1$, the expressions of the blocks $\mathbf{Q}_{i,i}$, $\mathbf{Q}_{i,i-1}$ and $\mathbf{Q}_{i,i+1}$ do not depend on i, what means that the Markov chain ξ_t belongs to the well known class of Quasi-Birth-and-Death processes, see [58].

Let this block be denoted as $\mathbf{Q}_{i,i} = \mathbf{Q}^0$, $\mathbf{Q}_{i,i-1} = \mathbf{Q}^-$ and $\mathbf{Q}_{i,i+1} = \mathbf{Q}^+$. The structure of generator is the following:

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_{0,0} & \mathbf{Q}_{0,1} & O & O & \dots \\ \mathbf{Q}_{1,0} & \mathbf{Q}^0 & \mathbf{Q}^+ & O & \dots \\ O & \mathbf{Q}^- & \mathbf{Q}^0 & \mathbf{Q}^+ & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

In order to compute the mean performance indices, an ergodicity condition has to be defined and verified.

4.3 Ergodicity Condition

The ergodicity condition is stated in the following theorem.

Theorem 12. *The considered Markov chain ξ_t is ergodic if the inequality*

$$\lambda b_1^{(N)} < N \tag{4.3.1}$$

is fulfilled and is non-ergodic if

$$\lambda b_1^{(N)} > N.$$

Here $b_1^{(N)} = \boldsymbol{\beta}^{(N)}(-S^{(N)})^{-1}\mathbf{e}$ is the average duration of service of a batch consisting of N customers and λ is the fundamental rate of the MAP.

Proof. It follows from [58] that the criterion of ergodicity of the Markov chain ξ_t is the fulfillment of inequality

$$\mathbf{y}Q^-\mathbf{e} > \mathbf{y}Q^+\mathbf{e}, \tag{4.3.2}$$

where the vector $\mathbf{y}$ is the unique solution of the system of linear algebraic equations

$$\mathbf{y}\left(Q^- + Q^0 + Q^+\right) = \mathbf{0},\ \mathbf{y}\mathbf{e} = 1. \tag{4.3.3}$$

It is easy to check that the matrix

$$\mathbf{V} = Q^{-} + Q^{0} + Q^{+}$$

has the following structure

$$\mathbf{V} = \begin{pmatrix} A & A' & & \\ & \ddots & \ddots & \\ & & \ddots & A' \\ A' & & & A \end{pmatrix},$$

where

$$A=\begin{pmatrix} S^{(1)} \oplus D_0 & & \mathbf{S}_0^{(1)} \otimes \boldsymbol{\beta}^{(N)} \otimes I_{\overline{W}} \\ & \ddots & \vdots \\ & & \ddots\, \mathbf{S}_0^{(N-1)} \otimes \boldsymbol{\beta}^{(N)} \otimes I_{\overline{W}} \\ & & S^{(N)} \oplus D_0 + \mathbf{S}_0^{(N)} \otimes \boldsymbol{\beta}^{(N)} \otimes I_{\overline{W}} \end{pmatrix},$$

$$A'=\begin{pmatrix} I_{M^{(1)}} \otimes D_1 & & \\ & \ddots & \\ & & I_{M^{(N)}} \otimes D_1 \end{pmatrix}.$$

The solution to the system (4.3.3) can be found rewriting the system in the form

$$\mathbf{yV} = \mathbf{0}, \quad \mathbf{ye} = 1. \tag{4.3.4}$$

It is clear that the vector $\mathbf{y}$ has the following structure:

$$\mathbf{y} = (\mathbf{y}_0, \mathbf{y}_1, \dots, \mathbf{y}_{N-1}),$$

where $\mathbf{y}_m = (\mathbf{y}_m^{(1)}, \dots, \mathbf{y}_m^{(N)}), m = \overline{0, N-1}$.

By direct substitution of this form of the vector $\mathbf{y}$ to system (4.3.4), it can be verified that the vectors $\mathbf{y}_m,\ m = \overline{0, N-1}$, have the following form:

$$\mathbf{y}_m = (\mathbf{0}, \dots, \mathbf{0}, \mathbf{y}_m^{(N)}),$$

where the vectors $\mathbf{y}_m^{(N)},\ m = \overline{0, N-1}$, satisfy the following system of equations:

$$\mathbf{y}_{m+1}^{(N)}\left[S^{(N)} \oplus D_0 + \mathbf{S}_0^{(N)} \otimes \boldsymbol{\beta}^{(N)} \otimes I_{\overline{W}}\right] + \mathbf{y}_m^{(N)}(I_{M^{(N)}} \otimes D_1) = \mathbf{0},$$
$$m = \overline{0, N-2},$$

$$\mathbf{y}_0^{(N)}\left[S^{(N)} \oplus D_0 + \mathbf{S}_0^{(N)} \otimes \boldsymbol{\beta}^{(N)} \otimes I_{\overline{W}}\right] + \mathbf{y}_{N-1}^{(N)}(I_{M^{(N)}} \otimes D_1) = \mathbf{0},$$

$$\sum_{m=0}^{N-1} \mathbf{y}_m^{(N)}\mathbf{e} = 1.$$

Again by direct substitution, it is possible to verify that the solution of this system of equations is the following:

$$\mathbf{y}_m^{(N)} = \frac{(\boldsymbol{\beta}^{(N)}(-S^{(N)})^{-1}) \otimes \boldsymbol{\theta}}{N b_1^{(N)}}, \ m = \overline{0, N-1}. \tag{4.3.5}$$

After substitution of (4.3.5) into inequality (4.3.2) and some algebraic manipulations, the inequality (4.3.1) is got. Theorem is proven.

Remark 13. *It has already been said that ergodicity (stability) condition for any queueing system is defined by its ability to reduce the customers number in the system in situation when this number is huge (the system is overloaded). For the system under study, when it is overloaded, the average number of customers arriving during the service time is equal to $\lambda b_1^{(N)}$ (here $b_1^{(N)}$ is the average service time of a batch of exactly N customers) while the number of customers departing from the system at service completion moment is given by N. Thus, an intuitively clear condition of the system ergodicity should be of form $\lambda b_1^{(N)} < N$ what coincides with strictly proven condition (4.3.1).*
The throughput of the system (the maximal intensity of customers flow that can be successfully processed by the system), which is one of the main performance measures of the system, is equal to $\frac{N}{b_1^{(N)}}$.

4.4 Key performance indices of the system

Further it will be assumed that inequality (4.3.1) is fulfilled. Then the stationary distribution of the Markov chain ξ_t exists.
The stationary state probabilities of the chain are denoted as

$$\boldsymbol{\pi}(i,m,r,\eta,\nu) = \lim_{t\to\infty} P\{i_t = i,\ m_t = m,\ r_t = r,\ \eta_t = \eta,\ \nu_t = \nu\},$$
$$i \geq 0,\ m = \overline{0,N-1},\ \ \nu = \overline{0,W},$$

with $\eta = \overline{1,M^{(0)}}$, if $r = 0$ and $\eta = \overline{1,M^{(r)}}$ if $r = \overline{1,N}$.

