Investigation of the performance of finite-source communication systems

Egyetemi doktori (PhD) értekezés

a szerző neve: Tóth Ádám
témavezető neve: Dr. Sztrik János

DEBRECENI EGYETEM
Természettudományi és Informatikai Doktori Tanács
Informatikai Tudományok Doktori Iskola
Debrecen, 2020.

Ezen értekezést a Debreceni Egyetem Természettudományi és Informatikai Doktori Tanács Informatikai Tudományok Doktori Iskola Informatikai Rendszerek és Hálózatok programja keretében készítettem a Debreceni Egyetem természettudományi/műszaki doktori (PhD) fokozatának elnyerése céljából.
Nyilatkozom arról, hogy a tézisekben leírt eredmények nem képezik más PhD disszertáció részét.
Debrecen, 20. . . . .

a jelölt aláírása

Tanúsítom, hogy Tóth Ádám doktorjelölt 2016- 2020 között a fent megnevezett Doktori Iskola Informatikai Rendszerek és Hálózatok programjának keretében irányításommal végezte munkáját. Az értekezésben foglalt eredményekhez a jelölt önálló alkotó tevékenységével meghatározóan hozzájárult. Nyilatkozom továbbá arról, hogy a tézisekben leírt eredmények nem képezik más PhD disszertáció részét.
Az értekezés elfogadását javasolom.

Debrecen, 20.. . . . . *a témavezető aláírása*

AZ ÉRTEKEZÉS CÍME

Értekezés a doktori (Ph.D.) fokozat megszerzése érdekében
az Informatikai tudományágban

Írta: Tóth Ádám okleveles mérnökinformatikus

Készült a Debreceni Egyetem Informatikai Tudományok Doktori Iskola
(Informatikai Rendszerek és Hálózatok programja) keretében

Témavezető: Dr. Sztrik János

Az értekezés bírálói:

Dr.	....
Dr.	....
Dr.	....

A bírálóbizottság:

elnök:	Dr.	....
tagok:	Dr.	....
	Dr.	....
	Dr.	....
	Dr.	....

Az értekezés védésének időpontja: 202… . ……………… … .

My publications

List of Papers [J]:

J1 Tóth Á., Bérczes T., Sztrik J. and Kvach A. S.
Simulation of finite-source retrial queuing systems with collisions and non-reliable server
Communications in Computer and Information Science 700 (**2017**), pp. 146–158, doi: 10.1007/978-3-319-66836-9_13

J2 Tóth Á., Bérczes T., Sztrik J. and Kuki A.
Comparison of two operation modes of finite-source retrial queuing systems with collisions and a non-reliable server by using simulation
Journal of Mathematical Sciences (**2019**), Vol. 237 (6), pp. 846–857, doi: 10.1007/s10958-019-04211-2

J3 Tóth Á., Sztrik J., Kuki A., Bérczes T. and Efrosinin D.
Reliability analysis of finite-source retrial queues with outgoing calls using simulation
The International Conference on Information and Digital Technologies 2019 (**2019**), IEEE (2019), pp. 504–511, doi: 10.1109/DT.2019.8813419

J4 Sztrik J., Tóth Á., Pintér Á. and Bács Z.
Simulation of finite-source retrial queues with two-way communications to the orbit
Information Technologies and Mathematical Modelling - Queueing Theory and Applications, Communications in Computer and Information Science; Vol. 1109 **(2019)** pp. 270–284, doi: 10.1007/978-3-030-33388-1_22

J5 Nazarov A. A., Sztrik J., Kvach A. S. and Tóth Á.
Asymptotic sojourn time analysis of finite-source M/M/1 retrial queuing systems with collisions and server subject to breakdowns and repairs
Annals of Operations Research (**2020**), Vol. 288, pp. 417–434, doi: 10.1007/s10479-019-03463-0

J6 Tóth Á., Bérczes T., Sztrik J., Kuki A. and Schreiner W.
The simulation of finite-source retrial queuing systems with collision and blocking
Journal of Mathematical Sciences (**2020**), Vol. 246 (4), pp. 548–559, doi: 10.1007/s10958-020-04759-4

Software developed during PhD:

- *Simulation programs:* Based on SimPack

My contribution report

Papers J1, J2, J3, J4 and J6 are mainly the work of the author. Paper J5 is the main work of A. A. Nazarov and A. S. Kvach where the author contributed with simulation results.

Papers not included in the thesis:

J7 Tóth Á., Bérczes T., Kuki A., Almási B., Schreiner W., Wang J. and Wang F.
Analysis of finite source cluster networks
Creative Mathematics and Informatics 25 (**2016**), No. 2, pp. 223–235

J8 Bérczes T., Sztrik J., Tóth Á. and Nazarov A. A.
Performance modeling of finite-source retrial queuing systems with collisions and non-reliable server using MOSEL
Communications in Computer and Information Science 700 (**2017**), pp. 248-258, doi: 10.1007/978-3-319-66836-9_21

1. Danilyuk E. Yu. and Fedorova E. A. and Moiseeva S. P.
   Asymptotic analysis of an retrial queueing system M/M/1 with collisions and impatient calls
   Automation and remote control 79(12): pp. 2136-2146. **(2018)**
2. Vygovskaya O and Danilyuk E and Moiseeva S
   Retrial queueing system of MMPP/M/2 type with impatient calls in the orbit
   Communications in Computer and Information Science 912 pp. 387-399. doi: 10.1007/978-3-319-97595-5_30, **(2018)**
3. Danilyuk E and Vygoskaya O and Moiseeva S
   Retrial queue M/M/N with impatient customer in the orbit
   Communications in Computer and Information Science 919 pp. 493-504. doi: 10.1007/978-3-319-99447-5_42, **(2018)**

J9 Kuki A., Sztrik J., Tóth Á. and Bérczes T.
A contribution to modeling two-way communication with retrial queueing systems
Communications in Computer and Information Science 912 (**2018**), pp. 236–247, doi: 10.1007/978-3-319-97595-5_19

1. Dragieva Velika I. and Phung-Duc Tuan
   A finite-source M/G/1 retrial queue with outgoing calls
   Annals of operations research pp. 1–21., **(2019)**

2. Dragieva Velika and Phung-Duc Tuan
   On the busy period in a finite-source retrial queue with outgoing calls
   In: Moiseev, Alexander; Nazarov, Anatoly; Dudin, Alexander (szerk.) Information Technologies and Mathematical Modelling. Queueing Theory and Applications Springer International Publishing, pp. 1–13., **(2019)**

J10 Tóth Á. and Karimi R.
Optimization of hadoop cluster for analyzing large-scale sequence data in bioinformatics
Annales Mathematicae et Informaticae (**2019**), Vol. 50, pp. 187–202
doi: 10.33039/ami.2019.01.002

J11 Kuki A., Bérczes T., Sztrik J. and Tóth Á.
Reliability analysis of a two-way communication system with searching for customers
The International Conference on Information and Digital Technologies 2019 (**2019**), IEEE (2019), pp. 260-265, doi: 10.1109/DT.2019.8813455

J12 Kuki A., Bérczes T., Sztrik J. and Tóth Á.
Modeling of a two-way communication system with a special searching for customers
Communications in Computer and Information Science 1141 (**2019**), pp. 3–14, doi: 10.1007/978-3-030-36625-4_1

J13 Kuki A., Sztrik J., Bérczes T., Tóth Á. and Efrosinin D.
Numerical analysis of non-reliable retrial queuing systems with collision and blocking of customers
Journal of Mathematical Sciences (**2020**), Vol. 248, pp. 1–13

J14 Attila Kuki, Tamás Bérczes, Ádám Tóth, János Sztrik
Numerical analysis of finite source Markov retrial system with non-reliable server, collision, and impatient customers
Annales Mathematicae et Informaticae (**2020**), Vol. 51, pp. 53–63
doi: 10.33039/ami.2020.07.008

My publications

List of Conference proceedings [C]:

C1 Novac O. C., Bérczes T., Kuki A., Tóth Á. and Schreiner W.
Modeling RF-based sensor networks by using dual-source retrial queuing systems
14th International Conference on Engineering of Modern Electric Systems. Konferencia helye, ideje: Oradea, Románia, 2017.06.01 – 2017.06.02 Danvers (MA): IEEE, **2017**. pp. 149–153

C2 Attila Kuki, Tamas Berczes, Janos Sztrik, Adam Toth
Reliability Analysis of a Two-Way Communication System with Searching for Customers
Proceedings of The International Conference on Information and Digital Technologies 2019 Zilina, Szlovákia : IEEE, **(2019)** pp. 255–260

C3 János Sztrik, Ádám Tóth, Ákos Pintér, and Zoltán Bács
Simulation of finite-source retrial queues with two-way communications to the orbit
In: A.A., Nazarov; S.P., Moiseeva; A.Yu., Matrosov; E.Yu., Lisovskaya Information technologies and mathematical modeling (ITMM-2019): Proceedings of the XVIII International Conference named after A.F. Terpugov Tomsk, Oroszország, **(2019)** pp. 104–109.

C4 Adam Toth, Janos Sztrik, Attila Kuki, Tamas Berczes, Dmitry Efrosinin
Reliability Analysis of Finite-Source Retrial Queues with Outgoing Calls Using Simulation
Proceedings of The International Conference on Information and Digital Technologies 2019 Zilina, Szlovákia : IEEE, **(2019)** pp. 521–528.

C5 Ádám Tóth, János Sztrik
Simulation of finite-source retrial queuing systems with collisions, non-reliable server and impatient customers in the orbit
Proceedings of the 11th International Conference on Applied Informatics (ICAI 2020) (**2020**), CEUR Workshop Proceedings 2650, pp. 408–419

List of Others [O]:

O1 Schreiner W., Bérczes T. and Tóth Á.
Analyzing cluster scheduling schemes by probabilistic model checking
Research Institute for Symbolic Computation (RISC), Johannes Kepler University, Linz, Austria. Technical report, **September 2014**

O2 Schreiner W., Bérczes T., Sztrik J. and Tóth Á.
Analyzing cluster scheduling schemes by probabilistic model checking
Research Institute for Symbolic Computation (RISC), Johannes Kepler University, Linz, Austria. Technical report, **October 2015**

O3 Schreiner W., Bérczes T., Sztrik J. and Tóth Á.
Modeling RF communication in sensor networks by probabilistic model checking
Research Institute for Symbolic Computation (RISC), Johannes Kepler University, Linz, Austria. Technical report, **October 2015**

Contents

1

Background and motivation

The first chapter introduces retrial queuing systems and the motivation of the research conducted in the thesis

Contents

1.1 Background

In our world, the phenomenon of waiting is presented in the congested, urbanized societies. In many fields of life, people experience the annoyance of having to wait in queues, which is often not a pleasant and satisfying act. Whenever a demand for a service surpasses the capacity of a service unit it easily leads the others to wait. Decisions in connection to the amount of capacity are very complicated and often impossible to make because it is unpredictable when customers will arrive or how much time will be needed to provide its appropriate service. Increasing the capacity of a system is often costly for example adding more servers but fewer servers can cause a higher possibility of congestion and larger waiting queues in the system. With the help of queueing theory, various system characteristics are predictably lessening the hardship of such decisions providing vital information such as the average number of customers in the system, the average waiting time of a customer, etc. In [65] queuing system is described as a service centre and a population of customers, that at some times enter the service centre to get service. In many cases, the service centre can only serve a limited number of customers at a time. If a new customer arrives and there is no free service capacity, the customer enters a waiting line and waits until the service facility becomes available. Figure 1.1 shows the elements of a simple classical queueing model.

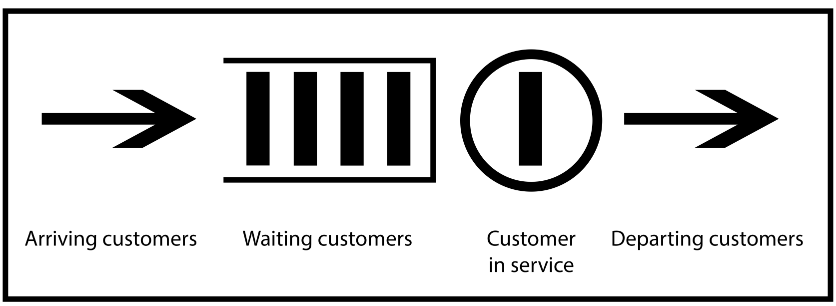

Figure 1.1: Classical queuing model

Based on queuing theory there are many applications, which have been documented in many fields like in probability, operation research, management science, communication systems, and industrial engineering. The following show some examples where queues occur and play an important role described by [1],[65].

- Supermarket: At the checkouts, customers usually wait from time to time. During peak-hours, more customers tend to be in the shops forming longer queues.

- Traffic lights: Customers spend a lot of time in front of traffic lights. An interesting topic is to manage waiting times in a way such that to be minimized.

- Computer: In a computer, the processor runs a set of tasks and each of them necessitates some computation time. Many papers devote to the interest of task completion times and there are other shared resources like disks, printers, etc. Usually, the jobs wait for these resources compelling to be queued.

- Dental clinic: At practitioner or in dental clinic patients are usually examined in the order of arrival. Often patients have to wait for doctors even if the appointment system exists.

- Call centre: A company frequently requires a call centre because it provides a channel for customers to contact the company. The call centre has a team structure where each operator helps a customer. In the case of an idle agent, a customer calling for help is immediately answered but might wait in a queue till an operator becomes available.

- Post Office: In a post office there may be specialized counters for stamps, packages, financial transactions, etc. Counters with the same specialisation could have one common queue or separate queue.

- Data communication: So much data traverses the Internet, that in the information-hungry world multiple communications channels are multiplexed onto and demultiplexed from data communications lines at a number of switches throughout the interconnection which forms a path. Along this path at many points, queues are built to manage data transmission. Also in this path, an advanced feature called Quality of Service can take place that prioritizes internet traffic for applications. Depending on the traffic types there can be different numbers of queues.

Since queueing theory is applied in different fields, often the terms job or task are used for a customer. The service centre is also called a processor or machine.

At some systems in front of the service provider instead of a physical waiting line, the members of the queue may be dispersed throughout an area. In that case, they wait for the server to come to them, like machines waiting to be repaired ([32]).

1.1.1 Retrial queuing models

Some queuing models assume that an arriving customer waits until being served because the waiting capacity is infinite. In some other models called loss models when the customer at the time of arrival sees that the service area is fully occupied leaves the system and is lost forever. However, in real life, there are various situations when customers instead of waiting temporarily leave the service facility awhile and attempts to be served after some random time. In that case, this customer resides in a virtual waiting room called *orbit* before launching its attempt to reach a server again. Those models which of possess an orbit can be modeled with retrial queues. Queuing systems with retrial queues are common and powerful tools modelling problems arising in major telecommunication systems, such as telephone switching systems, call centres, CSMA-based wireless mesh networks in frame level. Their importance can be viewed in the following works like in [9],[22],[23],[31],[36]. In this type of queuing system an incoming customer tries to reach the service unit after some time and remains in the system if the server is occupied or not available upon its arrival. For example in a call centre when every operator is busy and a customer initiates a phone call then it has to try to make a phone call again after some time. In the case of computer networks, TCP (Transmission Control Protocol) uses a retransmission mechanism. This occurs whenever a packet gets lost and TCP tries to retransmit this packet a while later [11],[12]. The importance of retrial queuing systems with infinite source are referred to the works [3],[7],[34],[39],[62],[68].

In a communication session where there are only a limited number of communication channels or other facilities the users (sources) usually fight for these resources. In many cases, there is a significant possibility of a conflict. Several sources launching uncoordinated attempts can produce collisions leading to the loss of the transmission and consequently the necessity for retransmission. It is very important to build up efficient procedures for preventing the conflict and corresponding message delay. There have been recent results on retrial queues with collision in [41],[42],[43],[44],[48].

Relatively just a small number of papers have been found during exploring scientific databases, in which queueing systems are investigated when the arriving

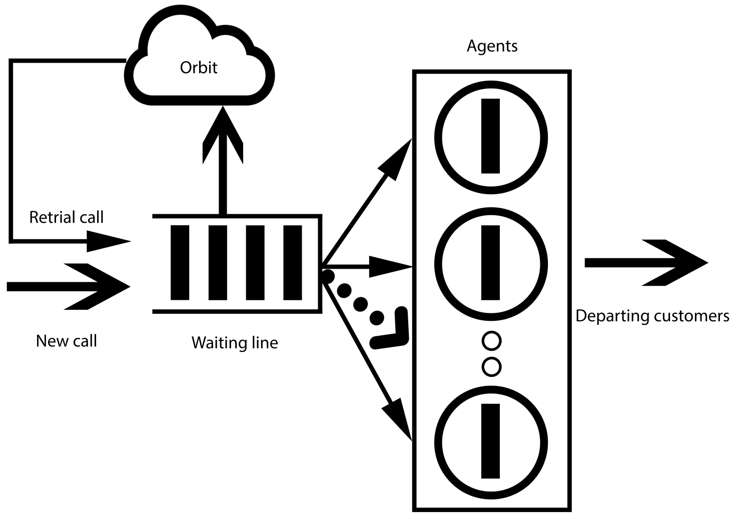

Figure 1.2: A retrial queueing model for call centres

calls either from the source or from the orbit cause collisions to the request under service and both are compelled to be delivered into the orbit like in [5],[16],[35],[40],[45],[53],[60].

In real-life applications of these types of systems in the retrial queuing literature, the server is usually assumed to be available perpetually. But these assumptions are quite unrealistic because errors, a power outage can supervene at any time. In the case of a wireless communication scenario where various factors have an effect on the transmission rate of the wireless channel and tend to be exposed to transmission failure, interruptions during transferring the packets like physical objects, radio frequency interference, electrical interference, or environmental factors. Thus it is very important to investigate retrial queueing systems with random server breakdowns and repairs. The non-reliable operation of the systems has a great influence on system characteristics and performance measures. Finite-source retrial queues with server breakdowns have been investigated in several recent papers, for example in [6],[58],[59], where the software MOSEL (Modeling, Specification and Evaluation Language) is used or in [28] where several homogeneous servers were modeled and analyzed by the help of Generalized Stochastic Petri nets (GSPNs) using retrial systems. There are other papers which studied a finite-source retrial queue where the server is subject to breakdowns like [20],[27],[28],[30],[37],[56],[61],[63],[64],[67],[69].

1.1.2 Two-way communication systems

Some of the first results connected with two-way communication retrial queues are gathered by Falin [21], who analyzes a system with a single server queue in which the outgoing and the incoming calls follow the same arbitrary service time distribution. M/M/1 and M/G/1 retrial queuing systems are investigated by [8],[10],[57] where the distribution of service times of incoming and outgoing calls follow two different distributions and server-orbit interaction is studied. The examination of the two-way communication retrial queuing system is quite a popular topic in recent years. This can be explained by the fact that using a two-way communication scheme is very helpful in many application fields modelling occurring real-life problems. Especially in the case of call centres where service units can perform certain other work in an idle state like selling, advertising, and promoting products including serving incoming calls. For example, a call centre of a credit card company where the operator may inform the customers about money payment or call them for commercial purposes. In such systems utilization of the service unit is always pivotal, see for example in [2],[4],[13],[19],[26],[38],[55],[66]. Results in connection with retrial queueing systems with two-way communication, where the source is infinite, are found in [8],[10],[17],[18],[49],[50],[54],[57]. When the population of customers is regarded as infinite then the probability of that a server calls a customer from the orbit is very small so under such circumstances it is suitable to model system with finite source. These real situations are the motivation for me to consider finite source retrial models with two-way communication.

Once the server becomes idle after some time it can call for customers inside and outside of the system which is called an outgoing call. Two types of outgoing calls are distinguished:

- the server may call a customer from the source to be served (primary outgoing call).
- The server can call a customer from the orbit, as well (secondary outgoing call).

In the thesis, two main different types of retrial queues are modeled and examined. In the first one, a finite-source retrial queuing system is considered which contains a non-reliable server and collisions can take place. The second one is a special retrial queueing system with the help of two-way communication where the server after becomes idle may call for customers from inside the system (from the orbit) or from outside the system (from finite source or infinite source). I aim to study the operation of these types of systems and to compare them with each other using various distribution of service time on performance measures like mean waiting time of an incoming call or utilization of the server. I am also mainly interested in how the different distributions modify the characteristics of the system. Different scenarios or operation modes are designed in order to compare the performance measures such as mean and variance of the waiting time, mean and variance of the number of customers in the system, mean and variance of the sojourn time in the orbit, mean and variance of time a customer spent in service of these modes or scenarios and estimations obtained by a developed simulation program package. The obtained results are illustrated in numerous figures emphasizing the differences or similarities.

The rest of the thesis is structured in the following way: Chapter 2 provides a short introduction about the general model which is used as a base model throughout the thesis. In that chapter, Section 2.2 introduces the simulation model and briefly how this is built up. I used a statistics package in order to obtain the desired estimate of mean and variance values of variables. Chapter 3 consists of the results in connection with the classical retrial queueing systems. In Section 3.1 with the help of the developed simulation program I illustrate how close the asymptotic result to the simulation results. In Section 3.2 I use the general model and I carry out a sensitivity analysis of the performance measures using various distributions to compare the steady-state distribution, mean waiting time, mean total uninterrupted service time of a customer, mean total interrupted service time of the investigated cases under different scenarios. In Section 3.3 I investigated the case when the service unit breaks down in a busy state and the service of the interrupted request is suspended and it continues after repairing the server. I compared the achieved results with the case when the interrupted request gets into the orbit instantaneously during a server failure. Blocking is introduced in the general model in Section 3.4 which means that the arriving customers cannot enter the system when the server is down and they return to the source and a new request generation process starts. With numerical results and with certain figures the effect of blocking is depicted. Chapter 4 presents the results in connection with retrial queuing systems with the help of two-way communication. In Section 4.1 I explore another system type of two-way communication. Here the service unit is reliable and able to make outgoing calls toward the customers residing in the orbit. I confer the results when the distribution of service time of the primary customers is disparate and I contrast them with the classical retrial queueing system in connection with the mean waiting time of primary customers and utilization of the service unit. I examine a system in Section 4.2 where the service unit becoming idle may call in a customer (outgoing call) from an infinite source after some exponentially distributed time. Four scenarios are differentiated in case of server failure and I study them using various distributions of failure and service time of primary customers on performance measures like mean waiting time of primary customers or utilization of the server. The service time of the primary customers and failure time of the server follows gamma, hypo-exponential, hyper-exponential, Pareto, and lognormal distribution with different parameters but with the same mean value. The novelty of this section is the examination of the effect of various distributions when the squared coefficient of variation is greater than one and when it is less than one.

Finally, Chapter 5 summarizes all the achieved results of the dissertation.

2

Simulation model

This chapter introduces the general simulation model and the operation of the developed simulation program used throughout the thesis

Contents

2.1 Description and notations

In this section, I introduce the system model in a general form. I consider a retrial queueing system of type M/G/1//N with collisions of the customers and an unreliable server (Figure 2.1), where the number of sources is denoted by N. Systems with retrial features are identified by a specific feature of arriving customers when the server is occupied. These customers stay in the system and spend their time in a virtual waiting room called the orbit. Each source can generate a request with a rate of λ/N, meaning that the distribution of inter-request time is exponential with parameter λ/N. Call generation is not possible until the end of the successful service of the customer by the source. There is no waiting queue thus in case of an idle server the service of an incoming customer starts instantly. The service times are supposed to be gamma distributed with parameter α and β. If the server is occupied with a customer, an arriving customer - either from the orbit or the source - will cause a collision with the customer under service and both requests are directed toward the orbit. The customers in the orbit are able to retry reaching the server after an exponentially distributed time with parameter σ/N. The server is non-reliable, so it is supposed to break down after an exponentially distributed time interval. In case of a busy server, the parameter is γ_0, otherwise (in case of an idle server) it is γ_1. The repair process starts immediately upon the breakdown. The repair time is an also exponentially distributed random variable with parameter γ_2. It is assumed that when the server is unavailable every source is eligible to generate customers and sends it to the unit, and these requests are forwarded to the orbit. Throughout the thesis, different cases are considered how the generation of request develops and what takes place until the end of the repair process of the server.