Let $\boldsymbol{\pi}(i,m,r)$ be the row vector of probabilities of the states belonging to the macro-state (i,m,r), $\boldsymbol{\pi}(i,m)$ be the row vector of probabilities of the states belonging to the extra-state (i,m), and $\boldsymbol{\pi}_i$ be the row vector of probabilities of the states belonging to the super-state i, $i \geq 0$.

The following theorem holds.

Theorem 14. *The stationary probability vectors $\boldsymbol{\pi}_i$ can be computed as follows:*

$$\boldsymbol{\pi}_i = \boldsymbol{\pi}_1 \boldsymbol{R}^{i-1},\ i \geq 2,$$

where $\boldsymbol{R}$ is solution to matrix equation

$$\mathbf{Q}^{+} + \boldsymbol{R}\mathbf{Q}^{0} + \boldsymbol{R}^2\mathbf{Q}^{-} = \boldsymbol{O}$$

having the spectral radius strictly less than 1 and the vectors $\boldsymbol{\pi}_0$ and $\boldsymbol{\pi}_1$ are defined as solution of the system

$$\boldsymbol{\pi}_0\mathbf{Q}_{0,0} + \boldsymbol{\pi}_1\mathbf{Q}_{1,0} = \mathbf{0},$$
$$\boldsymbol{\pi}_0\mathbf{Q}_{0,1} + \boldsymbol{\pi}_1(\mathbf{Q}^{0} + \boldsymbol{R}\mathbf{Q}^{-}) = \mathbf{0},$$

subject to normalizing condition

$$\boldsymbol{\pi}_0\mathbf{e} + \boldsymbol{\pi}_\mathbf{1}(\mathbf{I} - \boldsymbol{R})^{-\mathbf{1}}\mathbf{e} = \mathbf{1}.$$

Proof easily follows from [58], taking into account that the matrix $\boldsymbol{R}$ has been computed recursively, according to this relation:

$$\boldsymbol{R}_{k+1} = \left(\mathbf{Q}^{+} + \boldsymbol{R}_k^2 \mathbf{Q}^{-}\right) \left(-\mathbf{Q}^{0}\right)^{-1}, \quad k \geq 0,$$

assuming $\boldsymbol{R}_0 = \mathbf{0}$.

Once all the vectors $\boldsymbol{\pi}_i, \ i \geq 0$, have been computed, various performance measures of the system can be calculated:

- The average number of blocks of customers in the system, including one in service

$$L = \sum_{i=1}^{\infty} i\boldsymbol{\pi}_i \mathbf{e}.$$

- The average number of blocks of customers in the system, excluding one in service

$$\tilde{L} = \sum_{i=1}^{\infty} (i-1)\boldsymbol{\pi}_i \mathbf{e} = L - 1 + \boldsymbol{\pi}_0 \mathbf{e}.$$

- The average number of customers in the pool at an arbitrary moment

$$N^{(pool)} = \sum_{i=0}^{\infty} \sum_{m=1}^{N-1} m\boldsymbol{\pi}(i,m)\mathbf{e}.$$

- The average number of customers in service at an arbitrary moment

$$N^{(serv)} = \sum_{r=1}^{N} r \left[\boldsymbol{\pi}_1 \left(I - R\right)^{-1}\right]^{(r)} \mathbf{e},$$

where the denotation $[\boldsymbol{\pi}_i]^{(r)} = \sum\limits_{m=0}^{N-1} \boldsymbol{\pi}(i,m,r)$ has been used.

- The average number of customers in the system at an arbitrary moment

$$
\begin{aligned}
N^{(syst)} &= \sum_{i=1}^{\infty}\sum_{m=0}^{N-1}\sum_{r=1}^{N}((i-1)N+m+r)\boldsymbol{\pi}(i,m,r)\mathbf{e} \\
&\quad + \sum_{m=0}^{N-1} m\boldsymbol{\pi}(0,m)\mathbf{e} \\
&= N\tilde{L} + N^{(serv)} + N^{(pool)}.
\end{aligned}
$$

- The probability that an arbitrary customer immediately starts the service upon arrival

$$P_{imm} = \lambda^{-1}\boldsymbol{\pi}(0,N-1)(\mathbf{e}_{M_0} \otimes D_1\mathbf{e}_{\overline{W}}).$$

- The probability that the server is idle at an arbitrary moment

$$P_0 = \boldsymbol{\pi}_0(\mathbf{e}_{M_0} \otimes \mathbf{e}_{\overline{W}}).$$

- The probability that the server is idle at the arbitrary arrival moment

$$P_0^{(arrival)} = \lambda^{-1}\boldsymbol{\pi}_0(\mathbf{e}_{M_0} \otimes D_1\mathbf{e}_{\overline{W}}).$$

4.5 Waiting time distribution

In this section, the Laplace-Stilties Transform (LST) of the waiting time distribution is derived.

Waiting time is the time interval since the moment of arrival of an arbitrary customer to the system until the moment when this customer enters the service.

To get the *LST* of the stationary waiting time distribution, the “method of catastrophes” (also known as the method of additional events) [44, 68] can be used. This is a powerful method for the derivation of *LST* of distributions of quantities such as waiting

times, sojourn times, and busy period.
A catastrophe does not have any physical meaning and does not have any impact on the behavior of the queueing system that is being analyzed. The notion of the catastrophe in the context of a LST of a distribution is frequently employed due to its nice probabilistic interpretation which is briefly explained. It is assumed that, independently of the queueing system under study, there is a stream of catastrophes that arrive according to a Poisson process with rate, say, s. Here s is assumed to be real and positive. It is very easy to extend this to complex s, having real part that is positive. Suppose that ξ is a continuous random variable with distribution function $F_\xi(t)$. Then, it is obvious that the LST $\varphi(s) = \int\limits_0^\infty e^{-st} dF_\xi(t)$ of ξ gives the probability that a catastrophe from the stationary Poisson process with rate s will not arrive during the time given by ξ. The use of this probabilistic interpretation of a LST of a distribution greatly simplifies in obtaining an expression for the LST of the waiting time distribution under study. Therefore, this approach will be used below.

An arbitrary arriving customer can be tagged and its waiting in a queue can be monitored. Let $w(s)$ be the LST of the distribution of its waiting time or, in other words, the probability that a catastrophe from the stationary Poisson process with rate s will not arrive during the waiting time.
In order to derive expression for $w(s)$, other auxiliary denotations have to be introduced. Let $\boldsymbol{\beta}(0, m, s)$ be the column vector consisting of the LSTs of time until the tagged customer starts service, conditional that currently the server does not provide service (i.e., admission period is in a progress), m customers stay in the pool and the underlying processes of admission period and arrivals have the corresponding states. Let $\boldsymbol{\beta}(i, m, r, s)$ be the column vector consisting of the LSTs of time until the tagged customer starts service, conditional that currently the server provides service, there are i blocks in the system, m customers stay in the pool, current service is provided to the batch consisting of r customers and the underlying processes of service and arrivals have the correspond-

ing states.
Recursive formulas for the $LSTs$ $\boldsymbol{\beta}(0,m,s)$ and $\boldsymbol{\beta}(i,m,r,s)$ are given in the next two lemmas.