When the submission (the service of a request) is successful, the request goes back to the source. All the random variables involved in the model construction are assumed to be totally independent of each other.

2.2 Description of the simulation program

In this section, I develop simulation models based on using SimPack, which is a collection of C/C++ libraries and executable programs for computer simulation [24], to receive the desired performance measures. In this collection, various algorithms are supported connected with simulation including discrete event

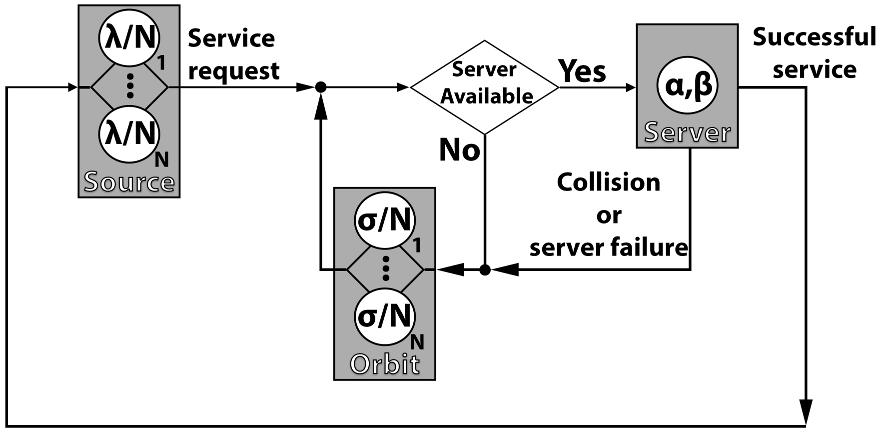

Figure 2.1: System model

simulation, continuous simulation, and combined (multi-model) simulation. I build a simulation model to quantify the performance of a system under study for various values of its input parameters. On the one hand, in some cases, performance measures are difficult and problematic to formulate or give exact formulas so the simulation can be a good alternative to approximate them. Because of the fact, that in many practical situations the state space of the describing Markov chain is enormous, to calculate the system measures in the traditional way of writing down and solving the underlying steady-state equations is nearly impossible. In order to simplify this procedure various software packages were evolved which are capable to describe and perform evaluation of complex systems, see for example [27],[28],[29],[30],[33]. But these types of applications are limited in terms of handling only exponentially distributed random variables or its memory usage is too high. On the other hand, one of the main reasons for the usage of simulation is that the user has the freedom what performance measures are calculated and how the model is built up. The first and most important step in building a simulation model of a system is to identify the basic events and each time with an occurrence of an event will alter the status of the system. During the time that elapses between two successive events, the system's status remains unchanged. In order to incorporate the basic events in the simulation model, every event is associated with timer information which will keep track of the time instants at which event will occur. In addition to that, a master clock is maintained to simply keep track of the simulated time. The simulation model, upon completion of processing an event, regroups all the possible events that will occur in the future, and finds the one with the smallest clock value. It then advances the time, i.e. the master clock, to this particular time when the next event will occur. It takes appropriate action as dictated by the occurrence of this event, and then repeats the process of finding the next event. The simulation model, therefore, moves through time by simply visiting the time instances at which events occur. In summary, using the timer information the model decides which of all the possible events will take place and the master clock is advanced to this time instant. The core of the simulation model centres around the manipulation of these events. In view of this, it suffices to monitor the changes in the system's status.

In my simulation environment, I applied random number generators of several distributions according to exponential, hyper-exponential, hypo-exponential, lognormal, gamma, and Pareto distribution. The following events may occur:

1. Arrival of a customer to the system: This event is always scheduled each time an arrival occurs.

2. Arrival of a customer to the server: This event will be triggered if the new arrival finds the server idle and its service is not interrupted by another arrival or server failure. Because of the existence of collision customers can come either from the orbit in numerous times or from the source.

3. Service completion at the server: This event will be triggered if the new arrival finds the server idle and its service is not interrupted by another arrival or server failure.

4. Arrival of a customer to orbit: This event will occur when a collision occurs and both the arriving customer and the customer under service get back into the orbit or the server breaks down while it is engaged. In that case, in some sections, the customer returns to the orbit without continuing its service.

5. Server breaks down: This event gives the time in the future when the server will be repaired and will become operational.

6. Server becomes operational: This event gives the time in the future when the server will break down. During this time the server is operational.

7. Server becomes idle: In some models, this event may take place after an exponentially distributed idle period.

In the simulation, typically, clocks are represented by real numbers. Therefore, it is not possible to have events occurring at the same time. I have implemented the system models and scheduling algorithms described above representing in Figure 2.1. The statistic module class derives from an adaptation of the statistics package, written by Andrea Francini in 1994 [25]. The statistic class is a statistical analysis tool capable of performing a quantitative estimate of the mean and variance values of the observed variables. The estimate of the mean value of a generic variable X is given by

$$\overline{X}(n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

where x_i equals to the observation of the variable X and n is the total number of observations collected. The distance between the estimated average, $\overline{X}$, and the actual statistical average ψ_x can be expressed in quantitative terms through the concepts of confidence interval and confidence level:

- If an average ψ_x with a confidence level equals to $1-\alpha$, then the corresponding confidence interval around $\overline{X}(n)$ must contain ψ_x with a probability of $1-\alpha$.

In mathematical terms the concept can be expressed as:

$$P(|\overline{X}(n) - \psi_x| \leq \Delta x) = 1 - \alpha, \qquad 0 < \alpha < 1$$

where Δx equals the half-amplitude of the confidence interval. Instead of this accuracy or relative precision has been defined:

$$\varepsilon = \frac{\Delta x}{\overline{X}(n)}$$

The statistical estimate of a stochastic process can be applied only if the process has already reached the condition of statistical stationarity. The collected observations during the initial transitional period must be discarded, as they may heavily deviate the results from the correct estimate values. A. Francini proposes two methods for calculating the initial transient, both based on the division of the observations collected in consecutive blocks of data, called batches

- Given a series of averages $\overline{X}_1(m_0), \overline{X}_2(m_0), \ldots,$ relating to batches of consecutive data, the initial transient can be considered as exhausted after b_0 batches, if the last k averages have an accuracy of δ.

This method involves assigning three parameters: m_0, k and δ. The second method, however, provides only the assignment of parameter m:

- Given a series of consecutive batches of data, the averages relating to them $\overline{X}_1(m), \overline{X}_2(m), ...,$ are used as secondary output data in the statistical analysis of the simulation results

The total average is obtained as:

$$\overline{\overline{X}}(k_b, m) = \frac{1}{k_b} \sum_{i=1}^{k_b} \overline{X_i}(m)$$

where k_b equals to the number of batches and m equals to the size of the batches.

Their sample variance then provides an estimate for the variance of a single $\overline{X_i}$:

$$S^2 = \frac{1}{k_b - 1} \sum_{i=1}^{k_b} (\overline{X_i}(m) - \overline{\overline{X}}(m))^2$$

The confidence interval of the estimator (at confidence level $1 - \alpha$) can be computed for mean $\overline{\overline{X}}$

$$\overline{\overline{X}}(k_b, m) \pm t_{k_b-1,1-\alpha/2} \frac{S^2}{\sqrt{k_b}}$$

where $t_{k_b-1,1-\alpha/2}$ is a $1 - \frac{\alpha}{2}$ critical value of the Student t distribution with $k_b - 1$ degrees of freedom. More from this method can be read in [14],[15],[46].

The simulations are performed with a confidence level of 99.9%. The relative half-width of the confidence interval required to stop the simulation run is 0.00001. The minimum number of observations to be collected before to check the initial transient closing condition is 100000. The maximum number of treatable observations is 1000000000. The size of a batch used to detect the initial transient duration is 1000. The number of transient batch means used to check the initial transient closing condition is 10. The accuracy level required to close the initial transient detection is 0.99. The initial size of a batch used during the stationary analysis is 10000.

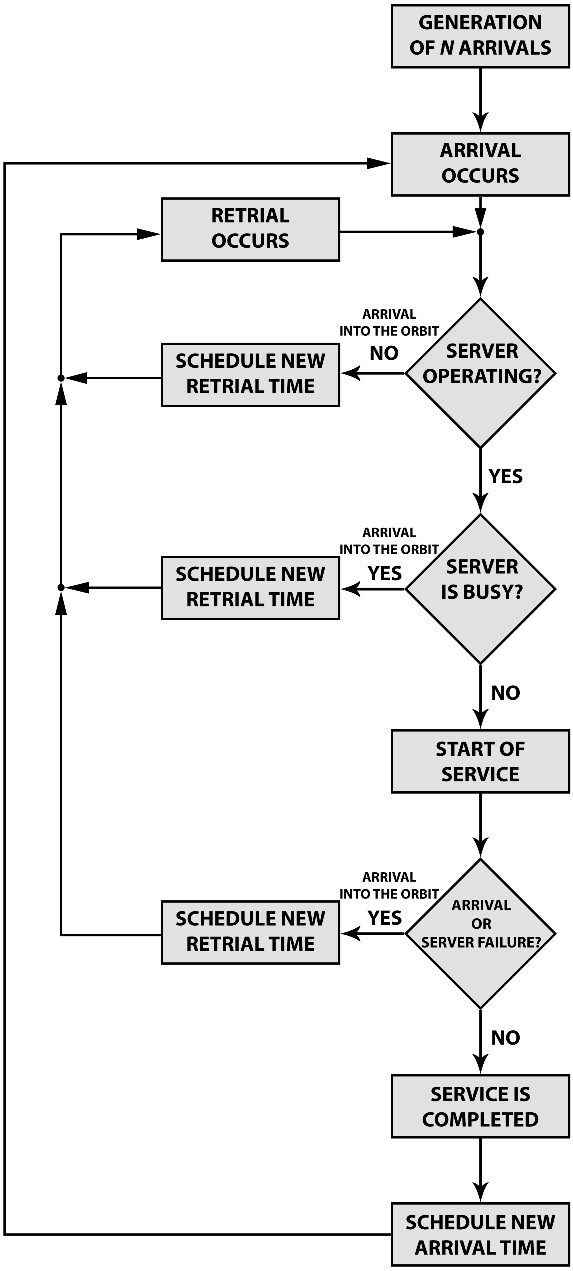

Figure 2.2: Simulation model

3

Finite-source retrial queuing systems

This chapter presents the obtained simulation results in connection with finite-source retrial queuing systems

Contents

3.1 M/M/1//N retrial queueing system with collisions and server subject to breakdowns and repairs

The present section aims to investigate a finite-source M/M/1 retrial queueing system with collisions of the customers [J5] where the server is subject to random breakdowns and repairs depending on whether it is idle or busy. An asymptotic method is applied under the condition that the number of sources tends to infinity while the primary request generation rate, retrial rate tends to zero and service rate, failure rates, repair rate are fixed. As the result of the analysis, it is shown that the probability distribution of the number of transitions/retrials of the customer into the orbit is geometric with a given parameter, and the normalized sojourn time of the customer in the system follows a generalized exponential distribution. It is also proofed that the limiting distributions of the sojourn time of the customer in the system and the sojourn/waiting time of the customer in the orbit coincide. The novelty of this investigation is the introduction of failure and repair of the server. Approximations of prelimit distributions obtained with the help of stochastic simulation by asymptotic one are considered and several illustrative examples show the accuracy and range of applicability of the proposed asymptotic method.

3.1.1 Model description and notation

Let us consider a retrial queueing system of type M/M/1//N with collisions of the customers and an unreliable server. The number of sources is N and each of them can generate a primary request with rate λ/N. A source cannot generate a new call until the end of the successful service of this customer. If a primary customer finds the server idle, he enters into service immediately, in which the required service time is assumed to be an exponentially distributed random variable with parameter μ. Otherwise, if the server is busy, an arriving (primary or repeated) customer involves into collision with a customer under service and they both moves into the orbit. The retrial time of the requests is assumed to be exponentially distributed with a rate of σ/N. We suppose that the server is unreliable and the lifetime is exponentially distributed random variable with failure rate γ_0 if the server is idle and with the rate γ_1 if it is busy. When the server breaks down, it is immediately sent for repair and the recovery time is assumed to be exponentially distributed with the rate γ_2. We deal with the case when the server is down all sources continue generation of customers and send it to the server, similarly customers may retry from the orbit to the server but all arriving customers immediately go into the orbit. Furthermore, in this unreliable model, we suppose the interrupted request goes to the orbit immediately and its next service is independent of the interrupted one. All random variables involved in the model construction are assumed to be independent of each other. Moreover, we can find the limiting probability distribution of the number of transitions of the tagged customer into the orbit, and, subsequently, the limiting probability distribution of the sojourn time of the customer in the orbit. With the help of a simulation program developed for these purposes, we can determine the accuracy and range of applicability of the asymptotic results in the prelimiting situation.

3.1.2 Numerical results and comparative analysis

My contribution in this paper is to illustrate how close the asymptotic results are to the simulation results. Since for a finite N the system is rather complicated it is almost hopeless to find the exact distributions analytically or even numerically. That is why we have developed a special simulation package written in C++ to get the estimations to the involved probabilities, means, and variances.

Using simulation we have reproduced the results obtained by a numerical procedure in [27]. Previously we have obtained that the probability distribution of the number of transitions of the tagged customer into the orbit is geometric with parameter q, defined in the [27]. Let us find out how close the limiting results are to the simulation results and at what values N this approximation is admissible.

Let us denote by $P_{as}(\nu = n)$ the asymptotic geometric distribution of probabilities with parameter q and by $P_s(\nu = n)$ the probability distribution of the number of transitions of the tagged customer into the orbit, obtained with the help of our simulation program. Furthermore, let us determine the accuracy (error) of approximation of distribution by mean of Kolmogorov distance Δ which for probability distributions $P_{as}(\nu = n)$ and $P_s(\nu = n)$ is defined as

$$\Delta = \max_{0 \leq i < \infty} \left| \sum_{n=0}^{i} (P_{as}(\nu = n) - P_s(\nu = n)) \right| .$$

Realizing the simulation program for

$$\lambda = 1, \qquad \mu = 1, \qquad \sigma = 4, \qquad \gamma_2 = 1$$

and applying the approximation [47], we provide the Kolmogorov distance Δ for various values N and $\gamma = \gamma_0 = \gamma_1$ in Table 3.1.

Table 3.1: Kolmogorov distance between distribution $P_s(i)$ and approximation of the geometric distribution $P_{as}(i)$ for various values of the parameters N and γ

	N=20	N=30	N=50	N=100	N=200
$\gamma = 0.05$	0.026	0.016	0.009	0.005	0.003
$\gamma = 0.1$	0.024	0.015	0.009	0.004	0.002
$\gamma = 0.5$	0.017	0.011	0.006	0.004	0.001

We can see, what is expected that by increasing N the Kolmogorov distance should decrease, but with this parameter setup, there is no essential reduction if $N > 50$.

3.2 Finite-source retrial queueing systems with collisions and non-reliable server

As described in Section 2.1 we consider a finite source retrial queueing system (Figure 2.1). In this model [J1], we suppose that the interrupted request gets into the orbit instantaneously and all of its services are independent of each other. When the submission is successful, the requests go back to the source.

The reason that we deal with rates λ/N and σ/N is that in [51],[52] similar systems were treated by an asymptotic method where N tends to infinity and it was proved that the number of customers in the systems follows a normal distribution.

The novelty of the investigation is to a carry sensitivity analysis of the performance measures using various distributions. Several figures show the effect of different distributions of service time on performance measures such as the mean number of customers in the system, mean and variance of response time, mean spent time in service of a customer, mean sojourn time in the orbit.

3.2.1 Scenarios

Scenario A: The following table shows the input parameters of Scenario A (see Table 3.2).

Table 3.2: Numerical values of model parameters

Case	N	λ/N	γ_0	γ_1	γ_2	σ/N	α	β
1	100	0.01	0.1	0.1	1	0.01	0.5	0.5
2	100	0.01	0.1	0.1	1	0.01	1	1
3	100	0.01	0.1	0.1	1	0.01	2	2

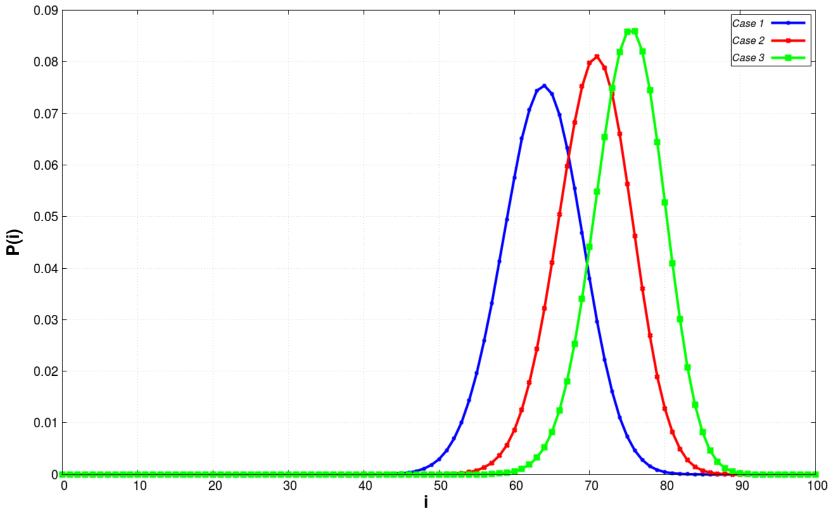

Figure 3.1: Comparison of steady-state distributions

Fig. 3.1 shows the steady-state distribution of the three investigated cases. It is observed the mean number of customers increases as α and β are getting larger. *Case 2* is a special case because when $\alpha = 1$ it represents the exponential distribution. From the shape of the curves, it is clearly visible that the steady-state distribution of the cases is normally distributed. The next table presents the considered performance measures in relation with the different cases (see Table 3.3).

In Table 3.3 the notations mean the followings:

E(NS): the mean number of customers,
Var(NS): the variance of the number of customers,
E(T): the mean response time,
Var(T): the variance of the response time,
E(W): the mean waiting time,
Var(W): the variance of the waiting time,
E(S): the total mean uninterrupted service time of a customer,
Var(S): the variance of total mean uninterrupted service time of a customer,

Table 3.3: Numerical results

Case	E(NS)	$Var(NS)$	E(T)	$Var(T)$	E(W)	$Var(W)$	E(S)	$Var(S)$	E(IS)
1	63.6842	27.9734	175.3073	65657.3454	174.5884	65434.6696	0.3147	0.1979	0.4041
2	70.5912	24.3012	239.9734	105273.4267	238.9734	104918.6389	0.4784	0.2289	0.5217
3	75.1825	21.2439	302.8106	151781.1411	301.5377	151277.6006	0.6472	0.2095	0.6257

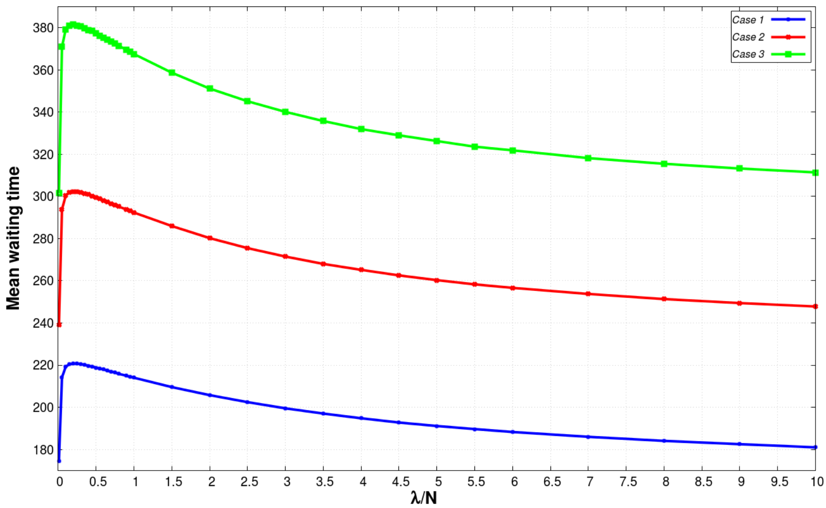

Figure 3.2: Mean waiting time vs. intensity of incoming customers

E(IS): the total mean interrupted service time of a customer.

Fig. 3.2 represents the conformation of mean waiting time. The same parameters are (see Table 3.2) used as in case of Fig. 3.1 but here the running parameter is λ/N. As it is expected with the increment of λ/N mean waiting time increases as well but an interesting phenomenon is noticeable namely after λ/N is greater than 0.1 mean waiting time starts to decrease.

Fig. 3.3 shows the effect of the inter-arrival time on the mean total uninterrupted service time of a customer. Because of the opportunity of collisions, it can easily happen that one job occurs multiple times in the service unit. Sooner or later its service is executed without being thrown out either because of an arriving customer or the failure of the server. Under the mean total uninterrupted service time of a customer, we mean the average of those specific last phase of service time intervals. When λ/N is 0.01 mean total uninterrupted service time of a customer is the highest then it starts to decrease as λ/N is increasing. This lasts till λ/N reaches 0.1, after it starts to increase.

In Fig. 3.4 the development of mean total interrupted service time of a customer can be seen. A customer can collide with another customer several times. This

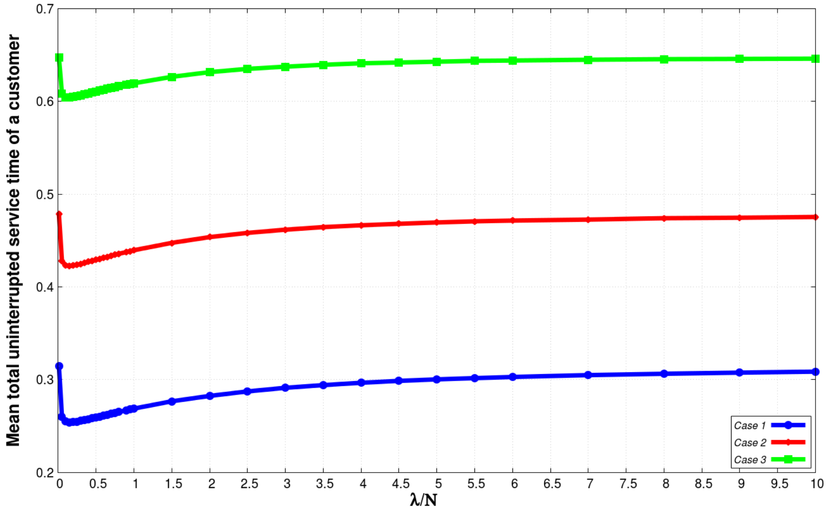

Figure 3.3: Mean total uninterrupted service time of a customer vs. intensity of incoming customers

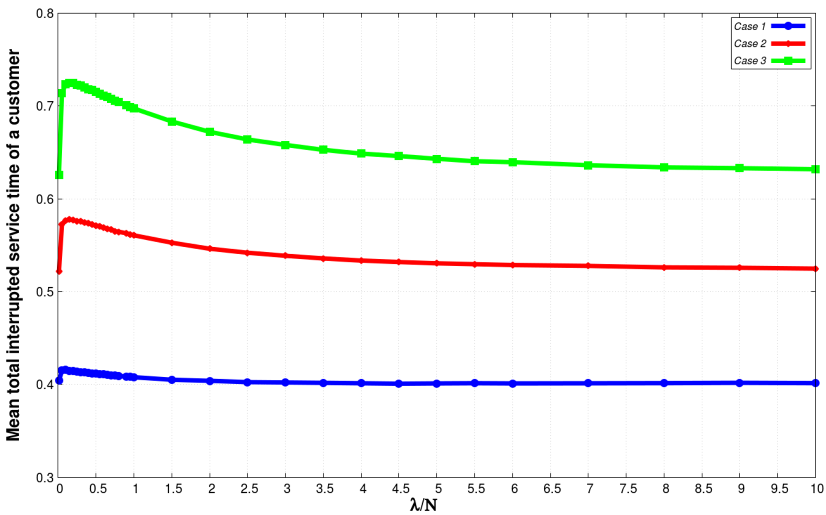

Figure 3.4: Mean total interrupted service time of a customer vs. intensity of incoming customers

performance measure includes the sum of every interrupted service time due to transmission toward the orbit because of collision or server failure divided by the number of departed requests. The opposite process happens compared to Fig. 3.3. It is worth mentioning when the service time is exponentially distributed then the sum of mean total interrupted service time of a customer and the mean total uninterrupted service time of a customer is equal to 1/parameter of service rate. This phenomenon can be explained by the following: due to the effect of collision uninterrupted service times constitute very low values and the average is also low. Because of exponential behaviour, every customer spends 1/parameter of service time (in this case 1) at the service unit and that is why the sum of mean total interrupted service time of a customer and the mean total uninterrupted service time of a customer is equal to one.

Scenario B: Scenario B is very similar to Scenario A except that now the distribution of inter-arrival times of the customers is not exponential but gamma-distributed. The next table (see Table 3.4) presents the input parameters of Scenario B.

Table 3.4: Numerical values of parameters of Scenario B

Case	N	α	β	γ_0	γ_1	γ_2	σ/N	α_1	β_1/N
1	100	1	1	0.1	0.1	1	0.01	0.5	0.01
2	100	1	1	0.1	0.1	1	0.01	1	0.01
3	100	1	1	0.1	0.1	1	0.01	2	0.01

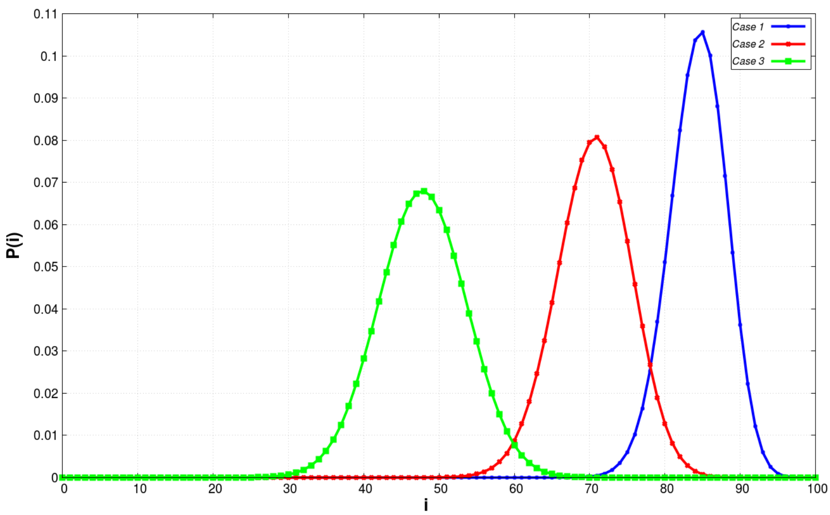

Figure 3.5: Steady-state distributions of Scenario B

Table 3.5: Numerical results

Case	E(NS)	$Var(NS)$	E(T)	$Var(T)$	E(W)	$Var(W)$	E(S)	$Var(S)$	E(IS)
1	84.3609	14.1827	270.0351	128831.5059	269.0354	128420.4087	0.4502	0.2039	0.5495
2	70.5912	24.3012	239.9734	105273.4267	238.9734	104918.6389	0.4784	0.2289	0.5217
3	47.7859	34.2376	183.0164	69830.9728	182.0164	69573.992	0.5462	0.2982	0.4538

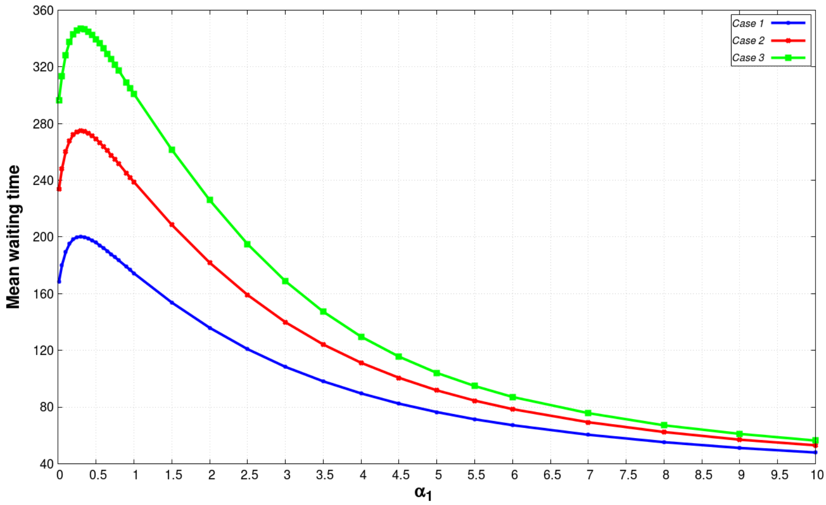

Figure 3.6: Mean waiting time vs. shape parameter

Fig. 3.5 displays the steady-state distribution of Scenario B. Now the service time is exponentially distributed ($\alpha = 1$) and the inter-arrival time is gamma-distributed. In these cases the steady-state distributions are still normally distributed and when α_1 is greater than 1 the mean number of requests in the system is significantly lower than in the other cases. In Table 3.8 the main performance measures can be found in connection with the cases.

In Fig. 3.6, 3.7 and 3.8 the service time distribution of the *Cases* is the following (Table 3.6).

Table 3.6: Parameters of service time

Case	α	β
1	0.5	0.5
2	1	1
3	2	2

All the other parameters are according to Table 3.3. The running parameter is α_1 so in this way, the impact of different distributions on the various performance measures can be discovered. First, the mean waiting time (Fig. 3.6), after an

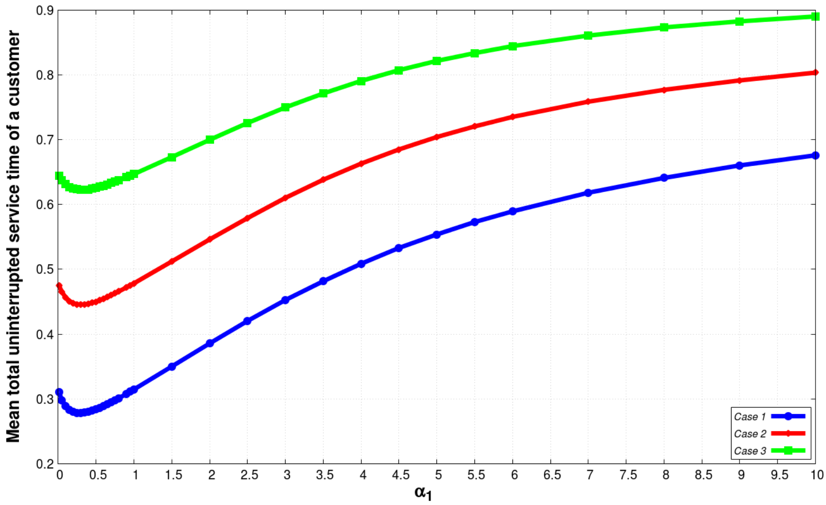

Figure 3.7: Mean total uninterrupted service time of a customer vs. shape parameter

initial jumps mean waiting time starts to monotonically decrease resulting that as α_1 is getting bigger the less time the customers spend in the system. In the end, the values of separate cases are almost equal.

As a consequence of Fig. 3.6 it is not surprising how the mean total uninterrupted service time and mean total interrupted service time of a customer change in the function of α_1. In those cases when the mean waiting time reaches the highest values the mean total uninterrupted service time is the lowest. The mean total interrupted service time acts the same as in case of the mean waiting time. It is interesting that after α_1 is higher than 4.5 we observe that in *Case 3* the values of the mean total interrupted service time of a customer is the lowest among the *Cases* despite the fact that at the very start it possesses the highest values of mean total interrupted service time of a customer.

Scenario C: In Scenario C not just inter-arrival but also the retrial time is generally distributed with the same parameters. Below Table 3.7 shows the input parameters of Scenario C.

Table 3.7: Numerical values of parameters of Scenario C

Case	N	α	β	γ_0	γ_1	γ_2	α_1	β_1/N
1	100	1	1	0.1	0.1	1	0.5	0.01
2	100	1	1	0.1	0.1	1	1	0.01
3	100	1	1	0.1	0.1	1	2	0.01

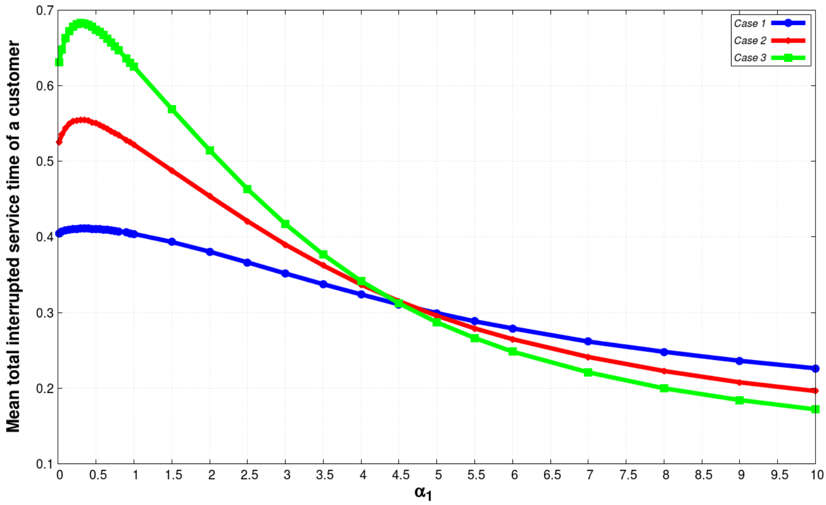

Figure 3.8: Mean total interrupted service time of a customer vs. shape parameter

Table 3.8: Numerical results

Case	E(NS)	$Var(NS)$	E(T)	$Var(T)$	E(W)	$Var(W)$	E(S)	$Var(S)$	E(IS)
1	81.5384	16.9599	220.8314	82735.764	219.8314	82377.8317	0.3398	0.1183	0.6602
2	70.5912	24.3012	239.9734	105273.4267	238.9734	104918.6389	0.4784	0.2289	0.5217
3	56.6635	30.1636	261.483	146264.4778	260.4827	145919.0907	0.626	0.3915	0.3743

Table 3.8 contains the main performance measures in connection with the cases.

This modification has no significant effect on the steady-state distribution (see Fig. 3.9). Of course the mean number of customers in the system is quite disparate but the distribution remains normal. Also when α_1 is less than 1 it results in a higher mean number of customers in the system compared to when it is more than one.

As in earlier in Scenario B in Fig. 3.6, 3.7 and 3.8 the service time distribution of the *Cases* is the following(Table 3.9).

Presently both inter-arrival and retrial distribution changes as α is increasing which results higher values of mean waiting time (Fig. 3.10). With the exception

Table 3.9: Parameters of service time

Case	α	β
1	0.5	0.5
2	1	1
3	2	2

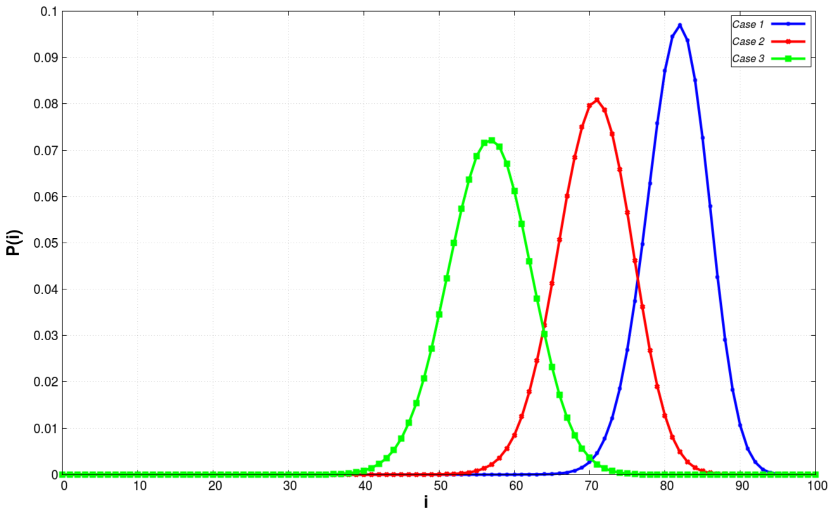

Figure 3.9: Steady-state distributions of Scenario C

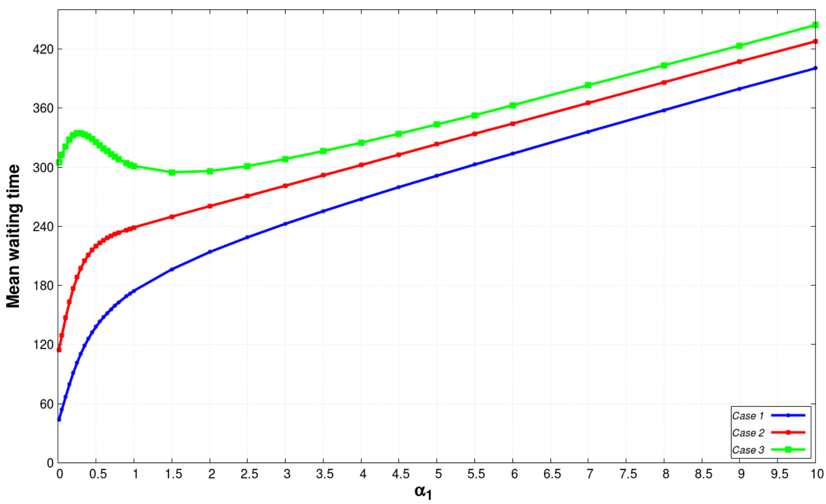

Figure 3.10: Mean waiting time vs. shape parameter

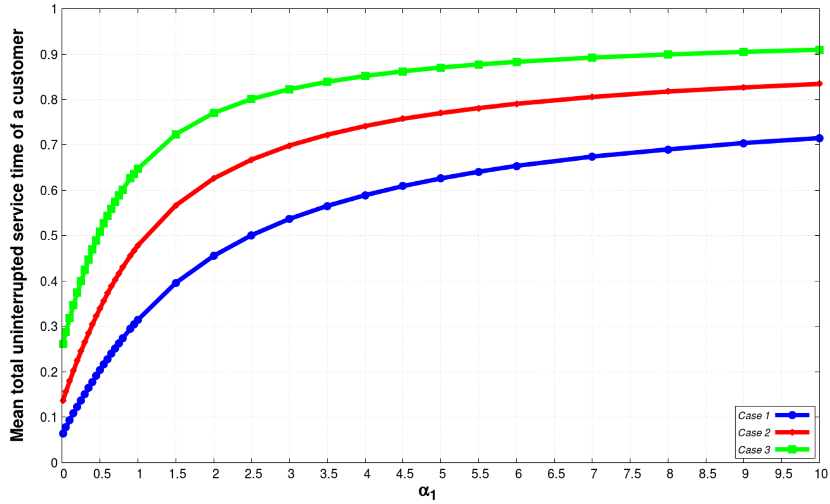

Figure 3.11: Mean total uninterrupted service time of a customer vs. shape parameter

of *Case 3* the others are monotonically increasing. This is true for the mean successful service time (Fig. 3.11) but now *Case 3* rises from the beginning.

The last figure shows the effect of the inter-arrival and inter-request time on the mean total interrupted service time of a customer. At the beginning the difference is quite high among the *Cases* especially *Case 3* has high values

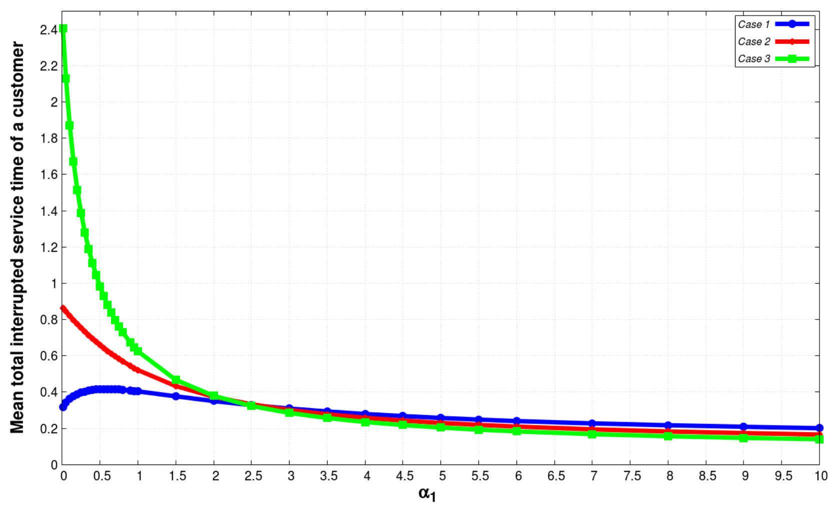

Figure 3.12: Mean total interrupted service time of a customer time vs. shape parameter

compared to the others but it is valid till α_1 is smaller than 1 then the values of the *Cases* are nearly identical.

3.3 Comparison of two operation modes in case of server failure

The same model is regarded in this section as in Section 3.2 with a little but essential modification. Two operation modes are considered in the case of server breakdown while it is occupied [J2]. Namely, these modes are the following:

- The interrupted request gets into the orbit instantaneously (same as in the previous section).
- The service of the interrupted request is suspended and it continues after repairing the server.

When the submission (the service of a request) is successful, the request goes back to the source. Similarly to the previous section, all the random variables involved in the model construction are assumed to be totally independent of each other. The novelty of this investigation is a comparison of the performance measures of these modes, and estimations obtained by the simulation are graphically illustrated showing the influence of the difference of the working modes on the mean and variance of the response time, mean and variance of the number of customers in the system, mean and variance of the sojourn time in the orbit, mean and variance of time a customer spent in service.

3.3.1 Scenarios

Scenario A: The following table shows the input parameters of Scenario A (see Table 3.10).

Table 3.10: Numerical values of model parameters

Case	N	λ/N	γ_0	γ_1	γ_2	σ/N	α	β
1	100	0.01	0.1	0.1	1	0.01	0.5	0.5
2	100	0.01	0.1	0.1	1	0.01	1	1
3	100	0.01	0.1	0.1	1	0.01	2	2

Figure 3.13 shows the steady-state distribution of the cases under consideration. *Case 2* is a special case because it yields the exponential distribution when the

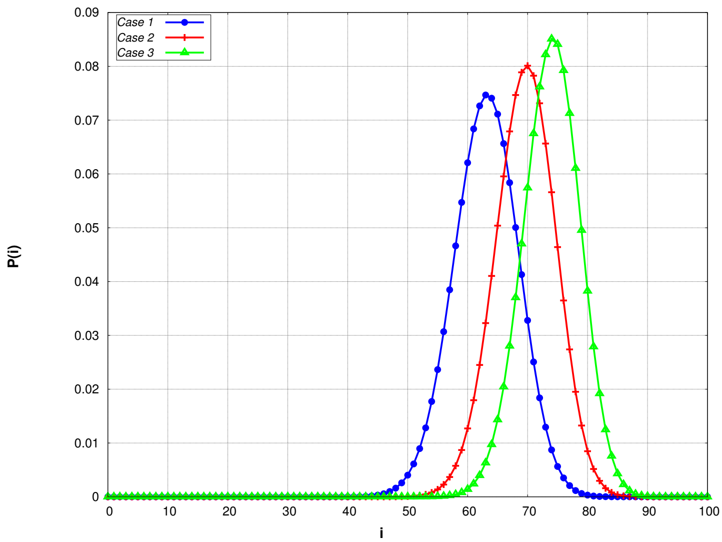

Figure 3.13: Comparison of steady-state distributions

Table 3.11: Numerical results of Scenario A

Case	E(NS)	$Var(NS)$	E(T)	$Var(T)$	E(W)	$Var(W)$	E(S)	$Var(S)$
1	63.0526	28.3198	170.5618	63092.8264	169.7553	62855.9266	0.3357	0.2255
2	69.6114	24.6949	228.9834	97974.6274	227.8834	97613.9701	0.5022	0.2523
3	73.9099	21.9244	283.1452	136396.9409	281.7728	135909.2373	0.6688	0.2238

parameter α is equal to 1. The mean number of customers increases with the increment of parameter α and β and the curves of the steady-state distribution follow the normal law. The following Table presents the considered performance measures in relation with the different cases (see Table 3.11).

The following three figures (Figure 3.14, 3.15 and 3.16) compare the mean waiting time of the two different operation modes in the *Cases* under consideration. Operation mode No. 1 corresponds to the mode when the interrupted requests get into the orbit instantaneously and under operation mode No. 2 we consider the mode when the service of the interrupted request is suspended and it continues after the server is repaired. In all cases, the results confirm the expectation that operation mode No. 2 results in a lower mean waiting time. When the values of parameter α and β are higher the difference between the modes is higher as well. With an increase in the arrival intensity, we should expect longer waiting times; however, after λ/N reaches 0.15, it begins to decrease monotonically, which is an interesting phenomenon.