Lemma 15. *The LSTs* $\boldsymbol{\beta}(0,m,s)\,,\ m=\overline{1,N-1}$, *are computed from the following backward recursion:*

$$\boldsymbol{\beta}(0,N-1,s)=(sI-T\oplus D_0)^{-1}\left(T_0\otimes\mathbf{e}_{\overline{W}}+\mathbf{e}_{M_0}\otimes D_1\mathbf{e}_{\overline{W}}\right),$$

$$\boldsymbol{\beta}(0,m,s)=(sI-T\oplus D_0)^{-1}\left(T_0\otimes\mathbf{e}_{\overline{W}}+(I_{M_0}\otimes D_1)\,\boldsymbol{\beta}(0,m+1,s)\right),$$
$$m=N-2,N-3,...,1.$$

Proof. The following formula

$$\boldsymbol{\beta}(0,m,s)=\int\limits_0^{+\infty}e^{(-sI+(T\oplus D_0))t}dt$$
$$\times\left\{T_0\otimes\mathbf{e}_{\overline{W}}+I_{M_0}\otimes D_1\boldsymbol{\beta}(0,m+1,s)\right\},\qquad(4.5.6)$$

is obvious from the following considerations. After the moment of the arrival of the tagged customer, which joins the pool and becomes the m-th customer in the pool, during some time $t,\ 0<t<\infty$, catastrophe does not arrive (probability of this event is e^{-st}); possible transitions of the underlying process of the admission period do not lead to completion of this period and their probabilities are given by the matrix e^{Tt}; possible transitions of the underlying process of arrivals do not lead to new customer arrival and their probabilities are obtained by the matrix e^{D_0t}.
Joint probability of the described events is equal to

$$e^{-st}e^{Tt}\otimes e^{D_0t}=e^{(-sI+(T\oplus D_0))t}.$$

After the moment t, during the interval $(t,t+dt)$ of infinitesimal length, one of two events can happen:
(i) Admission period expires (probabilities of this event under the fixed states of the underlying process of the admission period are

given by the vector $T_0 \otimes \mathbf{e}_{\bar{W}} dt$) and service of the tagged customer starts, therefore probability that a catastrophe does not arrive during the rest of the waiting time is equal to 1;
(ii) New customer arrives (probabilities of this event under the fixed states of the underlying process of arrivals are given by the matrix $I_{M_0} \otimes D_1 dt$). If $m < N-1$, this customers joins the pool and probabilities that a catastrophe will not arrive during the rest of the waiting time of the tagged customer are given by the vector $\boldsymbol{\beta}(0, m+1, s)$. If $m = N-1$, the batch consisting of N customers (including the tagged one) starts service. Integrating over t, (4.5.6) is got.
The statement of this lemma stems from (4.5.6) noting that

$$\int\limits_0^{+\infty} e^{(-sI+(T\oplus D_0))t} dt = (sI - (T \oplus D_0))^{-1}.$$

Lemma 16. *The LSTs* $\boldsymbol{\beta}(i, m, r, s)$, $r = \overline{1, N}$, *are sequentially computed from equations*

$$\begin{aligned}\boldsymbol{\beta}(1, m, r, s) =& C_r(s)\boldsymbol{\beta}(0, m, s) + B_r(s)\boldsymbol{\beta}(1, m+1, r, s),\\ & m = \overline{1, N-2},\end{aligned}$$

$$\begin{aligned}\boldsymbol{\beta}(1, N-1, r, s) =& C_r(s)\boldsymbol{\beta}(0, N-1, s)\\ &+ B_r(s)H_r(s)\left(\beta^{(N)}\left(sI - S^{(N)}\right)^{-1}\mathbf{S}_0^{(N)}\right)^{i-1},\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{\beta}(i, m, r, s) =& A_r(s)\boldsymbol{\beta}(i-1, m, N, s)\\ &+ B_r(s)\boldsymbol{\beta}(i, m+1, r, s),\ i > 1, m = \overline{1, N-2},\end{aligned}$$

$$\begin{aligned}\boldsymbol{\beta}(i, N-1, r, s) =& A_r(s)\boldsymbol{\beta}(i-1, N-1, N, s)\\ &+ B_r(s)H_r(s)\left(\beta^{(N)}\left(sI - S^{(N)}\right)^{-1}\mathbf{S}_0^{(N)}\right)^{i-1},\ i > 1,\end{aligned}$$

where

$$\begin{aligned}
A_r(s) &= \left(sI - S^{(r)} \oplus D_0\right)^{-1}\left(\left(\mathbf{S}_0^{(r)}\beta^{(N)}\right) \otimes I_{\overline{W}}\right),\\
B_r(s) &= \left(sI - S^{(r)} \oplus D_0\right)^{-1}\left(I_{M_r} \otimes D_1\right),\\
C_r(s) &= \left(sI - S^{(r)} \oplus D_0\right)^{-1}\left(\left(\mathbf{S}_0^{(r)}\tau\right) \otimes I_{\overline{W}}\right),\\
H_r(s) &= \left(\left(sI - S^{(r)}\right)^{-1}\mathbf{S}_0^{(r)}\right) \otimes \mathbf{e}_{\overline{W}}.
\end{aligned}$$

Proof is analogous to the proof of the previous lemma.

The next theorem is devoted to formulas for the computation of *LST* $w(s)$.

Theorem 17. *The LST* $w(s)$ *is computed as follows:*

$$\begin{aligned}
w(s) =& P_{imm} + \lambda^{-1}\left[\sum_{i=1}^{\infty}\sum_{r=1}^{N}\pi(i, N-1, r)(I_{M_r} \otimes D_1\mathbf{e})\right.\\
&\times (sI - S^{(r)})^{-1}\mathbf{S}_0^{(r)}(\beta^{(N)}(sI - S^{(N)})^{-1}\mathbf{S}_0^{(N)})^{i-1}\\
&+ \sum_{m=0}^{N-2}\pi(0, m)(I_{M_0} \otimes D_1)\boldsymbol{\beta}(0, m+1, s)\\
&\left.+ \sum_{i=1}^{\infty}\sum_{m=0}^{N-2}\sum_{r=1}^{N}\pi\left(i, m, r\right)\left(I_{M_r} \otimes D_1\right)\boldsymbol{\beta}\left(1, m+1, r, s\right)\right].
\end{aligned}$$

Proof obviously follows from the formula of total probability. Expression $(\beta^{(N)}(sI - S^{(N)})^{-1}\mathbf{S}_0^{(N)})$ defines the *LST* of service time of a block consisting of N customers, the vector $(sI - S^{(r)})^{-1}\mathbf{S}_0^{(r)}$ defines the *LST* of the residual service time of a batch consisting of r customers. The term P_{imm} accounts the possibility that the tagged customer starts service immediately upon arrival.

Moreover, starting from $w(s)$ and making appropriate derivations, the average waiting time can be computed. It is necessary to take into account that:

$$\begin{aligned}
\left(sI - S^{(r)}\right)^{-1} \mathbf{S}_0^{(r)}\Big|_{s=0} &= \mathbf{e}, \\
\left(\left(sI - S^{(r)}\right)^{-1} \mathbf{S}_0^{(r)}\right)'\Big|_{s=0} &= \left(S^{(r)}\right)^{-1} \mathbf{e}, \\
A_r(s)|_{s=0} &= \left(-S^{(r)} \oplus D_0\right)^{-1} \left(\left(\mathbf{S}_0^{(r)} \beta^{(N)}\right) \otimes I_{\overline{W}}\right), \\
(A_r(s))'\big|_{s=0} &= \left(S^{(r)} \oplus D_0\right)^{-1} A_r(0), \\
B_r(s)|_{s=0} &= \left(-S^{(r)} \oplus D_0\right)^{-1} \left(I_{M_r} \otimes D_1\right), \\
(B_r(s))'\big|_{s=0} &= \left(S^{(r)} \oplus D_0\right)^{-1} B_r(0), \\
C_r(s)|_{s=0} &= \left(-S^{(r)} \oplus D_0\right)^{-1} \left(\left(\mathbf{S}_0^{(r)} \tau\right) \otimes I_{\overline{W}}\right), \\
(C_r(s))'\big|_{s=0} &= \left(S^{(r)} \oplus D_0\right)^{-1} C_r(0), \\
H_r(s)|_{s=0} &= \mathbf{e}_{M_r\overline{W}}, \\
(H_r(s))'\big|_{s=0} &= \left(\left(S^{(r)}\right)^{-1} \mathbf{e}\right) \otimes \mathbf{e}_{\overline{W}},
\end{aligned}$$

and the derivative of vectors $\boldsymbol{\beta}\,(i, m, r, s)$ can be computed recursively in this way:

$$\begin{aligned}
\boldsymbol{\beta}'\,(1, m, r, s)|_{s=0} =& C_r(0)\boldsymbol{\beta}'\,(0, m, 0) + (C_r(s))'\big|_{s=0}\, \mathbf{e} + \\
&+ (B_r(s))'\big|_{s=0}\, \mathbf{e} + B_r(0)\boldsymbol{\beta}'\,(1, m+1, r, 0)\,, \\
&m = \overline{0, N-2}, r = \overline{1, N},
\end{aligned}$$

with initial condition

$$\begin{aligned}
\boldsymbol{\beta}'\,(1, N-1, r, s)|_{s=0} = & (C_r(s))'\big|_{s=0}\, \mathbf{e} + C_r(0)\boldsymbol{\beta}'\,(0, N-1, 0) + \\
&+ (B_r(s))'\big|_{s=0}\, \mathbf{e}_{M_r\overline{W}} + B_r(0)\ (H_r(s))'\big|_{s=0} + \\
&+ B_r(0)\mathbf{e}_{M_r\overline{W}} \left(-\,(i-1)\, b_1^{(N)}\right),\ r = \overline{1, N},
\end{aligned}$$

and

$$\begin{aligned}
\boldsymbol{\beta}'\,(i, m, r, s)|_{s=0} = & (A_r(s))'\big|_{s=0}\, \mathbf{e} + A_r(0)\boldsymbol{\beta}'\,(i-1, m, N, 0) + \\
&+ (B_r(s))'\big|_{s=0}\, \mathbf{e} + B_r(0)\boldsymbol{\beta}'\,(i, m+1, r, 0)\,, \\
&i > 1, m = \overline{0, N-2}, r = \overline{1, N}
\end{aligned}$$

with initial condition

$$\begin{aligned}\boldsymbol{\beta}'\left(i, N-1, r, s\right)\big|_{s=0} = & \left(A_r(s)\right)'\big|_{s=0}\mathbf{e}+\\ & + A_r(0)\boldsymbol{\beta}'\left(i-1, N-1, N, 0\right)+\\ & + \left(B_r(s)\right)'\big|_{s=0}\mathbf{e}_{M_r\overline{W}} + B_r(0)\left(H_r(s)\right)'\big|_{s=0}+\\ & + B_r(0)\mathbf{e}_{M_r\overline{W}}\left(-\left(i-1\right)b_1^{(N)}\right),\\ & i>1, r=\overline{1,N}.\end{aligned}$$

Finally, the average waiting time can be easily computed by the formula

$$W_1 = -w'(0).$$

4.6 Numerical results

The aim of this section is to demonstrate feasibility of the proposed algorithms for computation of steady-state distributions of the system states and the waiting time under any fixed set of the system parameters; to show effect of variation of the maximal number N of customers that can be processed simultaneously in a batch; to illustrate the high positive effect of the proposed discipline, which suggests that the idle period of the server may end via accumulation of N customers in the pool or via expiration of certain random amount of time, whichever occurs first, comparing to the standard in literature discipline that requires mandatory accumulation of N customers; to demonstrate the necessity to account the correlation in arrival process to avoid poor evaluation of performance of the system.

In derivations described in the previous sections, it has been assumed that the service time of a batch consisting of r customers has the PH distribution with irreducible representation $(\boldsymbol{\beta}^{(r)}, S^{(r)}),\ r=\overline{1,N}$.
To implement the numerical work, it is necessary to fix concrete dependence of the vectors $\boldsymbol{\beta}^{(r)}$ and sub-generators $S^{(r)}$ on r. Let it assume that the service time of an individual customer has

PH type distribution with irreducible representation $(\boldsymbol{\beta}, S)$ and the size of the vector $\boldsymbol{\beta}$ is M. Evidently, it is reasonable to set $(\boldsymbol{\beta}^{(1)}, S^{(1)}) = (\boldsymbol{\beta}, S)$. For $r > 1$, depending on the potential real world applications, one may think about many options. E.g., the service time of a batch consisting of r customers:

- does not depend on r and is identical, in stochastic sense, to the service time of an individual customer. This option is quite realistic, e.g., in modelling transportation systems. Travel time of the inter-city bus practically does not depend on the number of passengers in the bus;
- is the sum of r service times of individual customers;
- is the weighted sum of r service times of individual customers, e.g. their average value;
- is defined as the minimum of r service times of individual customers. This may be true in the system where, to guarantee quick delivering of some information, the latter is transmitted simultaneously in r channels;
- is defined as the maximum of r service times of individual customers. This kind of dependence takes place in modern telecommunication networks in some networks, e.g., in multirate wireless networks with protocol IEEE802.11 WLAN.