The results for the mean total uninterrupted service time of a customer and service of gamma distribution for $\alpha = \beta = 0.5; 1; 2$ are shown in Figure 3.17,

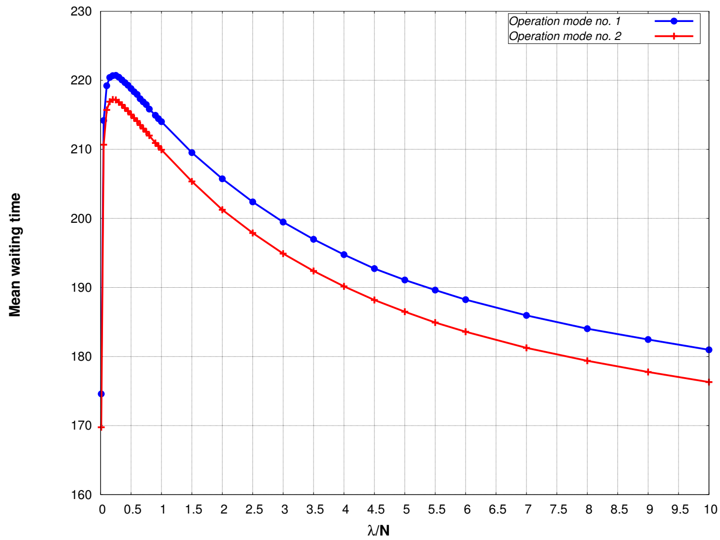

Figure 3.14: Mean waiting time vs. intensity of incoming customers of *Case 1*

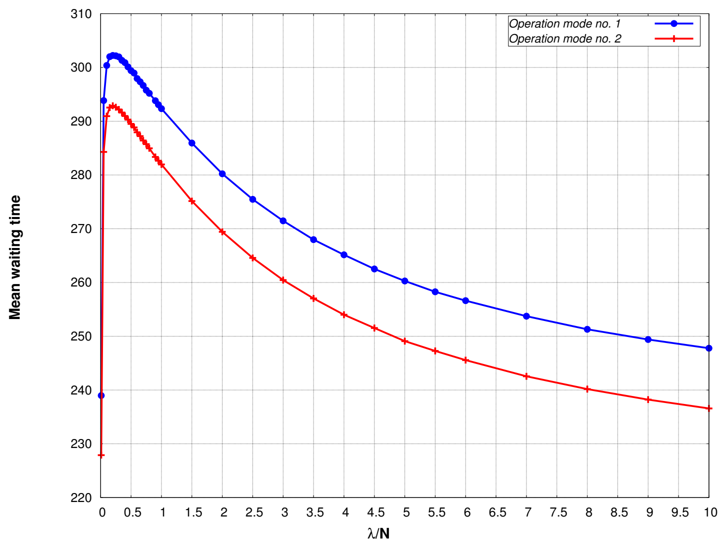

Figure 3.15: Mean waiting time vs. intensity of incoming customers of *Case 2*

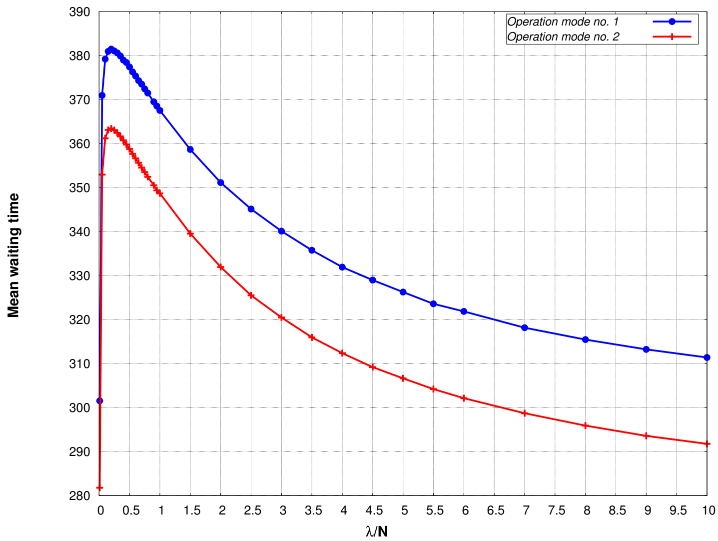

Figure 3.16: Mean waiting time vs. intensity of incoming customers of *Case 3*

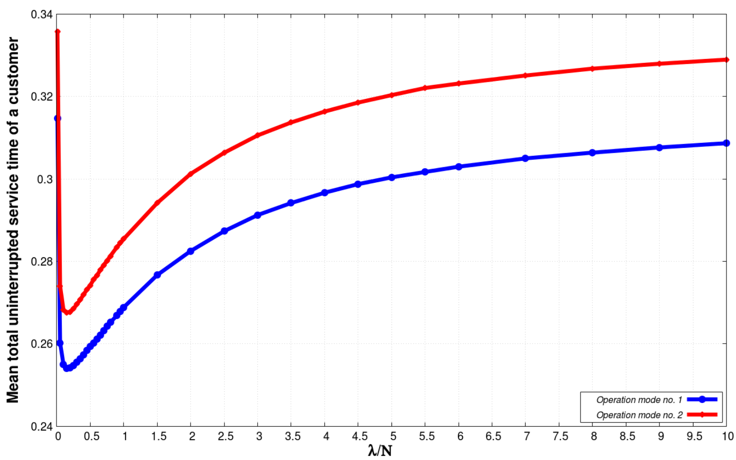

Figure 3.17: Mean total uninterrupted service time of a customer vs. intensity of incoming customers of *Case 1*

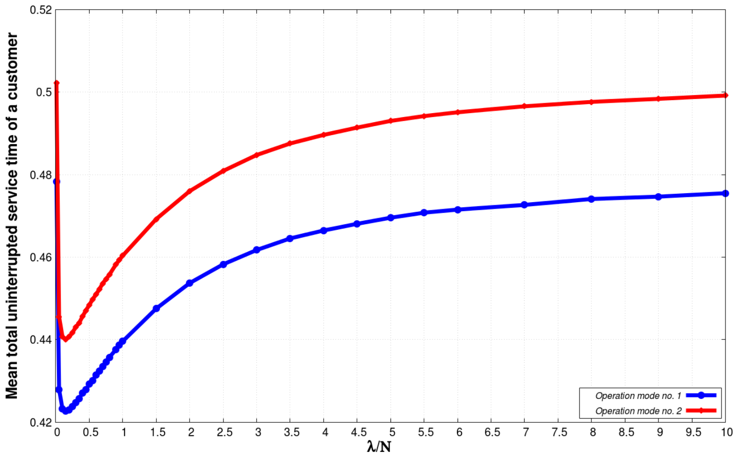

Figure 3.18: Mean total uninterrupted service time of a customer vs. intensity of incoming customers of *Case 2*

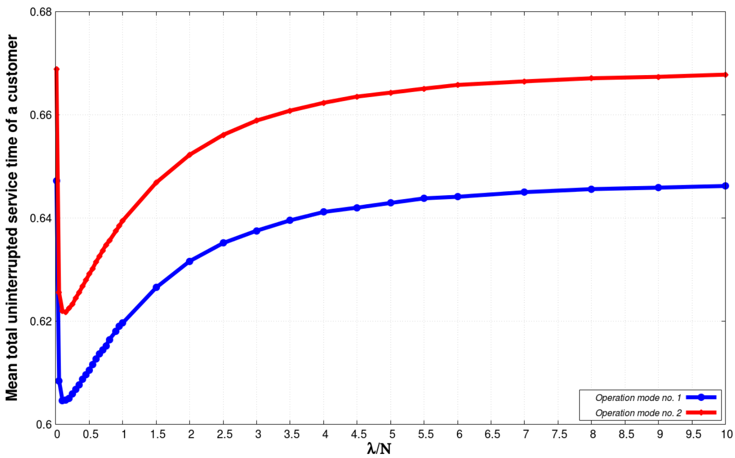

Figure 3.19: Mean total uninterrupted service time of a customer vs. intensity of incoming customers of *Case 3*

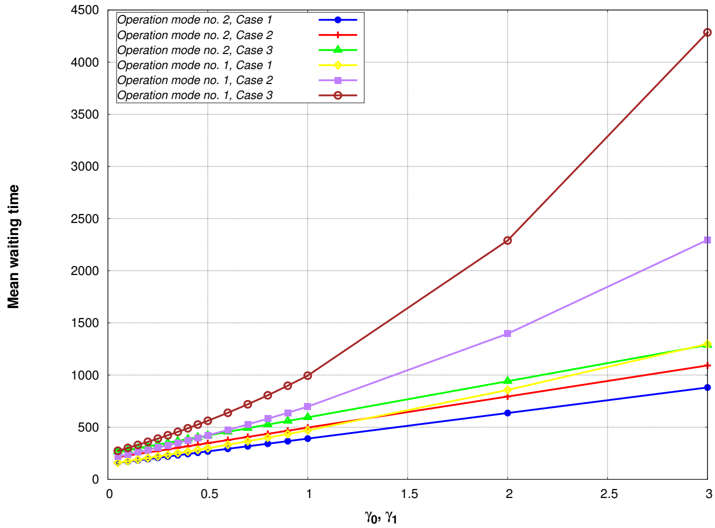

Figure 3.20: Mean waiting time vs. intensity of failure rate

3.18, 3.19. For any value of α and β using this parameter settings we experience minimum property characteristics. The mean total uninterrupted service time of a customer tends to be higher when α and β take higher values. In all *cases* using operation mode number 2 results higher amount of mean total uninterrupted service time which we expect to be true. Because of the effect of collisions, this mean value is not near $\frac{\alpha}{\beta}$ which would be the mean service time.

Figure 3.20 shows the mean waiting time as a function of the failure rate. As expected, the mean waiting time increases as the failure rate becomes higher in all *Cases*. The difference between the operation modes is quite obvious. While using operation mode No. 2 the growth seems linear the increment is drastic in case of operation mode No. 1. In *Case 3* the rise is quite significant.

Scenario B: In Scenario B the distribution of inter-arrival times of the customers is gamma distributed with parameter λ_1 and β_1. The following Table (see Table 3.12) presents the numerical values of parameters of Scenario B.

Table 3.12: Numerical values of parameters of Scenario B

Case	N	α	β	γ_0	γ_1	γ_2	σ/N	α_1	β_1/N
1	100	1	1	0.1	0.1	1	0.01	0.5	0.01
2	100	1	1	0.1	0.1	1	0.01	1	0.01
3	100	1	1	0.1	0.1	1	0.01	2	0.01

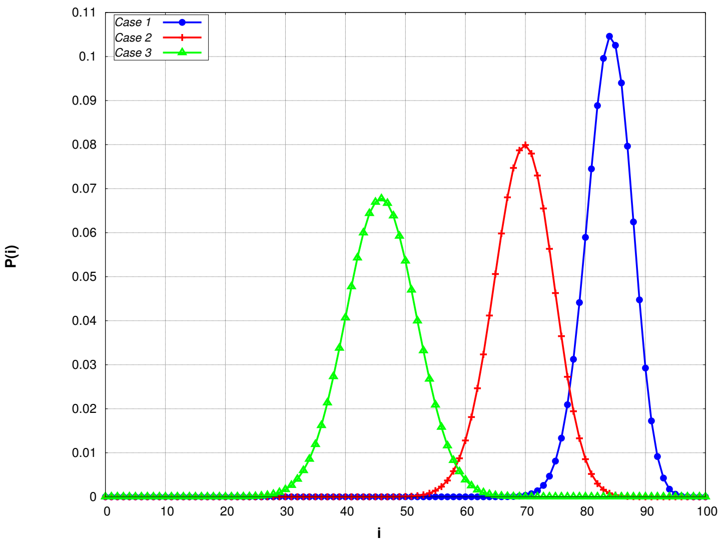

Figure 3.21: Steady-state distributions of Scenario B

Fig. 3.21 displays the steady-state distributions under Scenario B. Now the service time is supposed to be exponentially distributed ($\alpha = 1$). With this modification compared to Scenario A the steady-state distributions still follow a normal distribution, and as the value of α_1 increases, the mean number of requests in the system decreases. In *Case 3* the mean number of customers in the system is significantly lower among the *Cases*. In Table 3.13 an estimate for basic performance measures can be found in connection with the cases.

Table 3.13: Numerical results of Scenario B

Case	E(NS)	$Var(NS)$	E(T)	$Var(T)$	E(W)	$Var(W)$	E(S)	$Var(S)$
1	83.856	14.4707	259.7716	121328.2493	258.6719	120904.8333	0.4707	0.2232
2	69.5984	24.7715	228.9076	97997.6219	227.8074	97637.2906	0.5025	0.2525
3	45.9508	34.3607	170.0134	62628.4067	168.9133	62379.0441	0.5806	0.337

The running parameter is α_1 which help us to discover the impact of different distributions on the various examined performance measures. Fig. 3.22, 3.23 and 3.24 show a comparison of the mean waiting time between the two operation modes of the investigated *Cases*. As in Scenario A using operation mode No. 2 the interrupted requests stay at the server in case of server failure and their services continue after the server is ready to process jobs again consequently ensures lower mean waiting times. In all the cases the mean waiting time starts to increase till α_1 reaches 0.3, after which it monotonically decreases. With higher values of α and β, the differences between the operation modes are higher as well.

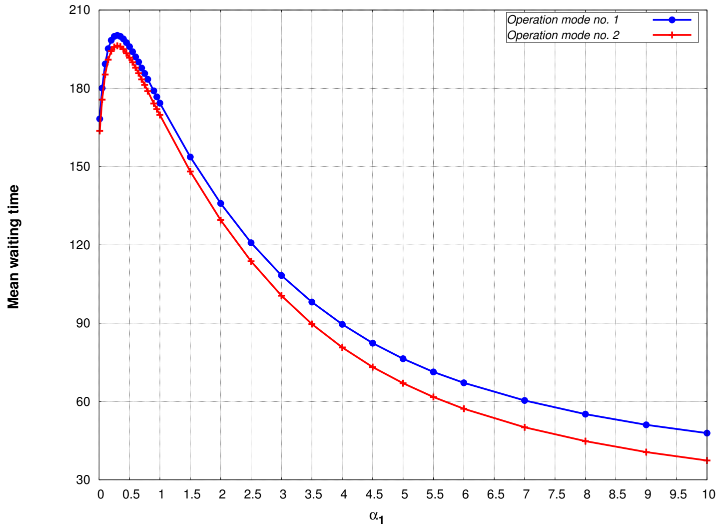

Figure 3.22: Mean waiting time vs. shape parameter, $\alpha = \beta = 0.5$

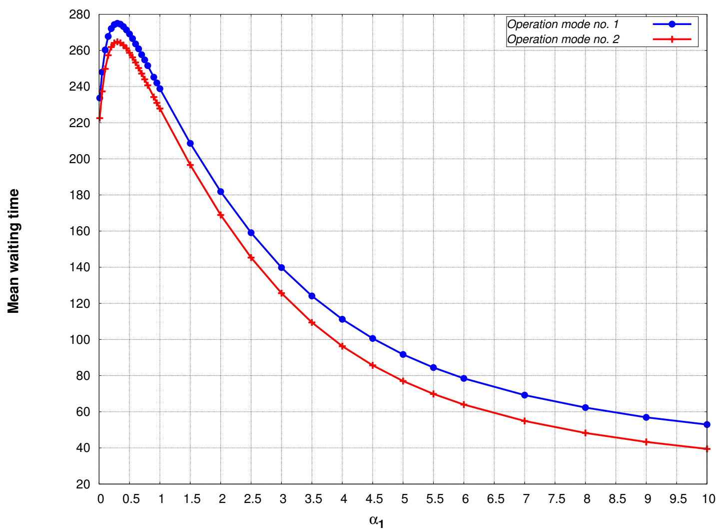

Figure 3.23: Mean waiting time vs. shape parameter, $\alpha = \beta = 1$

Fig. 3.25, 3.26 and 3.27 show the mean total uninterrupted service time of a customer vs. the shape parameter of the inter-arrival time using both operation mode. As we can see, in Scenario A here we get what we expected: the use of operation mode No. 2 provides greater values of mean total uninterrupted service time of a customer. The difference between the applied operation modes

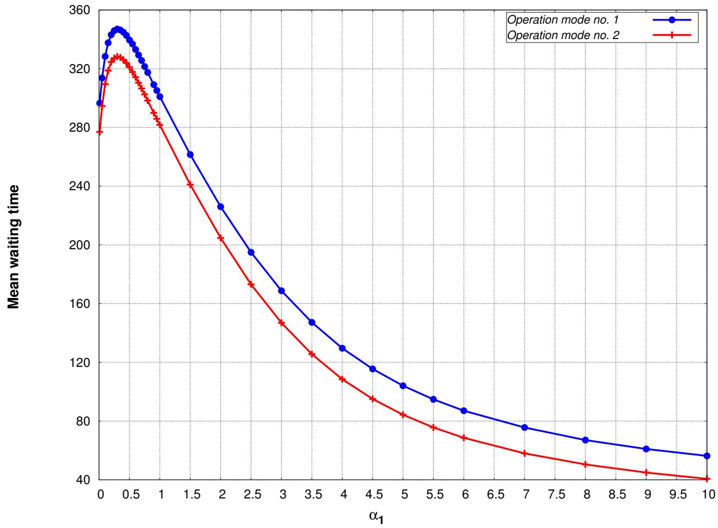

Figure 3.24: Mean waiting time vs. shape parameter, $\alpha = \beta = 2$

is quite large in all cases, especially if α and β are equal to 0.5. The mean total uninterrupted service time of a customer behaves inversely as compared to the mean waiting time, because while the mean waiting time of a customer increases, the mean total uninterrupted service time of a customer time decreases, and vice versa.

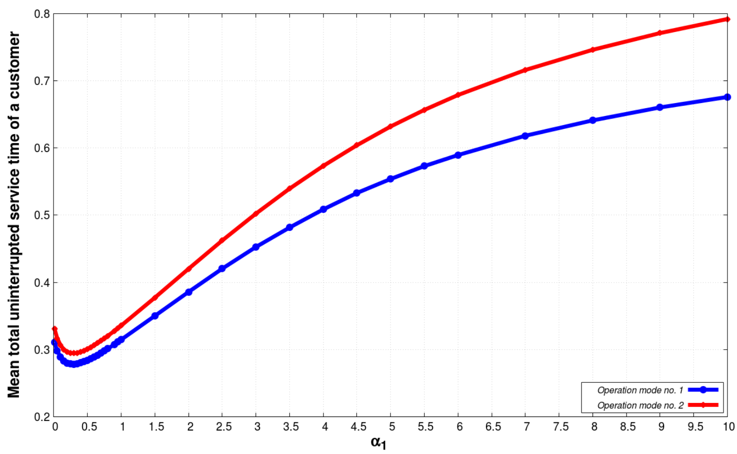

Figure 3.25: Mean total uninterrupted service time of a customer vs. shape parameter, $\alpha = \beta = 0.5$

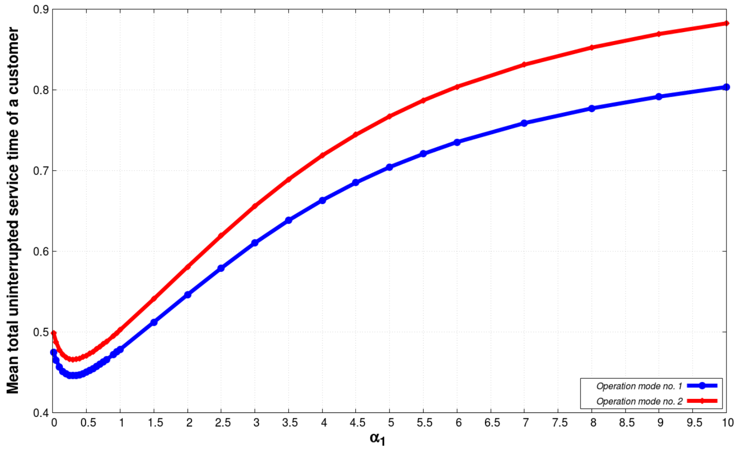

Figure 3.26: Mean total uninterrupted service time of a customer vs. shape parameter, $\alpha = \beta = 1$

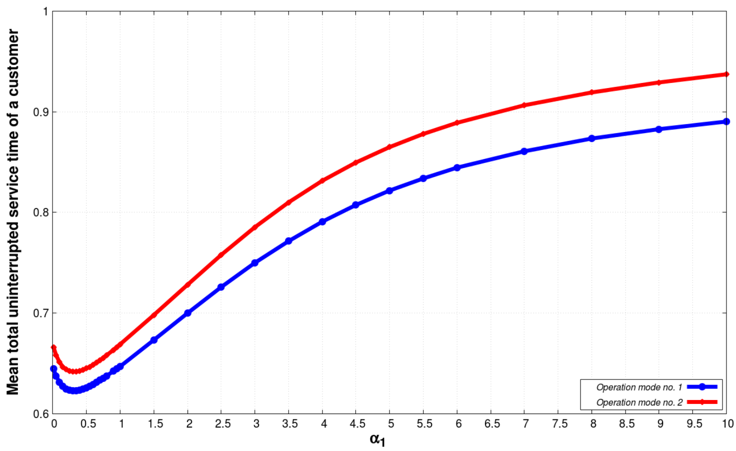

Figure 3.27: Mean total uninterrupted service time of a customer vs. shape parameter, $\alpha = \beta = 2$

3.4 Finite-source retrial queuing systems with collision and blocking

In this section, we also deal with a retrial queueing system with one server [J6]. The number of sources of calls is finite and collisions can take place. We assume that the failure of the server blocks the system's operation such that newly arriving customers cannot enter the system, contrary to an earlier section where the failure does not affect the arrivals. The novelty of this analysis is the inspection of the blocking effect on the performance measures using different distributions.

3.4.1 System model

The operation of the model can be seen in Figure 3.28. The same model is considered as in the previous two sections (Section 3.2, 3.3) but with a minor modification which basically changes the process of the system.

When a busy server breaks down the interrupted customer is transmitted to the orbit. Two types of operation behaviour are distinguished in the case of a server failure:

- Without blocking: requests are able to enter the system and these are forwarded to the orbit instantaneously.
- With blocking: the arriving customers cannot enter the system; they return to the source and a new request generation starts.

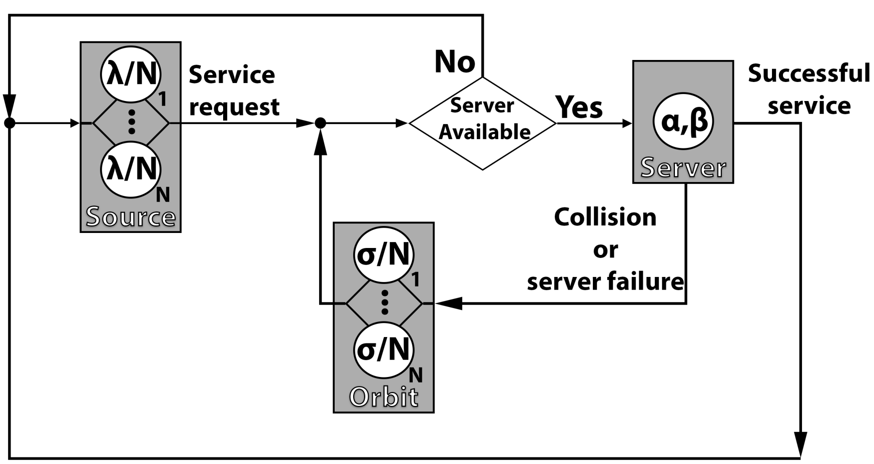

Figure 3.28: System model

3.4.2 Obtained results

The applied values of the input parameters are presented in Table 3.14.

Table 3.14: Numerical values of model parameters

Case studies								
No.	N	λ/N	γ_0	γ_1	γ_2	σ/N	α	β
Fig. 3.29	100	0.01	0.1	0.1	1	0.01	0.5,1,2	0.5,1,2
Fig. 3.30	100	0.01..6	0.1	0.1	1	0.01	0.5,1,2	0.5,1,2
Fig. 3.31	100	0.01..6	0.1	0.1	1	0.01	0.5	0.5
Fig. 3.32	100	0.01..6	0.1	0.1	1	0.01	1	1
Fig. 3.33	100	0.01..6	0.1	0.1	1	0.01	2	2
Fig. 3.34	100	0.01	0.05..0.5	0.05..0.5	1	0.01	0.5	0.5
Fig. 3.35	100	0.01	0.05..0.5	0.05..0.5	1	0.01	1	1
Fig. 3.36	100	0.01	0.05..0.5	0.05..0.5	1	0.01	2	2
Fig. 3.37	100	0.01	0.05..0.5	0.05..0.5	1	0.01	0.5	0.5
Fig. 3.38	100	0.01	0.05..0.5	0.05..0.5	1	0.01	0.5	0.5
Fig. 3.39	100	0.01	0.05..0.5	0.05..0.5	1	0.01	1	1
Fig. 3.40	100	0.01	0.05..0.5	0.05..0.5	1	0.01	2	2

Tables 3.15 and 3.16 indicate the numerical results of the main performance measures comparing the two possible operation methods whenever the server breaks down.