In the numerical results shown, the last option is namely fixed. In multi-rate protocols, several mobile stations share the same physical channel. Under the use of such protocols, a group of requests from users can be processed simultaneously in parallel and processing of the whole group is considered finished if processing of all individual requests belonging to this group is completed. Therefore, the length of the service period of a group has distribution of the maximum of several independent random variables, each of which represents the service time of an individual customer belonging to this group. Since the expectation of the maximum of a fixed number of independent random variables is less (and can be much less) than the sum of expectations of these random variables, the average time devoted to the service of an arbitrary customer under the proposed service discipline may be much less than such time under the classical service discipline. Thus, throughput of the system under the proposed service discipline is higher and other performance measures of the system may be much better compared to the classical admission discipline. In the numerical experiments that here are reported, advantages of the multi-rate transmission are quantitatively illustrated.

The service time of a batch consisting of r customers is defined as the maximum of r service times of individual customers. Because it has been assumed that the service time of an individual customer has PH type distribution with irreducible representation $(\boldsymbol{\beta}, S)$, the distribution of the maximum of r service times having such a distribution has to be recursively computed as made in the numerical sections of the previous chapters.

To illustrate the effect of correlation in arrival process, in experiments, three different $MAPs$ having the same fundamental rate $\lambda = 0.6$ but different coefficients of correlation of successive inter-arrival times will be taken in consideration (MAP_0, MAP_{02}, MAP_{038}).

As the main performance measure of the system under study in these experiments, the average waiting time W_1 of an arbitrary customer will be considered, while the other performance measures listed in Section 4.4 were computed as well.

It is worth to note that the numerous results of various numerical experiments show that the well known Little's formula is valid for the system under study in the following form:

$$W_1 = \lambda^{-1}(N\tilde{L} + N^{(pool)}).$$

In all the experiments, $\lambda = 0.6$ is fixed as the fundamental rate of the MAP, μ as intensity of exponential distribution of admission period, $0 < \mu \leqslant 25,$ distribution of service time of individual customer is Erlangian of order 2 with mean value equal to 1.

The dependency of W_1 on intensity μ at varying of the pool's capacity N is compared.

Figure 4.2 shows dependencies of W_1 on μ for MAP_0 and values of N equal to 2,3,4,5. The value of W_1 for $N = 1$ does not depend on μ and is equal to 1.125.

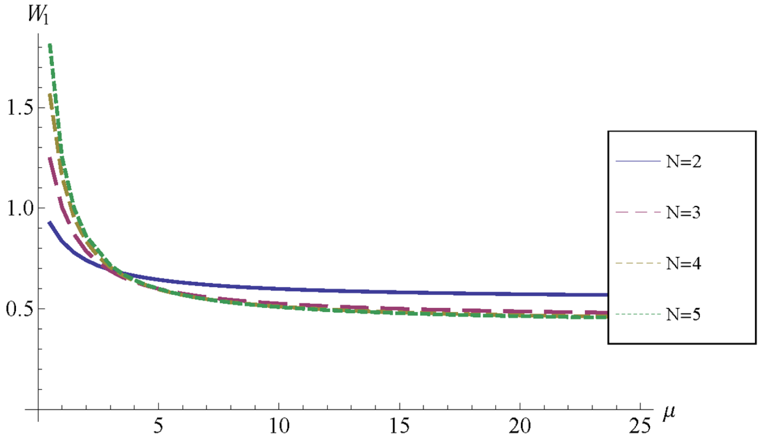


Figure 4.2 Average waiting time W_1 for different values of admission rate μ, and different dimensions N of the pool when corr=0

A careful examination of this figure reveals some interesting observations as summarized below:

- The introduction in the classical strategy, which assumes the possibility to start service only when the queue length reaches the level N, of the chance to be served also when a random admission period expires, essentially decreases the average waiting time. Classical strategy corresponds to the infinite length of the admission period (intensity of the admission period expiration equal to 0). It is evident from Figure 4.2, that the increase of μ essentially decreases the average waiting time.

- For small values of μ, small values of the pool capacity N are more preferable. However, when μ becomes larger than some value (about 3.5), large values of N become better. So, proper choice of μ is desirable for any value on N.

- Difference between values of W_1 for various N is significant, especially for small values of μ.

Figure 4.3 reports the behaviour of W_1 with respect to μ for $MAP_{0.2}$. In this case, for $N = 1$ the value of W_1 again does not depend on μ and is equal to 2.852.
The straight line for $N = 1$ is not represented in the figures to avoid suppression of the curves.

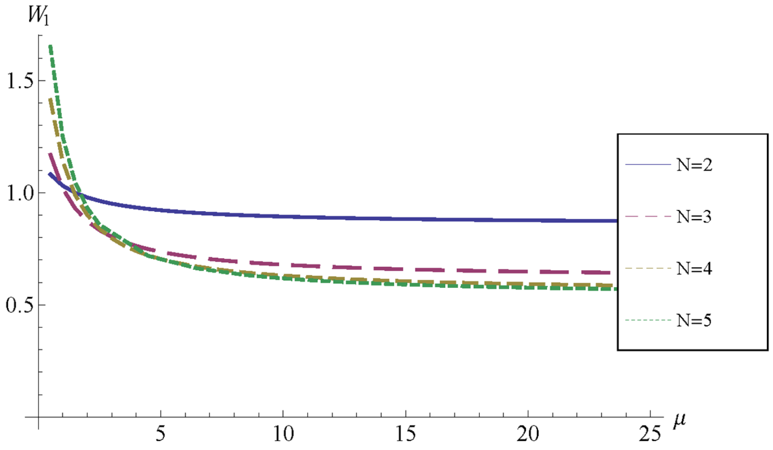


Figure 4.3 Average waiting time W_1 for different values of admission rate μ, and different dimensions N of the pool when corr=0.2

From Figure 4.3 the same conclusions of Figure 4.2 can be deduced. Moreover it can be observed that:

(i) Order of the curves for various values of μ can be different and more complicated than the one observed in Figure 4.2. Therefore, no "rule of thumb" can be formulated and computation of the average waiting time based on presented above results is mandatory for any available set of N and μ generated by a decision-maker who tries to optimize the system operation.

(ii) Correlation in arrival process increases the average waiting time.

The latter conclusion becomes much more evident after looking at Figure 4.4 that depicts W_1 as the function of μ for $MAP_{0.38}$. In this case, for $N = 1$, the value of W_1 is 77.648.

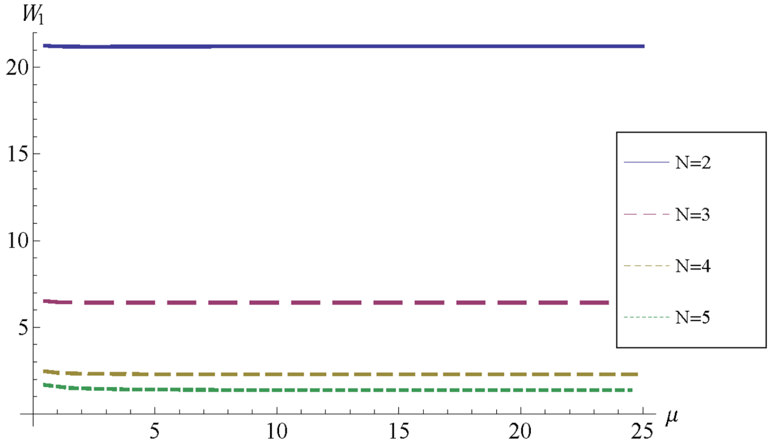


Figure 4.4 Average waiting time W_1 for different values of admission rate μ, and different dimensions N of the pool when corr=0.38

Figure 4.4 brightly illustrates the following two facts:

(i) Careful account of correlation in arrival process is vitally important to get correct evaluation of the system performance measures. Correlation is an important feature of the flows in modern telecommunication networks and it cannot be ignored by means of assuming that the arrival flow is described by the stationary Poisson arrival process. This ignorance can lead to huge errors.