Table 3.15: Numerical results when blocking is not applied

Case	E(NS)	Var(NS)	E(T)	Var(T)	E(W)	Var(W)	E(S)	Var(S)
1	63.6842	27.9734	175.3073	65657.3454	174.5884	65434.6696	0.3147	0.1979
2	70.5912	24.3012	239.9734	105273.4267	238.9734	104918.6389	0.4784	0.2289
3	75.1825	21.2439	302.8106	151781.1411	301.5377	151277.6006	0.6472	0.2095

Table 3.16: Numerical results when blocking is applied

Case	E(NS)	Var(NS)	E(T)	Var(T)	E(W)	Var(W)	E(S)	Var(S)	E(IS)
1	60.0320	28.5106	165.1955	63510.5573	164.4769	63287.935	0.3146	0.1977	0.4040
2	67.64	25.1115	229.8805	103125.2009	228.8805	102770.5034	0.4783	0.2288	0.5217
3	72.69	22.4004	292.6778	149545.4531	291.4051	149042.3599	0.6471	0.2095	0.6256

In Figure 3.29 the steady-state distribution of the investigated cases are represented when blocking is applied. The parameters of the service time of Case 2 are unique because for $\alpha = 1$ we have the exponential distribution. When the values of parameters α and β increase, this results in a higher mean number of customers, although the mean of the gamma distribution is the same in all cases ($\alpha/\beta = 1$). Looking carefully at Figure 3.29 it can be observed that all

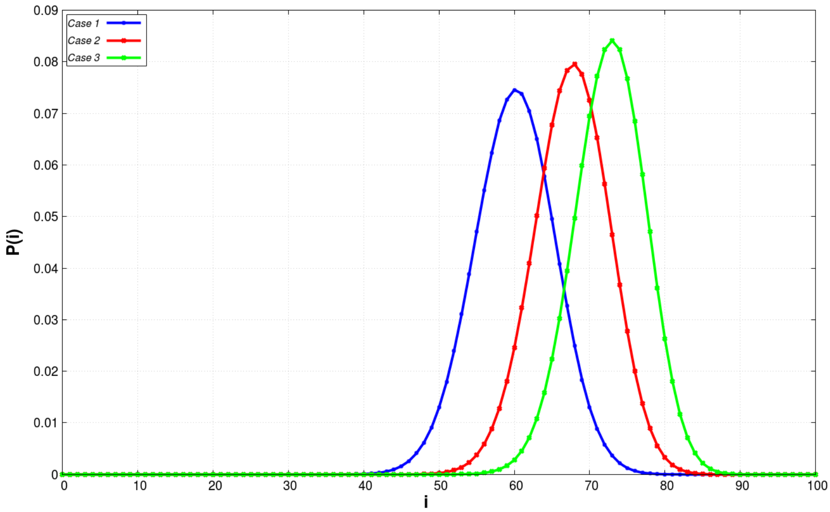

Figure 3.29: Comparison of steady-state distributions

cases correspond to the normal distribution. Figure 3.30 displays the mean waiting time of the customers as a function of the incoming generation rate when blocking is applied. Consequently, because the mean number of customers is higher in Cases 2 and 3 compared to Case 1 the mean waiting time is greater, too. It is a specialty that, when using retrial queues with finite-source, the mean waiting time has a maximum value. This is a general characteristic of retrial queues but depends on the parameter settings used.

In Figures 3.31, 3.32, and 3.33 the two operation methods are compared when the server is down. These figures ensure the expected behavior when blocking is applied, which results in a lower mean waiting time. This can be explained by the fact that customers cannot enter the system in case of server failure, such that these requests are rejected and directed back towards the source. These figures also show, besides higher values of α and β, that the difference between the operation methods is smaller. In Figure 3.33 the results are almost identical.

Figures 3.34, 3.35 and 3.36 present the mean spent time in the source of the customers as a function of the server failure rates. It may be seen, the results are equal, which indicates that the distribution of the service time has no effect on the mean spent time in the source of customers. The proposed behavior is noticeable in the case of blocking: the customers spend more and more time in the source if the server is more likely subjected to breakdowns. The mean spent time in the source of the customers is constant when requests can enter the system and the server does not function.

Figure 3.37 shows the average number of attempts of the customers to enter the

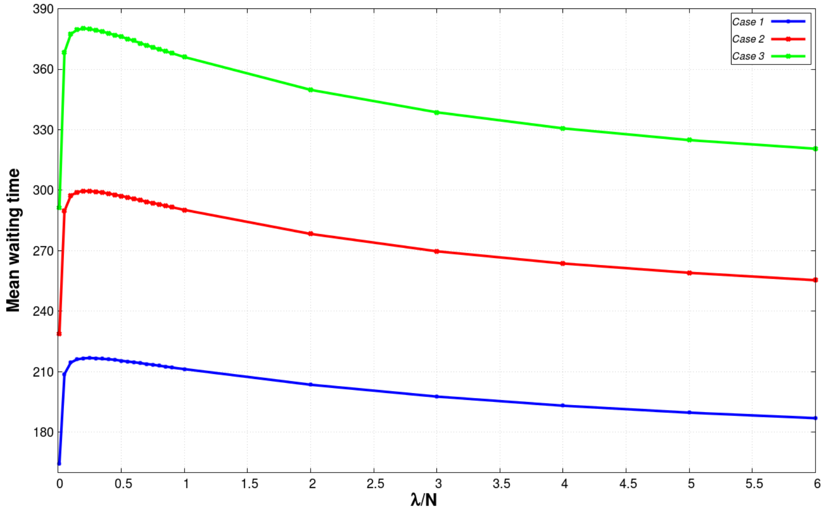

Figure 3.30: Mean waiting time vs. intensity of incoming customers

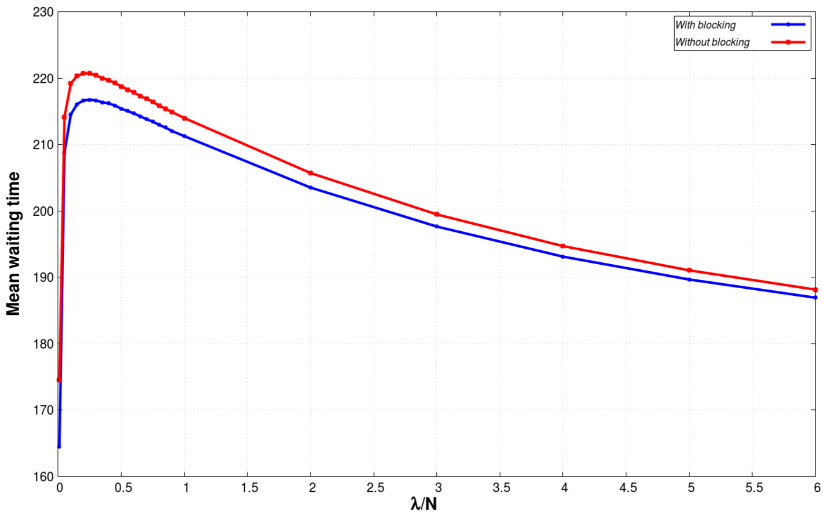

Figure 3.31: The effect of blocking, the parameters of service time: $\alpha = \beta = 0.5$

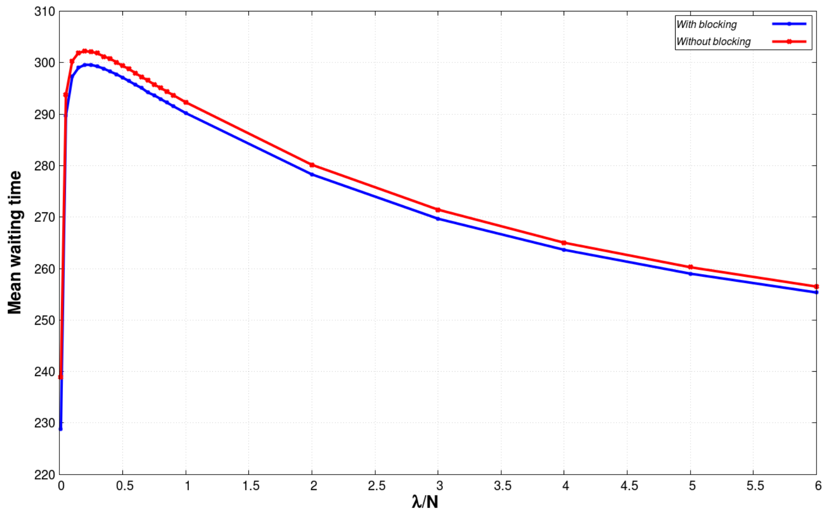

Figure 3.32: The effect of blocking, the parameters of service time: $\alpha = \beta = 1$

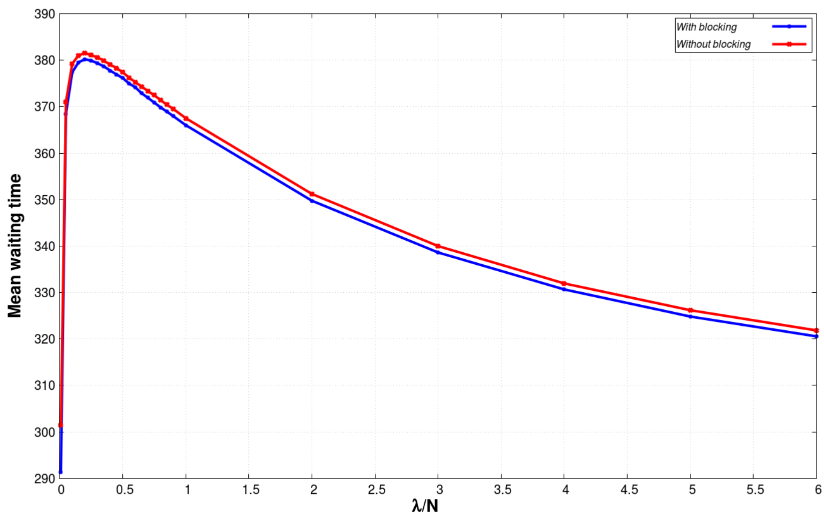

Figure 3.33: The effect of blocking, the parameters of service time: $\alpha = \beta = 2$

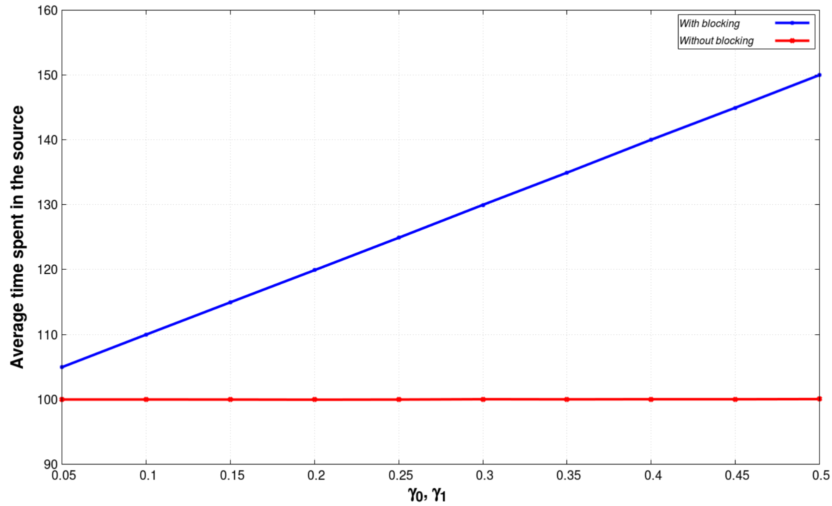

Figure 3.34: The effect of blocking, the parameters of service time: $\alpha = \beta = 0.5$

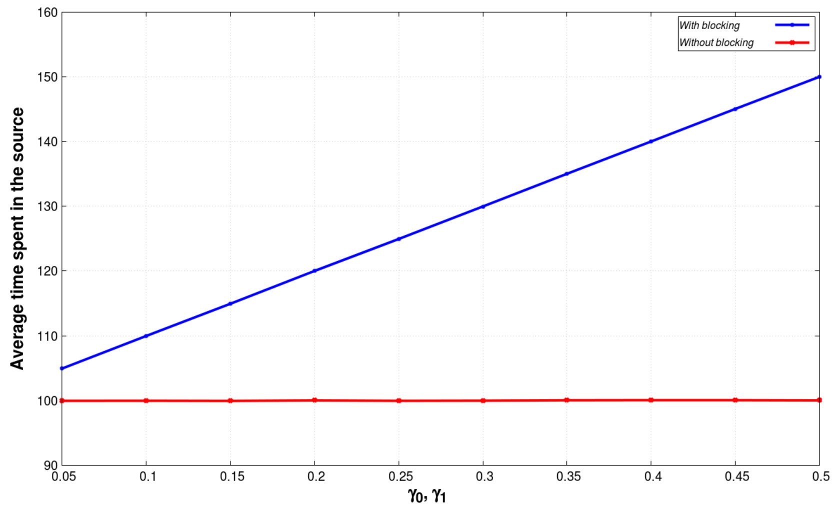

Figure 3.35: The effect of blocking, the parameters of service time: $\alpha = \beta = 1$

system. Consequently, if the service unit spends more time in a state of failure, then customers will be rejected more and more times when blocking is applied. Without blocking the customer needs to appear once at the system and it can enter irrespectively of the availability of the server. In the case of blocking, it is not surprising that by increasing parameters γ_0 and γ_1 the average number of attempts increases as well.

Figures 3.38, 3.39 and 3.40 demonstrate the mean waiting time of the customers

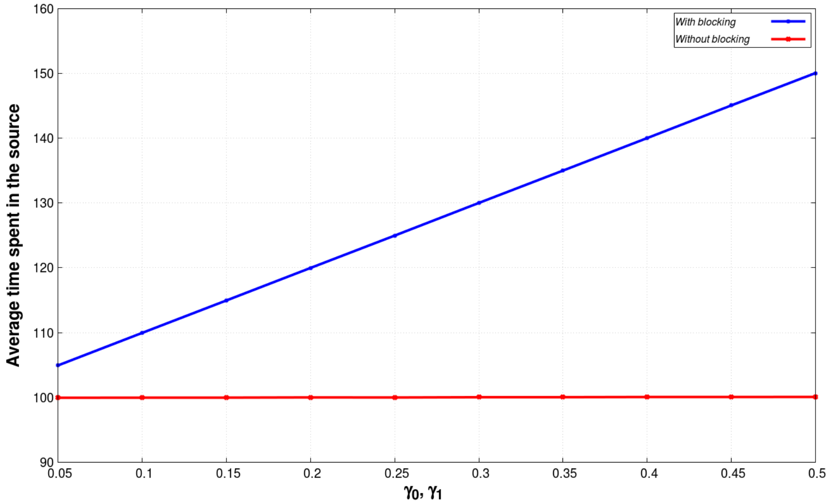

Figure 3.36: The effect of blocking, the parameters of service time: $\alpha = \beta = 2$

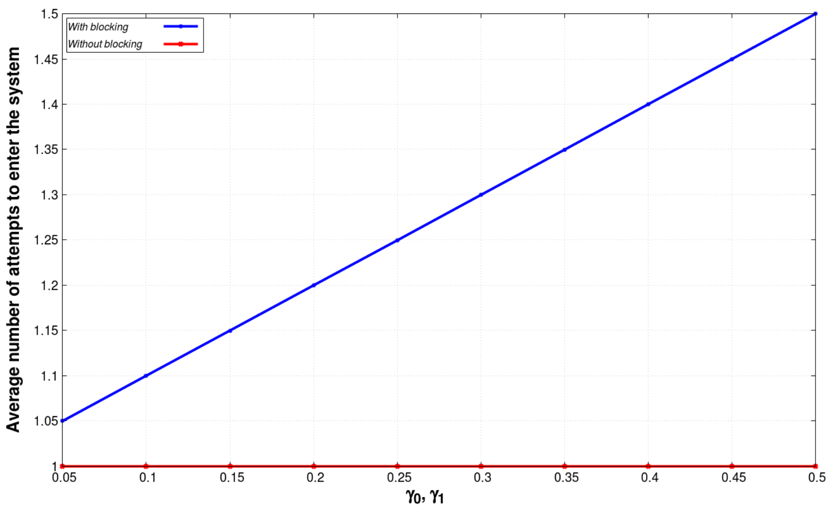

Figure 3.37: The effect of blocking, the parameters of service time: $\alpha = \beta = 0.5$

as a function of γ_0 and γ_1. In all of the three figures, higher parameters of server failure rates result in a greater mean waiting time in both operation methods, but especially in the case of non-blocking. A significant difference can be noted among the values of the mean waiting time using various sets of parameters for the service time. An increasing tendency is observed with the increment of values of α and β. The same phenomenon takes place as in the previous figures in connection with the mean waiting time that when $\alpha = \beta = 2$ the cases of

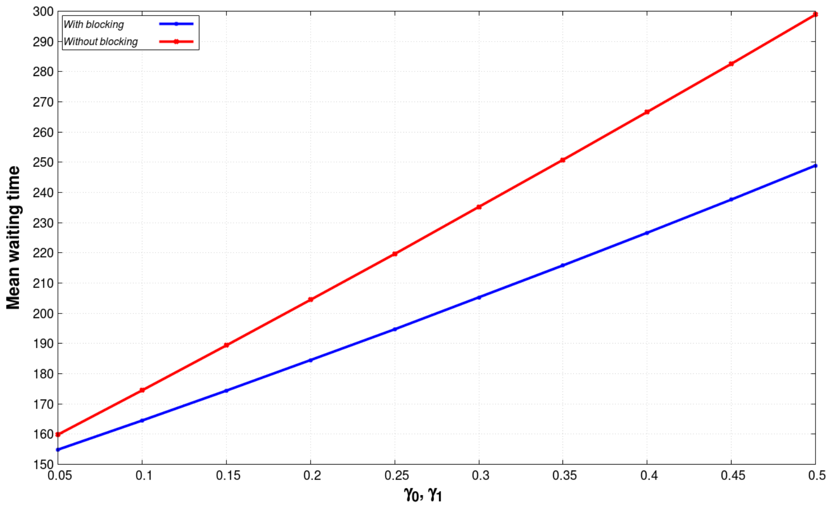

Figure 3.38: The effect of blocking, the parameters of service time: $\alpha = \beta = 0.5$

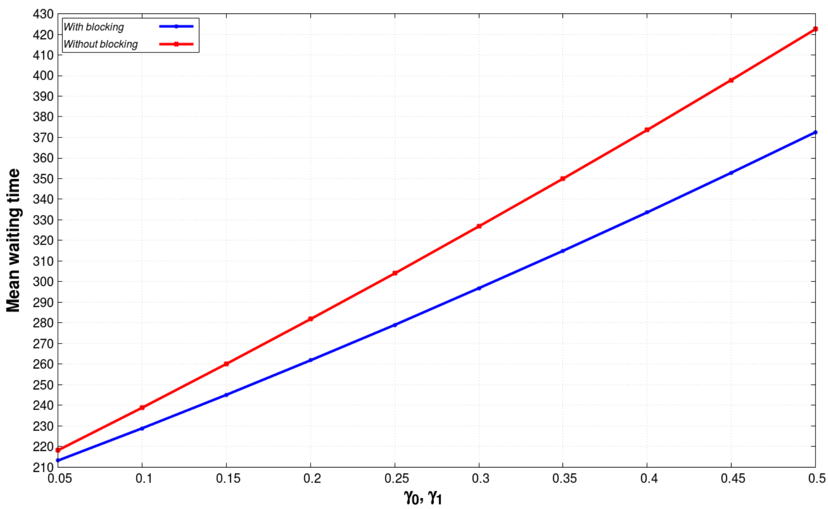

Figure 3.39: The effect of blocking, the parameters of service time: $\alpha = \beta = 1$

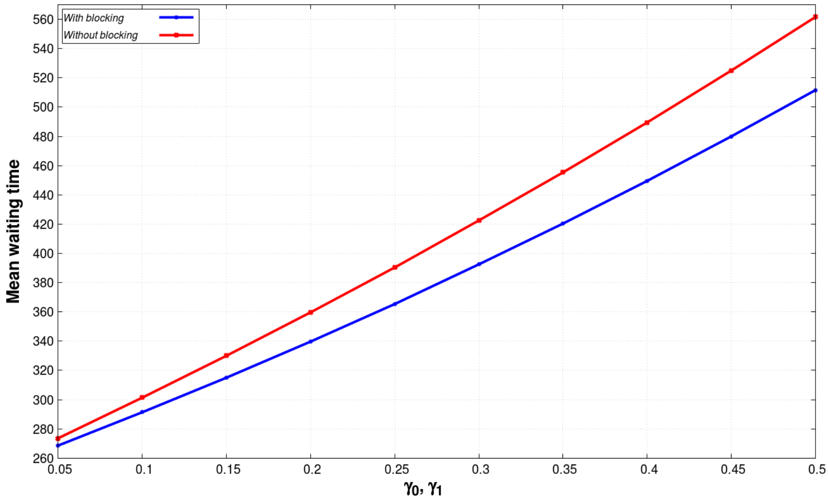

Figure 3.40: The effect of blocking, the parameters of service time: $\alpha = \beta = 2$

non-blocking and blocking are closer to each other compared to $\alpha = \beta = 0.5$.

4

Two-way communication systems

This chapter presents the results in connection with retrial queuing systems by the help of two-way communication

Contents

4.1 Finite-source retrial queues with two-way communications to the orbit

In this section, we examine a retrial queueing system of type $M/G/1//N$ with a reliable service unit that is capable to produce outgoing calls to the customers residing in the orbit [J4]. N customers are located in the source, where all of them can generate incoming, primary calls towards the server. The distribution of the inter-request times is exponential with a rate of λ/N. In default of waiting queue, an incoming customer either from the source or orbit finds the server in an idle state then its service begins instantly. The service times of incoming customers are assumed to be gamma, hypo-exponentially, hyper-exponentially, Pareto, and lognormal distributed with different parameters but with the same mean value. Customers return to the source after their service is terminated. If the server is busy, meaning that a request is under service, an incoming customer remains in the system and enters into the orbit. Customers located in the orbit are able to attempt to access the server again after an exponentially distributed time with parameter σ/N. On the other hand, when the server becomes idle it can make an outgoing call towards the customers in the orbit. It is performed after an exponentially distributed time with parameter ν. The service time of these outgoing customers follows gamma distribution with parameters α_2 and β_2. All the random variables involved in the model construction are assumed to be totally independent of each other.

4.1.1 Applied distributions and its parameters

In this Section, the reader gets an insight into the parameters of the applied distributions and the process of how to select them to execute a valid comparison. To do so our program is integrated with random number generators according to gamma, hyper-exponential, hypo-exponential, lognormal and Pareto distribution. These random number generators need input parameters that are different in every distribution, thus parameter selection is crucial. For a valid comparison, the goal is to achieve the same mean and variance in case of every distribution hence we take over every distribution and how the fitting process is accomplished.

4.1.1.1 Gamma distribution

Gamma distribution is a general type of statistical distribution and a random variable X has a gamma distribution if its density function is the following:

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{\beta(\beta x)^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} & \text{if } x \geq 0 \end{cases}$$

where $\beta > 0$ and $\alpha > 0$.

$$\Gamma(\alpha) = \int_0^\infty \mathrm{t}^{\alpha-1}\,\mathrm{e}^{-t}\,\mathrm{d}t$$

This is the so-called complete gamma function, which has two parameters: α is called the shape parameter and β is called the scale parameter. These two parameters are also the input parameters of the random number generator.

The coefficient $C_X^2 = \frac{Var(X)}{(EX)^2}$ is defined as the squared coefficient of variation of random variable X.

The mean value, variation, and the squared coefficient of variation can be calculated:

$$\overline{X} = \frac{\alpha}{\beta}, \qquad Var(X) = \frac{\alpha}{\beta^2}, \qquad C_X^2 = \frac{1}{\alpha}$$

For a predetermined mean value and variance to obtain parameters α and β the next calculation has to be done:

$$\alpha = \frac{1}{C_X^2}$$
$$\beta = \frac{\alpha}{\overline{X}}$$

4.1.1.2 Pareto distribution

A random variable X has a Pareto distribution if its density function is the following:

$$f(x) = \begin{cases} 0 & \text{if } x < k \\ \alpha k^{\alpha} x^{-\alpha-1} & \text{if } x \geq k \end{cases}$$

And the distribution function is:

$$F(x) = \begin{cases} 0 & \text{if } x < k \\ 1 - \left(\frac{k}{x}\right)^{\alpha} & \text{if } x \geq k \end{cases}$$

where $\alpha, k > 0$.

It has two parameters: α is called the shape parameter and k is called the location parameter. These two parameters are the input parameters of the random number generator.

The mean value, variation, and the squared coefficient of variation can be calculated:

$$\overline{X} = \begin{cases} \frac{k\alpha}{\alpha-1} & \text{if } \alpha > 1 \\ \infty & \text{if } \alpha \leq 1 \end{cases}$$

$$Var(X) = \frac{k^2\alpha}{\alpha - 2} - \left(\frac{k\alpha}{\alpha - 1}\right)^2$$

$$C_X^2 = \frac{(\alpha - 1)^2}{\alpha(\alpha - 2)} - 1, \qquad \alpha > 2.$$

For a predetermined mean value and variance to obtain parameters α and k the following interrelation is used:

$$\alpha = 1 + \frac{\sqrt{1 + C_X^2}}{\sqrt{C_X^2}}$$

$$k = \frac{\alpha - 1}{\alpha} \times \overline{X}$$

4.1.1.3 Lognormal distribution

Let $Y \in N(m, \sigma)$ a random variable with normal distribution, lognormal is a continuous distribution in which the logarithm of a variable having a normal distribution, namely $X = e^Y$ has lognormal distribution with parameters (m, σ). Its distribution and density function are the following:

$$F_x(x) = \Phi\Big(\frac{\ln(x) - m}{\sigma}\Big), \qquad x > 0.$$

$$f_x(x) = \frac{1}{\sigma x}\Phi\Big(\frac{\ln(x) - m}{\sigma}\Big), \qquad x > 0.$$

The mean value, variation, and the squared coefficient of variation can be calculated:

$$\overline{X} = e^{m+\frac{\sigma^2}{2}}, \qquad Var(X) = e^{2m+\sigma^2}(e^{\sigma^2} - 1)$$

$$C_X^2 = e^{\sigma^2} - 1.$$

To obtain the two parameters of the lognormal distribution the following interrelation is applied:

$$\sigma = \sqrt{\ln(1 + C_X^2)}$$
$$m = \ln(\overline{X}) - \frac{\sigma^2}{2}$$

4.1.1.4 Hypo-exponential distribution

Continuous statistical distribution, let $X_i \in Exp(\mu_i)(i = 1, ..., n)$ be independent exponentially distributed random variables. Then $Y_n = X_1 + ... + X_n$ has n-phase hypo-exponential distribution. Its density function is given by

$$f_{Y_n}(x) = \begin{cases} 0 & \text{if } x < 0 \\ (-1)^{n-1}\left[\prod_{i=1}^{n}\mu_i\right]\sum_{j=1}^{n}\frac{e^{-\mu_j x}}{\prod_{k=1,k\neq j}^{n}(\mu_j-\mu_k)} & \text{if } x \geq 0. \end{cases}$$

The mean value, variation, and the squared coefficient of variation can be calculated:

$$\overline{Y_n} = \sum_{i=1}^{n}\frac{1}{\mu_i}, \qquad Var(Y_n) = \sum_{i=1}^{n}\frac{1}{\mu_i^2}$$

$$C_{Y_n}^2 = \frac{\sum_{i=1}^{n}\left(\frac{1}{\mu_i}\right)^2}{\left(\sum_{i=1}^{n}\frac{1}{\mu_i}\right)^2}.$$

In our simulation program, we used the 2-phase hypo-exponential distribution where the parameters are the parameters of the two independent exponential distribution (μ_1, μ_2). For a predetermined mean value and variance to obtain parameters μ_1 and μ_2 the next equation system has to be solved:

$$\overline{X} = \frac{1}{\mu_1} + \frac{1}{\mu_2}$$
$$Var(X) = \frac{1}{\mu_1^2} + \frac{1}{\mu_2^2}$$

4.1.1.5 Hyper-exponential distribution

Suppose $X_1, X_2, \cdots, X_n$ are independent exponential random variables, where the rate parameter of X_i is λ_i. The random variable X can be one of the n independent exponential random variables $X_1, X_2, \cdots, X_n$ such that X is X_i with probability p_i with $p_1 + \cdots + p_n = 1$. Such a random variable X is said to follow a hyper-exponential distribution. Its density function is given by

$$f_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ \sum_{i=1}^{n} p_i\lambda_i e^{-\lambda_i x} & \text{if } x \geq 0. \end{cases}$$

Its distribution function is

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - \sum_{i=1}^{n} p_i e^{-\lambda_i x} & \text{if } x \geq 0. \end{cases}$$

In the case when for a random variable $X, C_X^2 > 1$ then the following two-moment fit is suggested

$$f_Y(t) = p\lambda_1 e^{-\lambda_1 t} + (1-p)\lambda_2 e^{-\lambda_2 t}.$$

Y is a 2-phase hyper-exponentially distributed random variable. The most commonly used procedure is the balanced mean method, that is

$$\frac{p}{\lambda_1} = \frac{1-p}{\lambda_2}.$$

To obtain the three parameters of the hyper-exponential distribution the following calculation is used:

$$p = \frac{1}{2}\left(\sqrt{\frac{C_X^2 - 1}{C_X^2 + 1}}\right)$$
$$\lambda_1 = \frac{2p}{\overline{X}}$$
$$\lambda_2 = \frac{2(1-p)}{\overline{X}}$$

4.1.2 Squared coefficient of variation is greater than one

The values of the input parameters are shown in Table 4.1. In this section,

Table 4.1: Numerical values of model parameters

N	$\boldsymbol{\sigma/N}$	$\boldsymbol{\nu}$	$\boldsymbol{\alpha_2}$	$\boldsymbol{\beta_2}$
100	0.01	0.02	1	1.1

these results are in connection with the effect of different service time distributions of incoming customers where the mean and variance are equal. We use hyper-exponential distribution if the squared coefficient of variation is greater than one, Table 4.2 shows the exact values of parameters of service time of incoming customers. Besides hyper-exponential, gamma, lognormal and Pareto distributions are also used for comparison.

Table 4.2: Parameters of service time of incoming customers

Distribution	Gamma	Hyper-exponential	Pareto	Lognormal
Parameters	$\alpha = 0.04$ $\beta = 0.04$	$p = 0.48$ $\lambda_1 = 0.961$ $\lambda_2 = 1.04$	$\alpha = 2.02$ $k = 0.505$	$m = -1.629$ $\sigma = 1.805$
Mean	1			
Variance	25			
Squared coefficient of variation	25			

Figure 4.1 shows the mean waiting time in function of arrival intensity of incoming customers. For these values of parameters regardless of the applied distribution, a maximum value of the mean waiting time can be seen. This maximum feature occurs for finite-source retrial queues, see for example [9], [19], [22], [38]. Differences can be observed among the values of mean waiting time especially in the case of using the gamma and Pareto distribution, even though the mean and variance are the same. In this figure, the effect of different distributions is clearly observable.

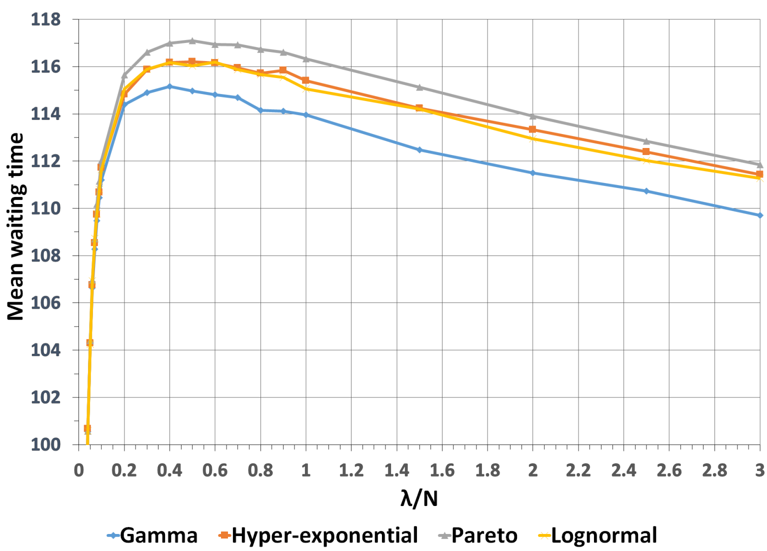

Figure 4.1: Mean waiting time vs. arrival intensity using various distributions

Figure 4.2 and 4.3 illustrates how the utilization of the server grows with the increment of the arrival intensity of incoming customers. The highest values can be found at gamma distribution but the differences of the applied distributions are as commensurable as in case of Figure 4.1. As the arrival intensity increases

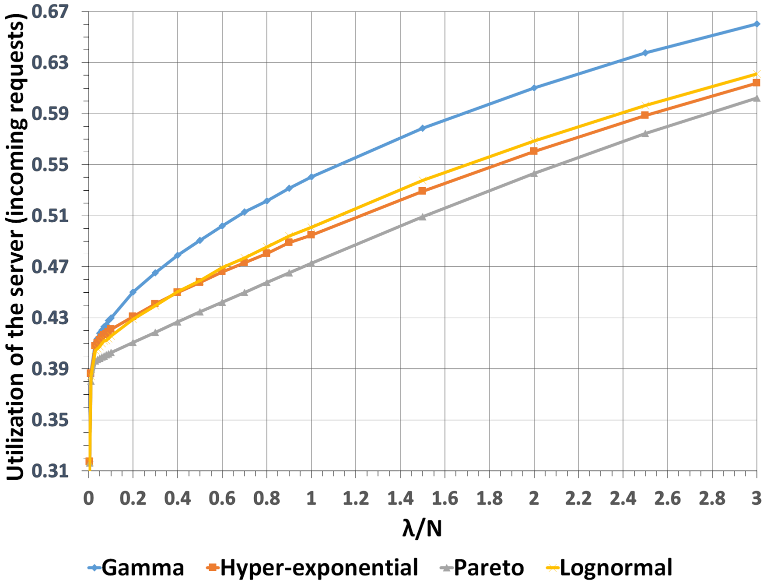

Figure 4.2: Utilization of server vs. arrival intensity using various distributions

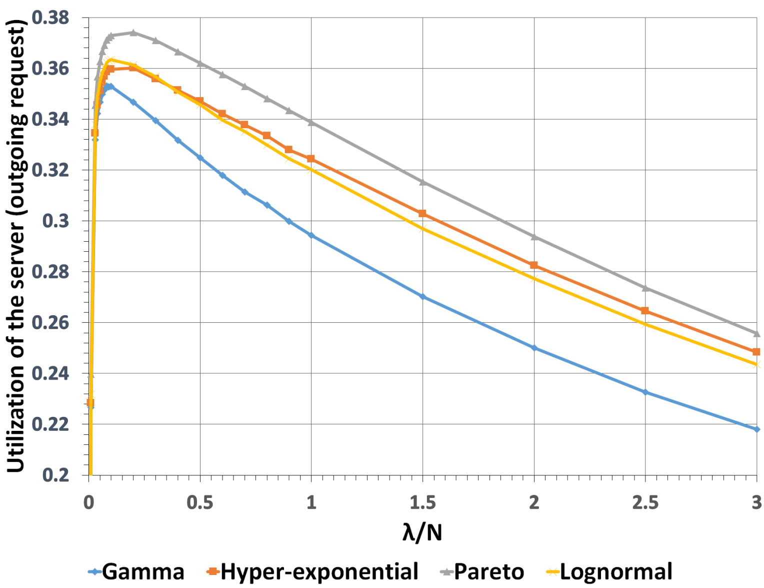

Figure 4.3: Utilization of server vs. arrival intensity using various distributions

the probability of performing outgoing call become less so outgoing requests spend less time at the service unit.

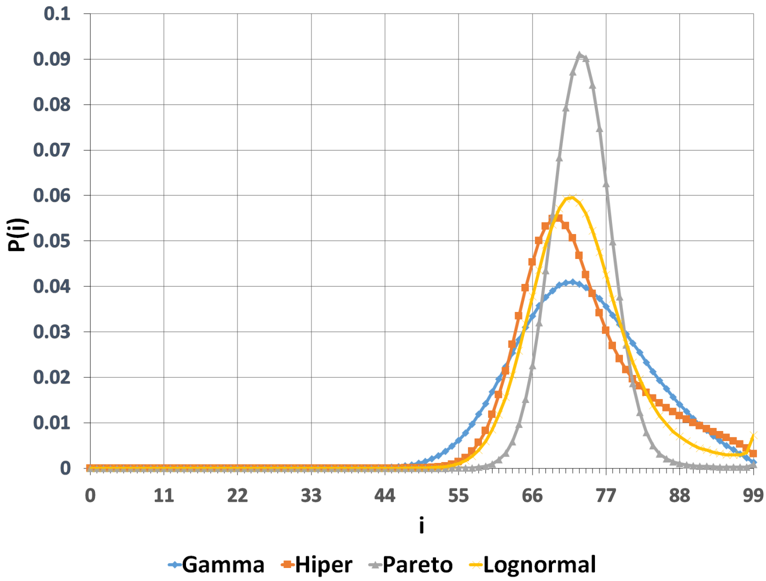

Figure 4.4: Comparison of steady-state distributions

On Figure 4.4 the comparison of steady-state distribution can be seen when the distribution of service time of the incoming customers is different. It represents the probability of how many customers residing in the orbit. Exploring the curves in more detail they correspond to normal distribution. The same parameter setting is used what Table 4.1 demonstrates where λ/N is 0.03.

To emphasize the importance of outgoing calls we compare our investigated model to the model without outgoing calls. This model is correlated with the classical retrial queuing system. Figure 4.5 depicts a comparison of the mean waiting time and due to the phenomena of outgoing call customers spend less time in the system, which is obvious looking at the curves. However, in our investigated model the utilization of the service unit (Figure 4.6) is much higher compared to the classical retrial queuing system therefore it spends less time in idle state. In this way, the efficiency of the server grows such that the mean waiting time decreases substantively. The distribution of service time of the incoming customer is gamma at Figure 4.5 and 4.6, but the ratio of difference is also true for the other distributions, too. On this figure under total utilization of the server, we mean the service of both the incoming and outgoing requests at the curve of with outgoing call.

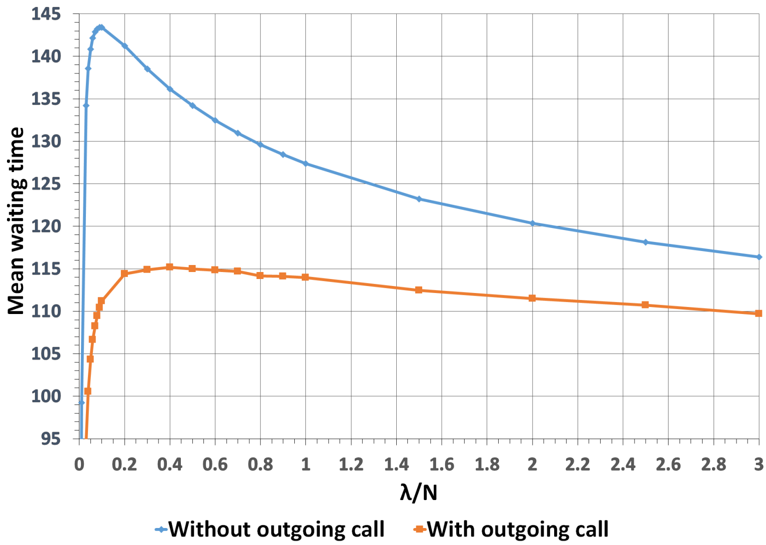

Figure 4.5: Comparison of our investigated model and the classical retrial queuing model on the mean waiting time

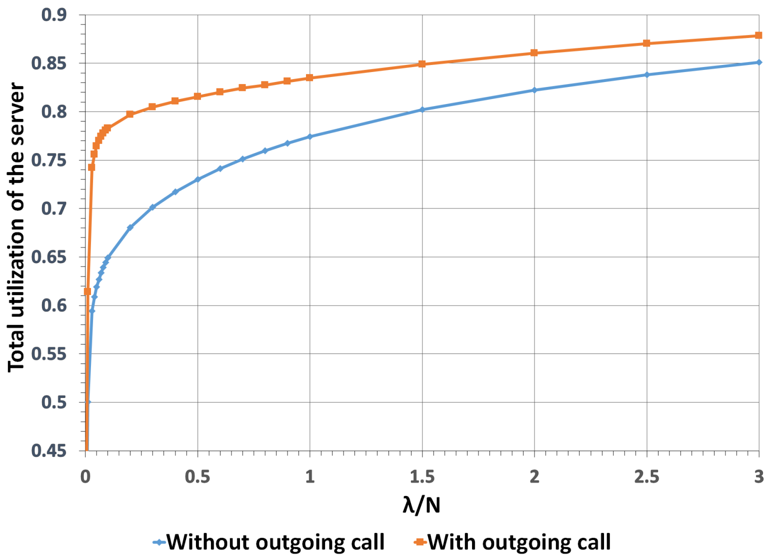

Figure 4.6: Comparison of our investigated model and the classical retrial queuing model on the utilization of server

4.1.3 Squared coefficient of variation is less than one

The same input parameters are used as in the previous section, see Table 4.1. The results are also in connection with the effect of different service time distributions of incoming customers where the mean and variance are equal. Instead of hyper-exponential distribution hypo-exponential distribution is used if the squared coefficient of variation is less than one. Table 4.3 illustrates the values of parameters of service time of incoming customers. In addition to hypo-exponential, we apply gamma, lognormal and Pareto distributions to perform sensitivity analysis.

Table 4.3: Parameters of service time of incoming customers

Distribution	**Gamma**	**Hypo-exponential**	**Pareto**	**Lognormal**
Parameters	$\alpha = 1.5504$ $\beta = 1.5504$	$\mu_1 = 1.3$ $\mu_2 = 4.333$	$\alpha = 2.597$ $k = 0.615$	$m = -0.249$ $\sigma = 0.705$
Mean	1			
Variance	0.6449704142			
Squared coefficient of variation	0.6449704142			

Figure 4.7 demonstrates the mean waiting time in the function of arrival intensity of incoming calls. Taking a closer look at the curves it can be stated that the values of mean waiting time are almost identical regardless of the applied distribution. With this parameter setting the interesting maximum value of the mean waiting time appears as in the previous section.

Figure 4.8 and 4.9 illustrate how the utilization of the server increases with the increment of the arrival intensity of incoming customers. As in the case of mean waiting time here using different distributions result in the same utilization. It seems that when the squared coefficient of variation is less than one using different distributions, it does not affect the performance measures and the obtained results are nearly identical.

Similarly to the previous section, we compare the results between the classical retrial queuing system and our investigated model. On Figure 4.8 and 4.9 the same tendency can be observed like when the the squared coefficient of variation is greater than one, namely values of mean waiting time is lesser when the server can make outgoing calls. But this also affects the utilization of the service unit because with the help of the outgoing calls server spends less time without satisfying the needs of the customers. As in the previous section the service time of the incoming customer follows gamma distribution at Figure 4.10 and 4.11, but the ratio of difference is also true for the other distributions, too.

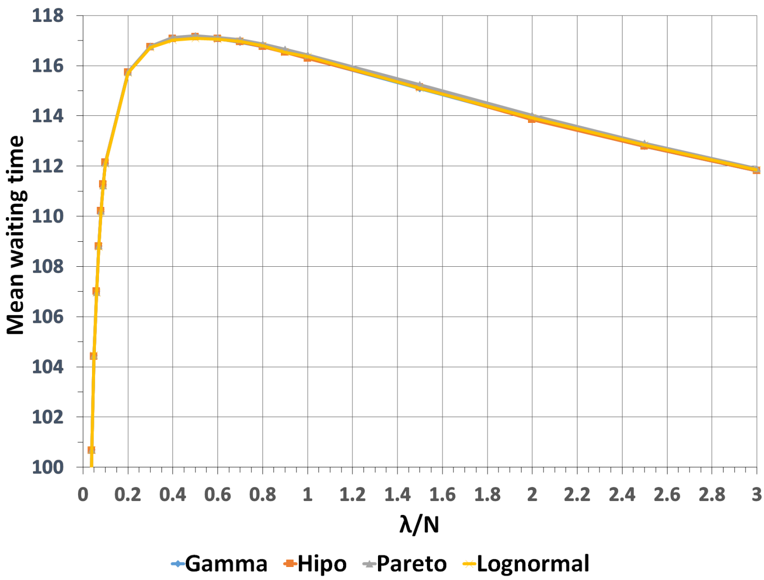

Figure 4.7: Mean waiting time vs. arrival intensity using various distributions

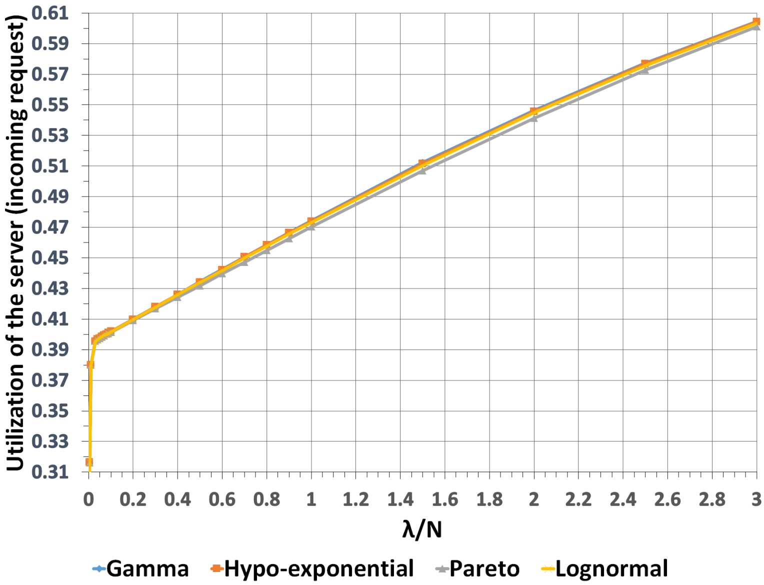

Figure 4.8: Utilization of server vs. arrival intensity using various distributions

From Figure 4.5, 4.6, 4.10 and 4.11 it can be said that the utilization of service

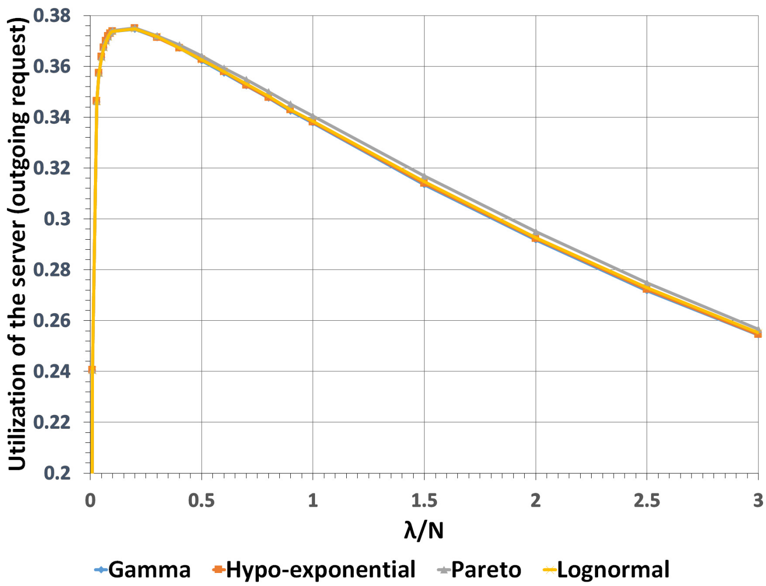

Figure 4.9: Utilization of server vs. arrival intensity using various distributions