(ii) The use of batch service can significantly help to improve quality of the system operation. For $N = 1$ it results $W_1 = 77.648$. For $N = 2$, W_1 is about 21, for $N = 3$, W_1 is about 6.4, for $N = 4$, W_1 is about 2.3, for $N = 5$, W_1 is less than 2.

All the three MAPs considered above are artificially constructed to illustrate effect of correlation and have two states of the underlying process $\nu_t,\ t \geqslant 0$.

Let the experiment for the MAP obtained be repeated as the result of fitting real world traces, see [26]. The underlying process

ν_t of this MAP has five states and is defined by the matrices D_0 and D_1 given by formulas:

$$D_0 = diag\{59620.6, 113826.1, 7892.6, 123563.2, 55428.2\},$$

$$D_1 = \begin{pmatrix} -59793.13 & 38.8 & 30.85 & 0.88 & 102.00 \\ 16.76 & -114709.36 & 97.52 & 398.90 & 370.08 \\ 281.48 & 445.97 & -9487.09 & 410.98 & 456.06 \\ 23.61 & 205.74 & 58.49 & -124162.13 & 311.09 \\ 368.48 & 277.28 & 7.91 & 32.45 & -56114.32 \end{pmatrix}.$$

These matrices are obtained based on information about the generator of the underlying process ν_t and the expression for D_1 as the diagonal matrix presented in [26]. The original matrices are scaled to get the MAP having the same fundamental rate $\lambda = 0.6$ as the three MAPs of order 2 which were used to build Figures 4.2-4.4. The coefficients of correlation and variation of the MAP are not changed under the scaling and are as follows: $c_{cor} = 0.141684$ and $c_{var}^2 = 1.46354$. Thus, the considered MAP of order 5 has the coefficients of correlation and variation intermediate between the values of these coefficients for MAP_0 and $MAP_{0.2}$.

Therefore, one may anticipate that the value of the average waiting time W_1 for various values of N should be intermediate between the values of W_1 for MAP_0 and $MAP_{0.2}$. However, for $N = 1$ it results that $W_1 = 1.125$ for MAP_0, $W_1 = 2.852$ for $MAP_{0.2}$, and $W_1 = 6.63691$ for the considered MAP of order 5.

High value of W_1 for the MAP of order 5 is easily explained by existence of two states of the underlying process ν_t, in which intensity of generation of customers is much higher than the intensity of generation in other states. Such irregularity in arrivals implies that sometimes the server is idle but sometimes it is highly loaded. It is important to note that the results of computations for $N = 2, 3, 4, 5$ presented in Figure 4.5 show that the negative effect of irregularity in arrivals is essentially mitigated by providing service in groups.

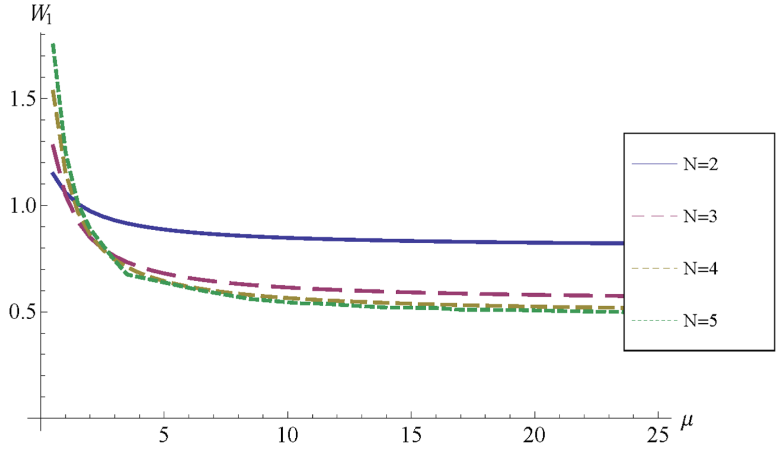


Figure 4.5 Average waiting time W_1 for different values of admission rate μ, and different dimensions N of the pool for the MAP of order 5

This confirms importance of the analysis here presented. Advantages of group service are illustrated and algorithmic tool is provided for optimal choice of the pair N and μ in situations when the use of large values of N is restricted technically or economically.

In telecommunications systems N can be interpreted as the level of multiplexing or the number of mobile stations, which can share the channel to access point, and the value of N can be limited by available bandwidth.

In applications to transportation systems, N can be interpreted as the capacity of vehicles or minivans which can be leased for passengers delivering.

In applications to manufacturing systems, N can be interpreted as the capacity of pallets used for providing technological operations like heating or cooling some details, etc.

After the choice of the appropriate value of the parameter N based on the restrictions on the waiting time of customers and cost of using the corresponding capacity of bandwidth, the presented results allow to fix also the suitable value of the parameter μ. The choice of μ is not trivial. If one chooses small value by μ, he benefits from high coefficient of utilization of the used capacities (bandwidth, vehicles, pallets, etc) but risks to provide poor quality of service. Long waiting time can cause that information to be transmitted may be outdated, passengers to be delivered to airport can miss the flight, details to be processed can lose required properties, etc. But if one chooses large value by μ, he benefits from the usage of advantages of group service and provides good quality of service, but the coefficient of utilization of the used capacities may be low. One may observe in Figures 4.2, 4.3, 4.5 that for large values of μ difference between the values of W_1 for $N = 4$ and $N = 5$ is quite small.

Therefore, the presented results can be useful to find some trade-off between quality of provided service and provider's expenditures.

Conclusion

The main goal of this research has been to study performance indices of $MAP/PH/1$ systems with novel admission strategies, in order to essentially reduce server idle time during periods when there are customers in the system and, thus, to increase system throughput as well as to provide better quality of service for customers. This has been achieved by means of providing the service to customers not individually, but in groups, and also introducing innovative admission strategies.

Significant advantages of the first two proposed disciplines have been numerically shown, comparing every result with the corresponding ones valid for classical discipline and another present in literature. In particular, it has been demonstrated that, under a proper choice of the length of admission period and the size of the groups, the performance measures considered for the numerical examples (the average number of customers in the system and the probability that an arbitrary customer gets service without visiting the orbit) give much better values, for several values of correlation.