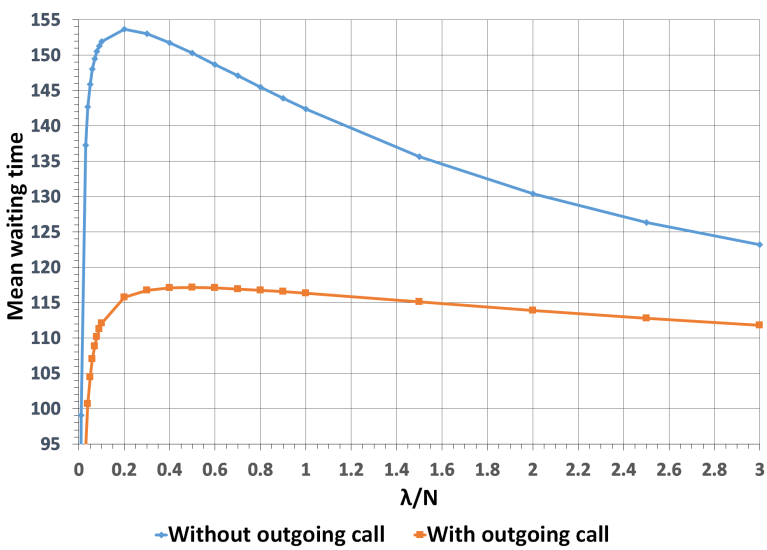

Figure 4.10: Comparison of our investigated model and the classical retrial queuing model on the mean waiting time

unit escalates when outgoing calls are performed, but it also results lesser mean waiting time of incoming customers. With a proper parameter setting in the

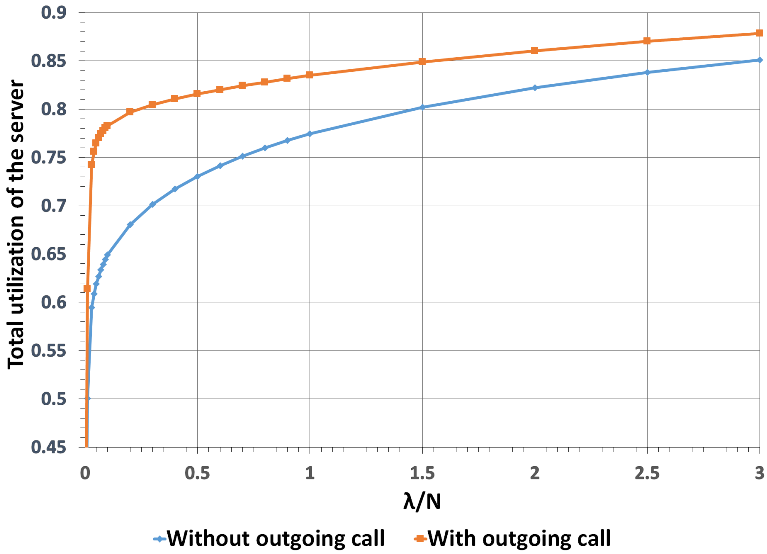

Figure 4.11: Comparison of our investigated model and the classical retrial queuing model on the utilization of server

case of outgoing calls the utilization of the server is much higher in a way that customers spend less time in the orbit. In the case of with outgoing calls total utilization of the server includes both incoming and outgoing requests occupying the service unit.

4.2 Finite-source retrial queuing systems with outgoing calls

Our objective is to examine the operation of a system when customers from outside can enter the system containing a non-reliable server [J3]. The novelty of this section is to compare different scenarios in connection with server breakdown using various distributions of failure on performance measures like mean waiting time of primary customers or utilization of the server. We do not have knowledge about that any paper dealt with this model in case of a non-reliable server or applying different distributions except for exponential.

4.2.1 System model

In this model, N customer resides in the source, which is finite such that the system is stable in every moment. Each customer can produce a call (primary customers) towards the server with a rate of λ/N, so the inter-request times are exponentially distributed with parameter λ/N. Our model does not contain a queue, therefore the service of primary customers starts instantaneously if the server is idle. The service time of the primary customers follows gamma, hypo-exponential, hyper-exponential, Pareto, and lognormal distribution with different parameters but with the same mean value. When the submission (the service of a request) is successful, the request goes back to the source. Customers located in the orbit may retry their requests for service after a random time. The distribution of this period is exponential with parameter σ/N. The server can break down during its operation or in idle state according to gamma, hypo-exponential, hyper-exponential, Pareto, and lognormal distribution time with the same mean value. Restoration starts instantly upon the breakdown and that time is also an exponentially distributed random variable with rate γ_1. The idle server after some exponentially distributed period can make outgoing calls towards the customers (secondary) from an infinite source. It is performed after an exponentially distributed idle period with parameter γ. The service of these customers can take place if no primary customers arrive from the finite source or from the orbit and the server is not in a failed state upon their arrivals. Otherwise, they are cancelled and they return without entering the system. The service time of these types of customers follows gamma distribution with parameters α_2 and β_2. We differentiate four scenarios in case of server failure:

- Scenario 1: Primary customers are forwarded immediately towards the orbit and the secondary customers leave the system without service.
- Scenario 2: Primary customers are forwarded immediately towards the orbit and the secondary customers remain at the service area during the recovery of the service unit.
- Scenario 3: Primary customers remain at the service area during the recovery of the service unit and the secondary customers leave the system without service.
- Scenario 4: Both the primary and the secondary customers remain in the service area during the recovery of the service unit.

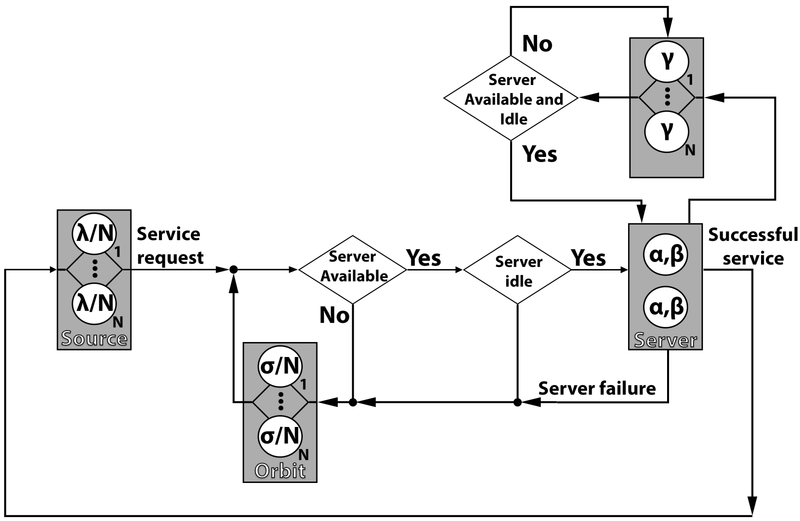

Figure 4.12: System model

4.2.2 Simulation results

4.2.2.1 Different distributions of service time of primary customers

Squared coefficient of variation is greater than one: The applied values of the input parameters are presented in Table 4.4, the failure time of the server is exponentially distributed with rate γ_0 in this case.

Table 4.4: Numerical values of model parameters

N	λ/N	γ_0	γ_1	σ/N	γ	α_2	β_2
100	0.01	0.05	0.5	0.01	0.8	1	1

We investigate the effect of different service time distributions where the mean and variance are equal, Table 4.5 shows the parameters of service time of primary customers. In this subsection, the squared coefficient of variation is greater than one, because using hyper-exponential distribution regardless of the parameters the squared coefficient of variation is always greater than one.

Besides hyper-exponential, gamma, lognormal and Pareto distributions are used for comparison.

Table 4.5: Parameters of service time of primary customers

Distribution	Gamma	Hyper-exponential	Pareto	Lognormal
Parameters	$\alpha = 0.037$ $\beta = 0.015$	$p = 0.482$ $\lambda_1 = 0.385$ $\lambda_2 = 0.416$	$\alpha = 2.018$ $k = 1.261$	$m = -0.751$ $\sigma = 1.826$
Mean	2.5			
Variance	169			
Squared coefficient of variation	27.04			

The mean waiting time is shown in the function of the arrival intensity of primary customers on the next figures, where all four scenarios appear. Interestingly, pronounced differences can be observed especially on Figure 4.13 and 4.14. Despite the fact that the mean and variance are the same, results clearly illustrate the effect of various distributions. The highest values are experienced in the case of Pareto distribution and the lowest in the case of gamma distribution. When primary customers remain at the server (Figure 4.15 and 4.16) during failure values of mean waiting time is much closer compared to the first two scenarios. With suitable parameter settings, we experience the maximum property characteristic of a finite-source retrial queueing system.

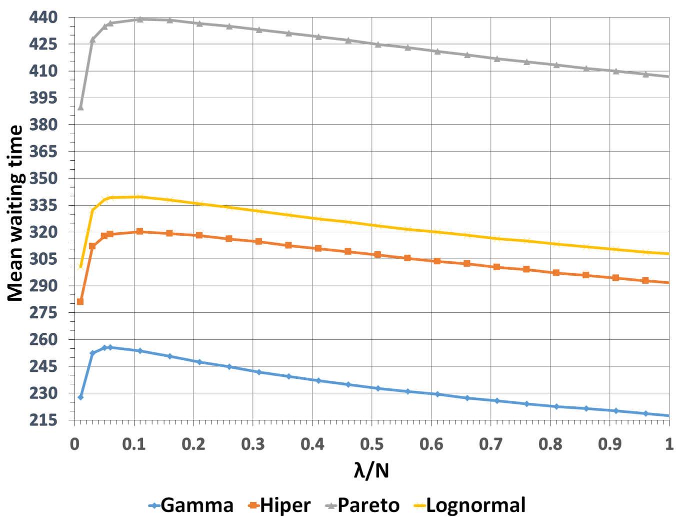

Figure 4.13: Mean waiting time vs. arrival intensity using various distributions in Scenario 1

The next four Figures illustrate how the utilization of the server is larger when the arrival intensity of primary customers increases. This performance measure

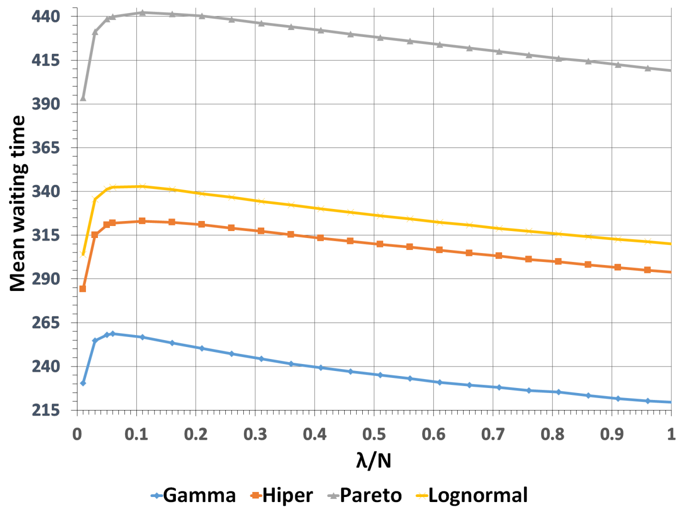

Figure 4.14: Mean waiting time vs. arrival intensity using various distributions in Scenario 2

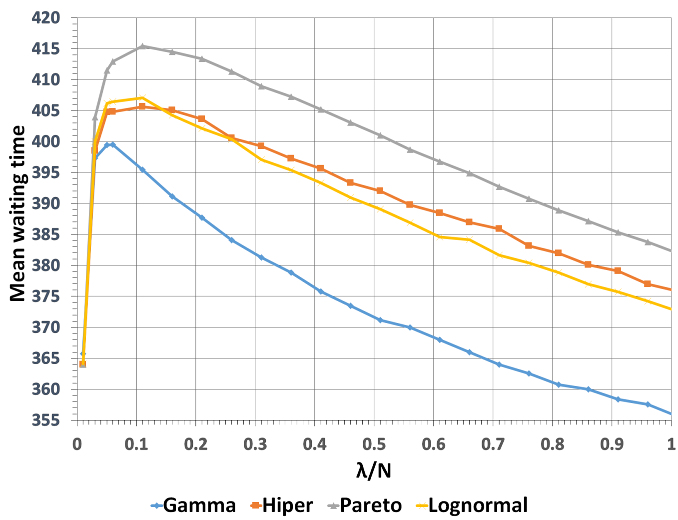

Figure 4.15: Mean waiting time vs. arrival intensity using various distributions in Scenario 3

contains the service of outgoing calls, too. In case of Figure 4.17 and 4.18 using Pareto distribution we observe the highest values indicating that here happens the most interruptions while in Figure 4.19 and 4.20 the highest values can be found at gamma distribution where primary customers spend the least in the

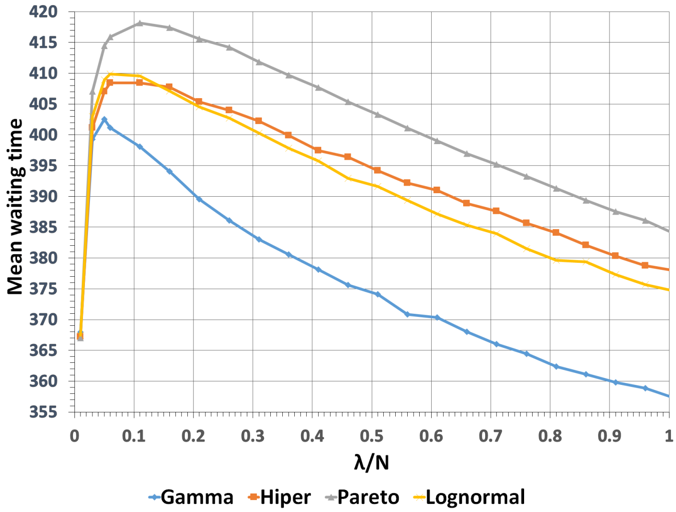

Figure 4.16: Mean waiting time vs. arrival intensity using various distributions in Scenario 4

system. Naturally in these scenarios, the utilization of the server is higher than in Scenario 1 or in Scenario 2.

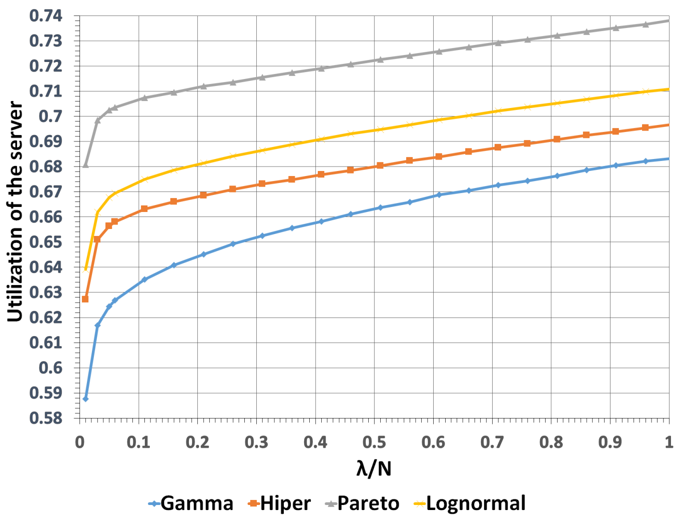

Figure 4.17: Utilization of server vs. arrival intensity using various distributions in Scenario 1

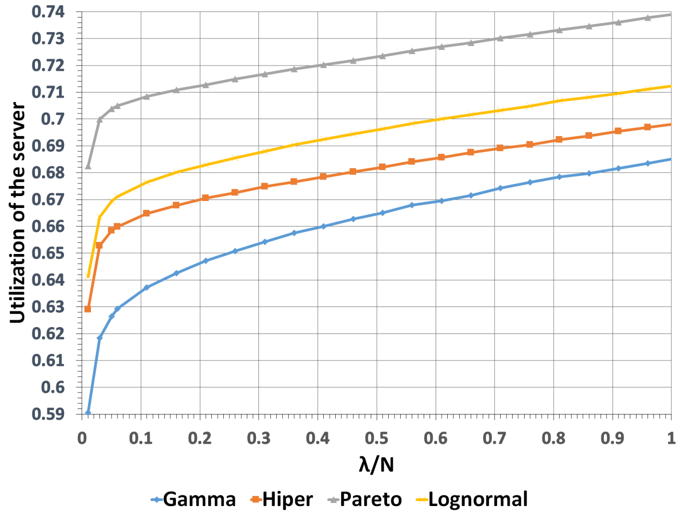

Figure 4.18: Utilization of server vs. arrival intensity using various distributions in Scenario 2

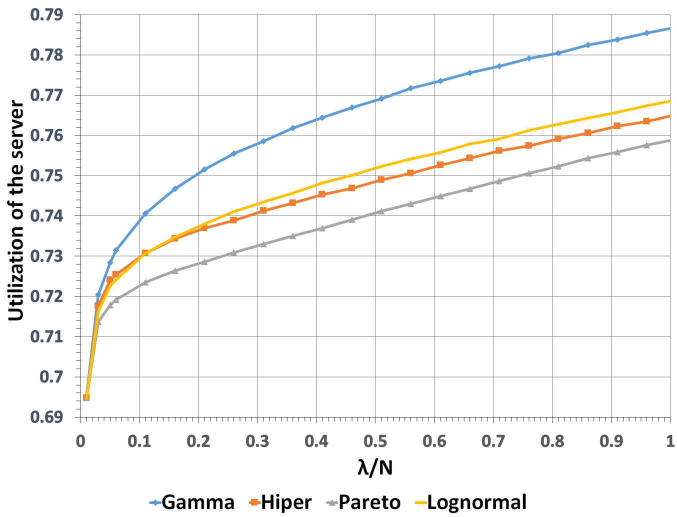

Figure 4.19: Utilization of server vs. arrival intensity using various distributions in Scenario 3

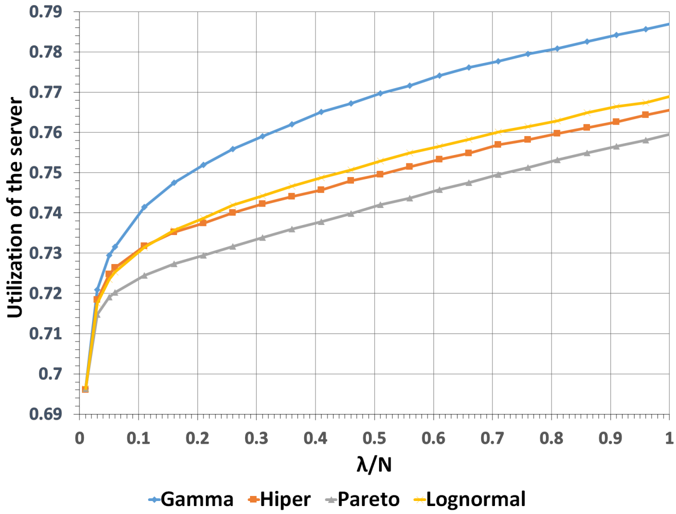

Figure 4.20: Utilization of server vs. arrival intensity using various distributions in Scenario 4

Squared coefficient of variation is less than one In the previous section we see that when the squared coefficient of variation of the service time of primary customers is more than 1, the differences are quite high among the performance measures. The question arises whether it is true for using other parameter settings for example when the squared coefficient of variation of the service time of primary customers is less than one. We use the same parameters like in the previous section (Table 4.4), and Table 4.6 contains the changed parameters of service time of primary customers. Instead of hyper-exponential, we use this time hypo-exponential distribution because the squared coefficient of variation is always less or equal to one.

Table 4.6: Parameters of service time of primary customers

Distribution	Gamma	Hypo-exponential	Pareto	Lognormal
Parameters	$\alpha = 1.8$ $\beta = 0.72$	$\mu_1 = 0.6$ $\mu_2 = 1.2$	$\alpha = 2.673$ $k = 1.565$	$m = 0.695$ $\sigma = 0.665$
Mean	2.5			
Variance	3.472			
Squared coefficient of variation	0.555			

Comparing Figure 4.13, 4.14, 4.15 and 4.16 with Figure 4.21, 4.22, 4.23 and 4.24 the contrast is quite obvious, namely the values of mean waiting time are almost identical regardless of the distribution. However, when the squared coefficient of variation is more than one it results lower mean waiting time in all Scenarios compared to previous section except the Pareto distribution.

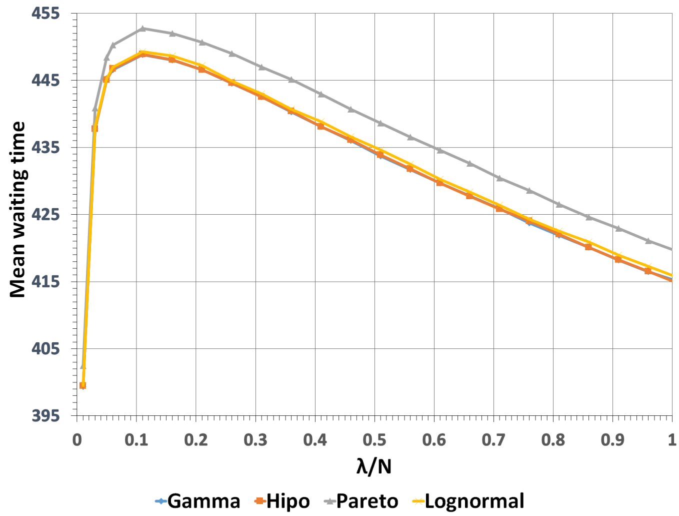

Figure 4.21: Mean waiting time vs. arrival intensity using various distributions in Scenario 1

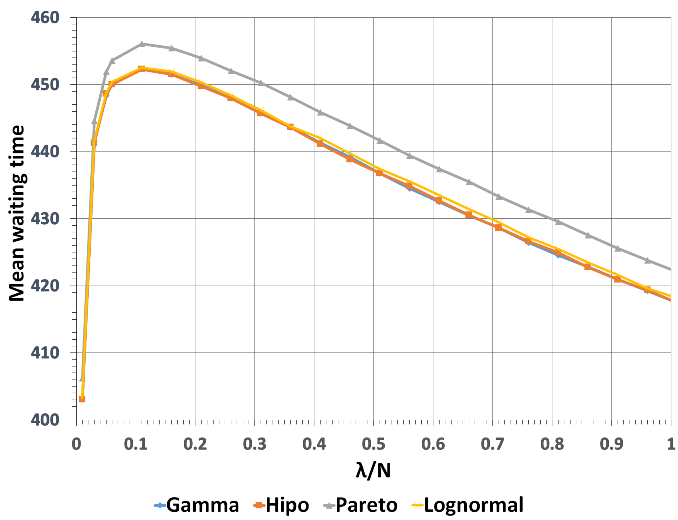

Figure 4.22: Mean waiting time vs. arrival intensity using various distributions in Scenario 2

Figure 4.25, 4.26, Figure 4.27 and 4.28 present the utilization of server in function of arrival intensity of primary customers, respectively. After observing the results in connection of the mean waiting time it is no wonder that the received results

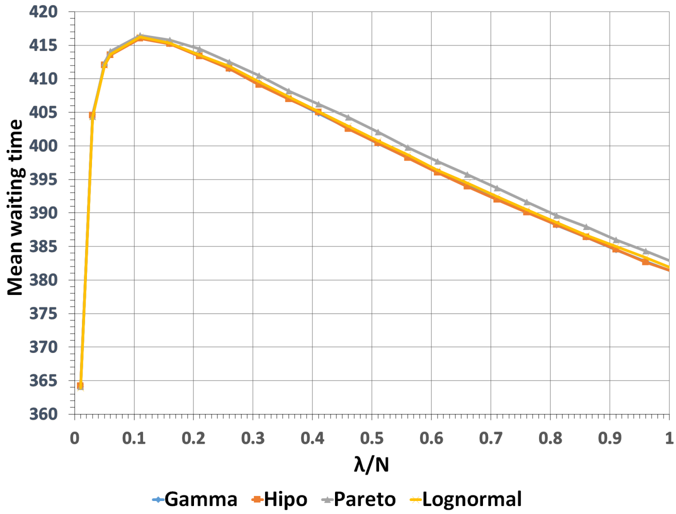

Figure 4.23: Mean waiting time vs. arrival intensity using various distributions in Scenario 3

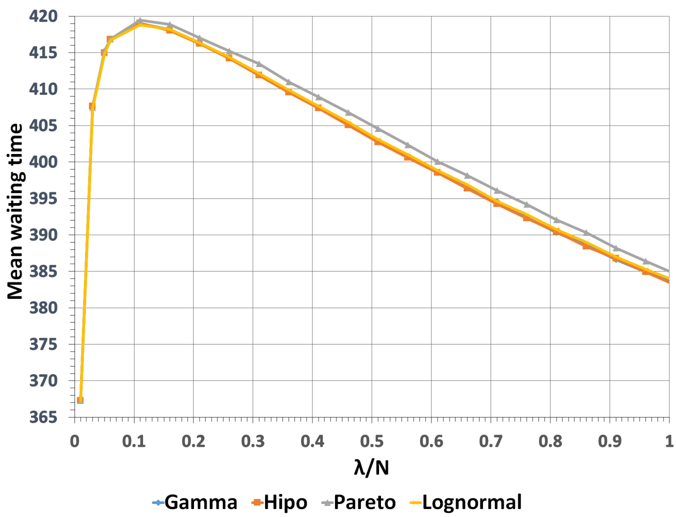

Figure 4.24: Mean waiting time vs. arrival intensity using various distributions in Scenario 4

are almost identical, this is also true if we take a closer look at the applied Scenarios.