As regards the third model analysed, the average waiting time of an arbitrary customer before receiving the service has been computed as the main performance measure of the system. Also in this case, the advantages of the new discipline have been numerically shown, over the classical one without possibility of starting the service earlier than the number of customers in the system reaches the predefined threshold value, at varying of admission rate and dimension of the groups, for several values of correlation.

The results can be used for optimal matching of the system parameters and parameters defining strategy of customers admission.

In conclusion, the analysis of the admission strategies proposed for the $MAP/PH/1$ queueing systems, introduced in this work of thesis, confirms the significant improvements that these disciplines can bring to the specified queueing systems, comparing to the classical strategy.

For this reason, these strategies have an high potential to be applied to real life communication systems and not only.

Bibliography

[1] M. Ajmone Marsan, G. De Carolis, E. Leonardi, R. Lo Cigno, M. Meo, Efficient estimation of call blocking probabilities in cellular mobile telephony networks with customer retrials, IEEE J. Sel. Areas in Commun. 19 (2) (2001) 332-346.

[2] A.S. Alfa, K.P. Sapna Isotupa, An $M/PH/k$ retrial queue with finite number of sources, Computers and Operations Research 31 (9) (2004) 1455-1464.

[3] J.R. Artalejo, Accessible bibliography on retrial queues: Progress in 2000-2009, Mathematical and Computer Modelling, 51 (2010) 1071-1081.

[4] J.R. Artalejo, A. Gomez-Corral, Modelling Communication Systems with Phase Type Service and Retrial Times, IEEE Communications Letters 11 (12) (2007).

[5] J.R. Artalejo, Gomez-Corral A., Retrial queueing systems: A computational approach, Berlin-Heidelberg: Springer, 2008.

[6] S. Asmussen, Applied Probability and Queues, New York: Springer, 2003.

[7] N.T.J. Bailey, On Queueing Processes with Bulk Service, Journal of the Royal Statistical Society, Series B 61 (1954) 80-87.

[8] A. Banerjee, U.C. Gupta, S.R. Chakravarthy, Analysis of a finite-buffer bulk-service queue under Markovian arrival pro-

cess with batch-size-dependent service, Computers and Operations Research 60 (2015) 138-149.

[9] P.P. Bocharov, R. Manzo, A.V. Pechinkin, Analysis of a two-phase queueing system with a Markov arrival process and losses, Stability Problems for Stochastic Models. Journal of Mathematical Sciences 131 (2005) 5606-5613.

[10] P.P. Bocharov, R. Manzo, A.V. Pechinkin, Two-phase queueing system with a Markov arrival process and blocking, Stability Problems for Stochastic Models. Journal of Mathematical Sciences 132 (2006) 578-589.

[11] L. Breuer, A.N. Dudin, V.I. Klimenok, A retrial $BMAP/PH/N$ system,Queueing Systems 40 (2002) 433-457.

[12] L. Breuer, V.I. Klimenok, A.A. Birukov, A.N. Dudin, U. Krieger, Modeling the access to a wireless network at hot spots, European Transactions on Telecommunications 16 (2005) 309-316.

[13] E. Brockmeyer, H.L. Halstrom, A. Jensen, The life and works of A. K. Erlang, The Copenhagen Telephone Company (1948) Copenhagen, Denmark.

[14] A. Brugno, A.N. Dudin, R. Manzo, Analysis of a Strategy of Adaptive Group Admission, Journal of Ambient Intelligence & Humanized Computing (2016) 1-13.

[15] A. Brugno, A.N. Dudin, R. Manzo, Retrial Queue with Discipline of Adaptive Permanent Pooling, Applied Mathematical Modelling (2015) - Under revision.

[16] A. Brugno, C. D'Apice, A.N. Dudin, R. Manzo, Analysis of a MAP/PH/1 Queue with Flexible Group Service, International Journal of Applied Mathematic and Computer Science 27 (1) (2017) 119-131.

[17] G. Casale, E.Z. Zhang, E. Smirn, Trace data characterization and fitting for Markov modeling, Performance Evaluation 67 (2010) 61-79.

[18] S.R. Chakravarthy, A finite capacity $GI/PH/1$ queue with group services, Naval Research Logistics Quarterly 39 (1992) 345-357.

[19] S.R. Chakravarthy, Analysis of the $MAP/PH/1/K$ queue with service control, Applied Stochastic Models and Data Analysis 12 (1996) 179-191.

[20] S.R. Chakravarthy, Analysis of a priority polling system with group services, Commun. Statist. Stochastic Models 14 (1998) 25-49.

[21] S.R. Chakravarthy, The batch Markovian arrival process: a review and future work, in: A. Krishnamoorthy, N. Raju, V. Ramaswami (Eds.), Advances in Probability Theory and Stochastic Processes, Notable Publications Inc., New Jersey (2001) 21-29.

[22] S.R. Chakravarthy, A.S. Alfa, A finite capacity queue with Markovian arrivals and two servers with group services, J. of Appl. Math. and Stochastic Analysis 7 (1994) 161-178.

[23] S.R. Chakravarthy, L. Bin, A finite capacity queue with non-renewal input and exponential dynamic group services, INFORMS Journal on Computing 9 (1997) 276-287.

[24] S.R. Chakravarthy, A.N. Dudin, A multiserver retrial queue with $BMAP$ arrivals and group services, Queueing Systems 42 (2002) 5-31.

[25] M. Chaudhry, J. Templeton, A first course in bulk queues, Wiley, New York, 1983.

[26] A. Chydzinski, Transient analysis of the $MMPP/G/1/K$ queue, Telecommunication Systems 32 (2006) 247-262.

[27] J.W. Cohen, Basic problems of telephone traffic theory and the influence of repeated calls, Philips Telecommunication Review 18 (1957) 49-100.

[28] C. D'Apice, R. Manzo, A Finite Capacity $BMAP_K/G_K/1$ Queue with the Generalized Foreground-Background Processor-Sharing Discipline, Automation and Remote Control 67 (2006) 428-434.

[29] C. D'Apice, R. Manzo, A.V. Pechinkin, A Finite $MAP_K/G_K/1$ Queueing System with Generalized Foreground-Background Processor-Sharing Discipline, Automation and Remote Control 65 (2004) 1793-1799.

[30] B.T. Doshi, Queueing systems with vacations - a survey, Queueing Systems 1 (1986) 29-66.

[31] A.N. Dudin, R. Manzo, R. Piscopo, Single Server Retrial Queue with Adaptive Group Admission of Customers, Computers and Operations Research 61 (2015) 89-99.

[32] A.N. Dudin, V. Klimenok, Queueing system $BMAP/G/1$ with repeated calls, Mathematical and Computer Modelling 30 (1999) 115-128.

[33] A.N. Dudin, V.I. Klimenok, Multi-dimensional Quasi-Toeplitz Markov chains, Appl. Math. and Stochast. Analysis 12 (1999) 393-415.

[34] A.N. Dudin, V. Klimenok, A retrial $BMAP/SM/1$ system with linear repeated requests, Queueing Systems 34 (2000) 47-66.

[35] O. Dudina, Ch. Kim, S. Dudin, Retrial Queueing System with Markovian Arrival Flow and Phase Type Service Time Distribution, Computers and Industrial Engineering 66 (2013) 360-373.

[36] G. Falin, A survey of retrial queues, Queueing Systems. Theory and Applications 7 (1990) 127-168.

[37] G. Falin, J.G.C. Templeton, Retrial Queues (Series: Monographs on Statistics and Applied Probability), Chapman & Hall (1997).

[38] Y. Fangi, Performance evaluation of wireless cellular networks under more realistic assumptions, Wireless Commun. and Mobile Computing 5 (8) (2005) 867-885.

[39] C.H. Foh, M. Zukerman, J.W. Tantra,A Markovian framework for performance evaluation of IEEE 802.11, IEEE Trans. Wireless Commun. 6 (2007) 1276-1285.

[40] D. Guha, V. Goswami, A.D. Banik, Algorithmic computation of steady-state probabilities in an almost observable $GI/M/c$ queue with or without vacations under state dependent balking and reneging, Applied Mathematical Modelling 40(5) (2016) 4199-4219.

[41] D.P. Heyman, D. Lucantoni, Modelling multiple IP traffic streams with rate limits, IEEE/ACM Transactions on Networking 11 (2003) 948-958.

[42] Indra, Sharda, A batch arrival two-state $M/M/1$ queueing system with latest arrival run (RUN) having maximum effective length one, Int. J. of Information and Management Sciences, 15 (2004) 71-80.

[43] Indra, Vijay Kumar, Analysis of a two-dimensional bulk arrival queueing model with exhaustive and non-exhaustive service policy, AryaBhatta Journal of Mathematics and Informatics (2010).

[44] H. Kesten, J.Th. Runnenburg, Priority in Waiting Line Problems, Amsterdam, Mathematisch Centrum, 1956.

[45] C.S. Kim, V. Klimenok, A.N. Dudin, Optimization of Guard Channel Policy in Cellular Mobile Networks with Account of Retrials, Computers and Operation Research, 43 (2014) 181-190.

[46] C.S. Kim, V. Klimenok, V. Mushko, A.N. Dudin, The $BMAP/PH/N$ retrial queueing system operating in Markovian random environment, Computers and Operations Research 37 (2010) 1228-1237.

[47] C.S. Kim, V.V. Mushko, A.N. Dudin, Computation of the steady state distribution for multi-server retrial queues with phase type service process, Annals of Operations Research 201 (2012) 307-323.

[48] A. Klemm, C. Lindermann, M. Lohmann, Modelling IP traffic using the batch Markovian arrival process, Performance Evaluation 54 (2003) 149-173.

[49] V.I. Klimenok, A.N. Dudin, Multi-dimensional asymptotically quasi-Toeplitz Markov chains and their application in queueing theory, Queueing Systems 54 (2006) 245-259.

[50] V.I. Klimenok, D.S. Orlovsky, A.N. Dudin, A $BMAP/PH/N$ system with impatient repeated calls, Asia-Pacific Journal of Operational Research 24 (2007) 293-312.

[51] V.I. Klimenok, D:S. Orlovsky, C.S. Kim, The $BMAP/PH/N$ retrial queue with Markovian flow of breakdowns, European Journal of Operational Research 189 (2008) 1057-1072.

[52] L. Kosten, On the influence of repeated calls in the theory of probabilities of blocking, De Ingenieur 59 (1947) 1-25.

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

[54] D. Lucantoni, New results on the single server queue with a batch Markovian arrival process, Communication in Statistics-Stochastic Models 7 (1991) 1-46.

[55] A. Mèszáros, J. Papp, M. Telek, Waiting time distribution in a poisson queue with a general bulk service rule, Management Science, 21 (1975), 777-782.

[56] R. Mukhtar, S. Hanly, A model for TCP behaviour over cellular radio channels with link layer error recovery, Proc. IEEE Global Telecommunications Conference 3 (2001) 1776-1780.

[57] M. Neuts, A General Class of Bulk Queues with Poisson Input, The Annals of Mathematical Statistics 38 (1967) 759-770.

[58] M. Neuts, Matrix-geometric solutions in stochastic models, The Johns Hopkins University Press, Baltimore, 1981.

[59] M. Neuts, Structured stochastic matrices of $M/G/1$ type and their applications, Marcel Dekker, New York, 1989.

[60] S. Ouazine, K. Abbas, Development of computational algorithm for multiserver queue with renewal input and synchronous vacation, Applied Mathematical Modelling 40(2) (2016) 1137-1156.

[61] Y. Sakuma, A. Inoie, An approximation analysis for an assembly-like queueing system with time-constraint items, Applied Mathematical Modelling 38 (2014) 5870-5882.

[62] M. Schleyer, K. Furmans, An analytical method for the calculation of the waiting time distribution of a discrete time $G/G/1$-queueing systems with batch arrivals, OR Spectrum 29 (2007) 745-763.

[63] Sharda, A queueing problem with batch arrivals and correlated departures, Metrika, 20 (1973) 81-92.

[64] W. Sun, S. Li, E. Cheng-Guo, Equilibrium and optimal balking strategies of customers in Markovian queues with multiple vacations and N-policy, Applied Mathematical Modelling 40(1) (2016) 284-301.

[65] H. Takagi, Queueing analysis: a foundation of performance evaluation, North-Holland, 1991.

[66] N. Tian, Z.G. Zhang, Vacation queueing models: theory and applications, Springer, New York, 2006.

[67] K. Tianbo, Q. Wu, C. Williamson, MRMC: a multi-rate multi-channel MAC protocol for multi-radio wireless LANs. Proc. of WiNCS (2005) 1-8.

[68] D. van Dantzig, Chaines de Markof dans les ensembles abstraits et applications aux processus avec regions absorbantes et au probleme des boucles, Ann. de l'Inst. H. Poincare 14 (1955) 145-199.

[69] P. Tran-Gia, M. Mandjes, Modeling of customer retrial phenomen in cellular mobile networks, IEEE J. Sel. Areas in Commun 15 (1997) 1406-1414.

[70] T. Yang, J.G.C. Templeton, A survey on retrial queues, Queueing Systems 2 (1987) 201-233.

[71] D. Y. Yang, F. M. Chang, J. C. Ke, On an unreliable retrial queue with general repeated attempts and J optional vacations, Applied Mathematical Modelling 40(4) (2016) 3275-3288.

[72] Q. Ye, High accuracy algorithms for solving nonlinear matrix equations in queueing models, Advances in Algorithmic Methods for Stochastic Models - Proceedings of the 3rd International Conference on Matrix Analytic Methods, G. Latouche and P.G. Taylor (Editors), Notable Publications Inc. NJ. (2000) 401-415.

[73] Y. Zhang, J. Wang, F. Wang, Equilibrium pricing strategies in retrial queueing systems with complementary services, Applied Mathematical Modelling 40(11) (2016) 5775-5792.