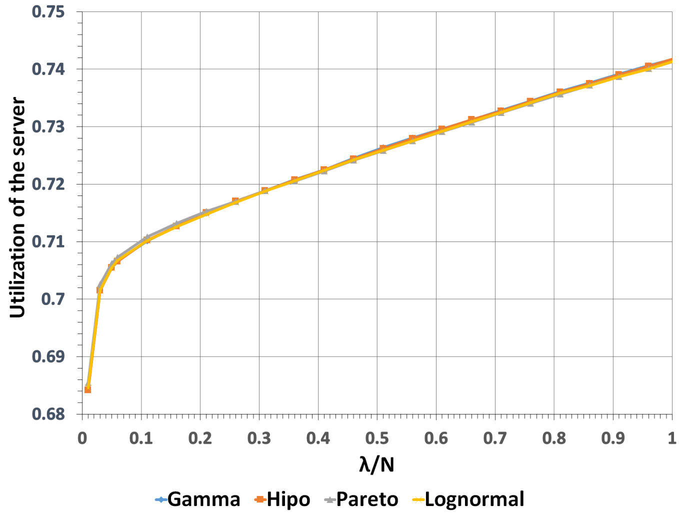

Figure 4.25: Utilization of server vs. arrival intensity using various distributions in Scenario 1

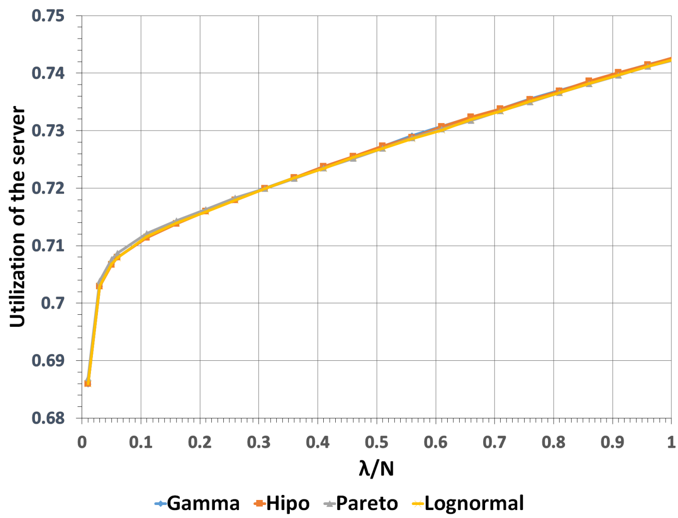

Figure 4.26: Utilization of server vs. arrival intensity using various distributions in Scenario 2

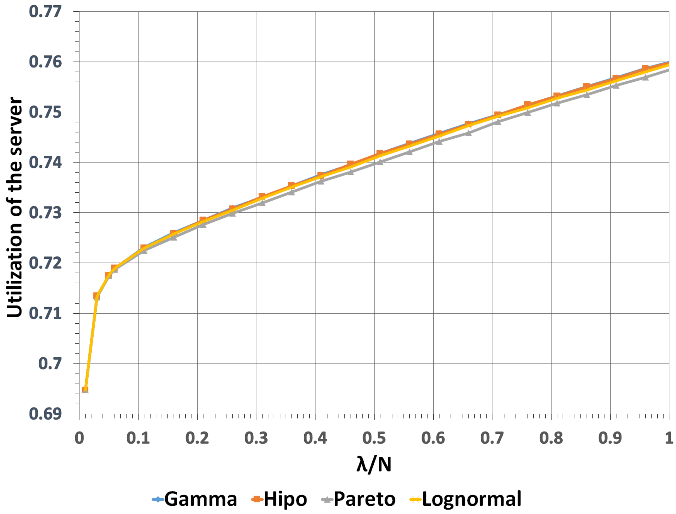

Figure 4.27: Utilization of server vs. arrival intensity using various distributions in Scenario 3

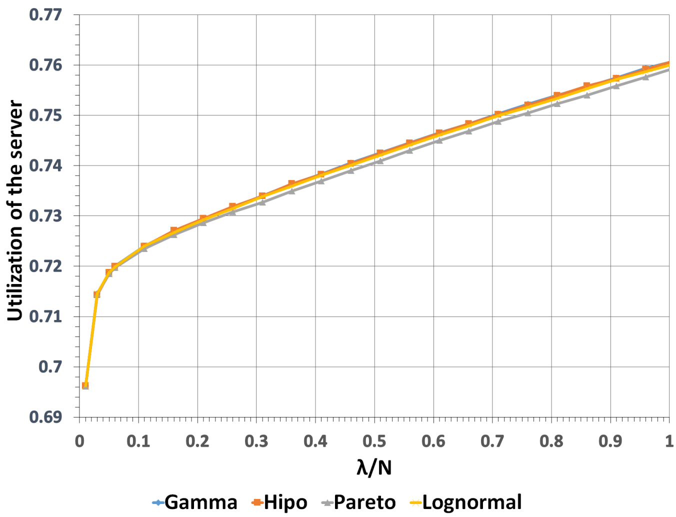

Figure 4.28: Utilization of server vs. arrival intensity using various distributions in Scenario 4

4.2.2.2 Different distributions of failure time of the server

Squared coefficient of variation is greater than one In this section, we investigate the effect of the failure time of the server on the mean waiting time.

We are interested in carrying out a sensitivity analysis to identify how much the various distributions alter the characteristics of the system. To do so for failure time we utilize gamma, hyper-exponential, lognormal and Pareto distributions having the same mean and variance so we carefully selected the parameters at the adequate distribution. The service of primary customers, in this case, is exponential with rate of $\alpha_1 = 1$. The same investigation will be carried out as in the previous section so first, we examine the case when the squared coefficient of variation is greater than one.

Table 4.7: Numerical values of model parameters

N	λ/N	α_2	β_2	γ_1	σ/N	γ	α_1
100	0.01	1	2.5	0.5	0.01	0.8	1

For the easier understanding the numerical values of parameters are collected in Table 4.7 and the parameters of failure time in Table 4.8.

Table 4.8: Parameters of failure time of the server

Distribution	Gamma	Hyper-exponential	Pareto	Lognormal
Parameters	$\alpha = 0.312$ $\beta = 0.056$	$p = 0.362$ $\lambda_1 = 0.129$ $\lambda_2 = 0.228$	$\alpha = 2.146$ $k = 2.984$	$m = 1.003$ $\sigma = 1.198$
Mean	5.588			
Variance	100			
Squared coefficient of variation	3.2			

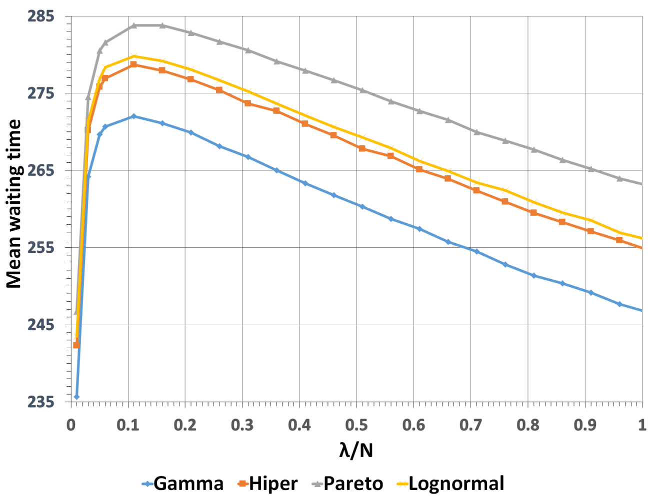

Figure 4.29: Mean waiting time vs. arrival intensity using various distributions in Scenario 1

The previous four figures show the mean waiting time of primary customers

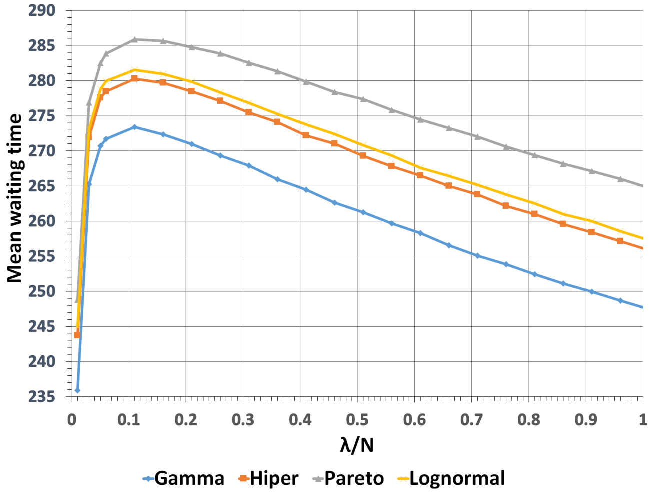

Figure 4.30: Mean waiting time vs. arrival intensity using various distributions in Scenario 2

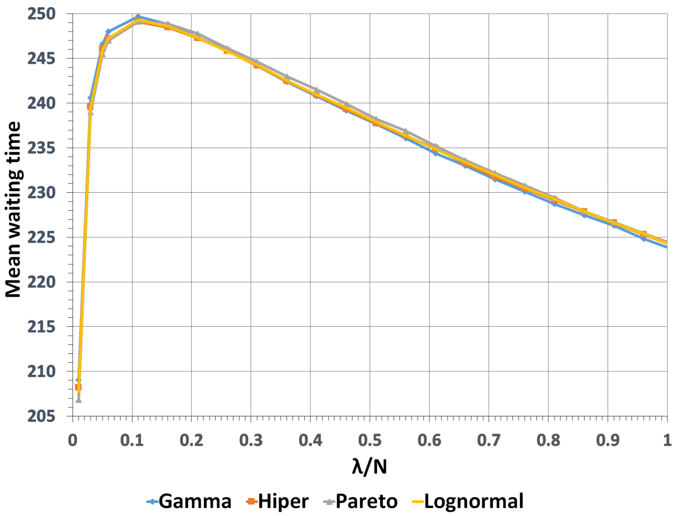

Figure 4.31: Mean waiting time vs. arrival intensity using various distributions in Scenario 3

vs. the arrival of primary customers in respect of all Scenarios. In Figure 4.29 and 4.30 the difference is evident among the applied distributions apparently in the case of Pareto where the values of mean waiting time are the highest. But in Scenario 3 and Scenario 4 this divergence disappears and the received graphs

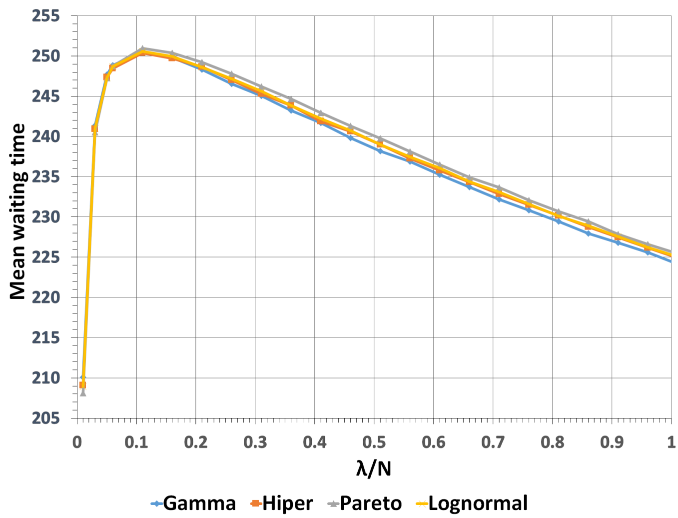

Figure 4.32: Mean waiting time vs. arrival intensity using various distributions in Scenario 4

reflect near identity.

Squared coefficient of variation is less than one In this section, we change the parameters of failure of the server to examine the mean waiting time when the squared coefficient of variation is less than one.

Table 4.9: Parameters of failure time of the server

Distribution	Gamma	Hypo-exponential	Pareto	Lognormal
Parameters	$\alpha = 1.232$ $\beta = 0.2204$	$\mu_1 = 0.2$ $\mu_2 = 1.7$	$\alpha = 2.494$ $k = 3.347$	$m = 1.426$ $\sigma = 0.771$
Mean	5.588			
Variance	25.346			
Squared coefficient of variation	0.8116			

Table 4.9 contains the overview of these parameters, mean value remains the same as in the previous section. The value of other parameters are unchanged and Table 4.7 contains them.

On the next four Figures the mean waiting time of the primary jobs are displayed as a function of the primary generation rate. On Figure 4.33, 4.34, 4.35 and 4.36 the difference is very moderate compared to Figure 4.29, 4.30, 4.31 and 4.32. The lines are identical in case of Scenario 3 and 4. These graphs show that the distribution of the failure time has minor impact on the performance measures when the squared coefficient of variation is less than one.

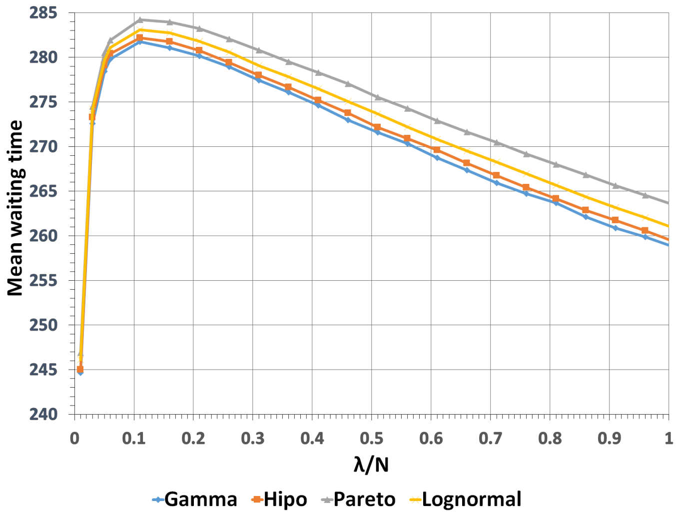

Figure 4.33: Mean waiting time vs. arrival intensity using various distributions in Scenario 1

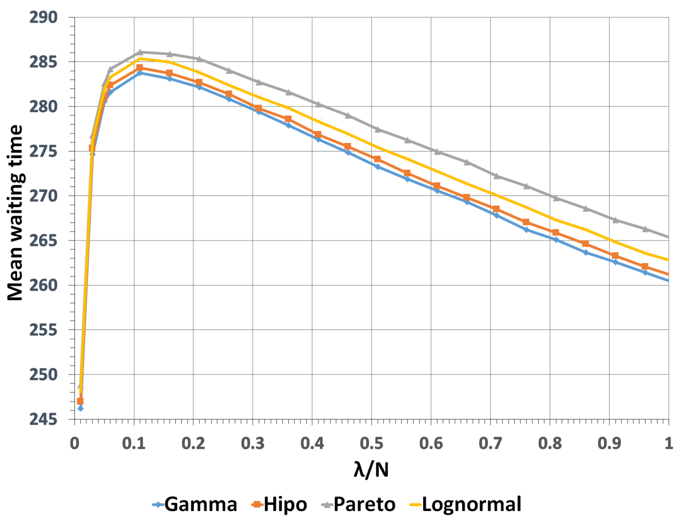

Figure 4.34: Mean waiting time vs. arrival intensity using various distributions in Scenario 2

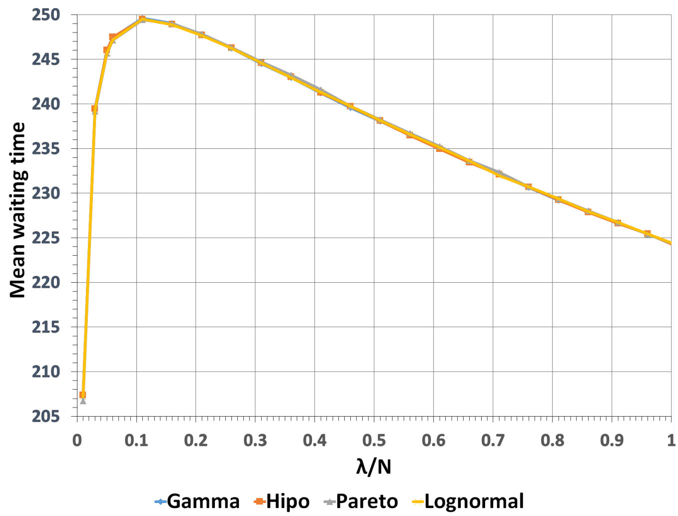

Figure 4.35: Mean waiting time vs. arrival intensity using various distributions in Scenario 3

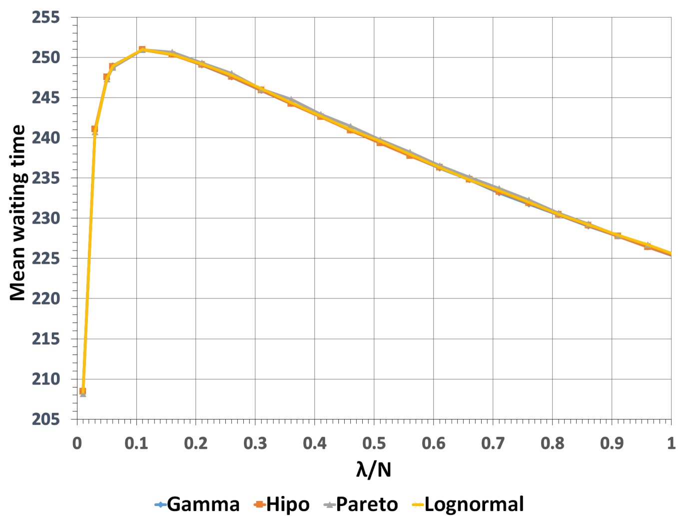

Figure 4.36: Mean waiting time vs. arrival intensity using various distributions in Scenario 4

5

Conclusions

This chapter concludes and summarizes the achieved results

Contents

5.1 Summary

The work presented in this thesis aimed at developing a simulation program package based on SimPack [24] creating simulation models. In particular, discrete event simulation relying on various algorithms were developed to investigate the efficiency of retrial queuing systems. In chapter 1 I gave a brief, concise introduction on queuing systems and their importance playing a vital role in many application fields. Then a short introduction is demonstrated in section 1.1.1 about the background of retrial queuing systems which is a specific queuing model and about two-way communication systems which is quite a popular topic over the past few years. These types of systems perform an important role in many territories in our life starting from various protocols like TCP to major communication systems as call centres. A great number of companies provide a variety of services including via telephone. In addition, a large call centre serves thousands of calls per day, each of which demands a response within seconds, and besides that agents may attempt to initiate phone calls to sell, advertise and promote products and services of a business. One of the most important performance measures for a call centre is the utilization of the server (agent) in order to minimize the probability that the customer hangs up.

In chapter 2 a general retrial queuing model is introduced, which serves as a base model, and gives an overall overview of the implemented simulation model. A retrial queuing system of type M/G/1//N with collisions of the customers and an unreliable server is regarded and the number of sources is denoted by N. Each of them can generate a request with a rate of λ/N. In the absence of a waiting queue, the service of an arriving customer starts immediately in the case of an idle server. Otherwise, when a collision occurs or the server is located in a failed state and these requests wait for an exponentially distributed time in the orbit to launch another attempt to attain the server. The server has gamma-distributed service time with parameters α and β.

Chapter 3 consists of the results in connection with standard retrial queuing systems. In section 3.1 I have reproduced the results obtained by the numerical procedure in [27] and I have calculated the probability distribution of the number of transitions of the customer into the orbit. To measure how similar two probability distributions are obtained by the simulation program and by the asymptotic method we use Kolmogorov distance. As we expected with the increment of N Kolmogorov distance decrease though with the parameter setting we use no significant reduction is observed if $N > 50$.

In Section 3.2 I investigated the general model as described in Chapter 2 with the addition that requests gets back to the orbit in case of server failure. Different cases have been defined to carry out a sensitivity analysis of the performance measures using various distribution. I compared the main performance measures such as the mean waiting time, the mean total uninterrupted service time of a customer, the mean total interrupted service time of a customer among the cases to show the effect of different distributions of service, inter-arrival and retrial times. With the help of the simulation program, I am able to illustrate the steady-state distribution of every scenario for all the investigated cases and all of them tend to be normally distributed even if the distribution of retrial or arrival time is not exponentially distributed. As it is expected with the increment of arrival rate the mean waiting time increases as well and with the used parameter setting we experience the maximum property characteristic of a finite-source retrial queueing system. It is worth mentioning when the service time is exponentially distributed then the sum of the mean total interrupted service time of a customer and the mean total uninterrupted service time of a customer is equal to 1/parameter of service rate. This phenomenon is due to the effect of collision because uninterrupted service times constitute very low values which results in a low average, too. I observe that the growth of λ and β eventuate greater values of the investigated performance measures in every scenario and every case. I proved this with numerical results containing in tables.

In the next Section (see 3.3) two operation modes are considered in case of server failure when a customer is under service:

- Operation mode number 1: the interrupted request gets into the orbit instantaneously (this happened in the previous section).

- Operation mode number 2: the service of the interrupted request is suspended and it continues after repairing the server.

The effect of two different operation modes was analyzed regarding the main performance measures. We got what we expected that it was clearly recognizable that applying operation mode number 2 gives more beneficial results compared to operation mode number 1. In all cases and all scenarios it turns out that the steady-state distribution of the number of customers in the system is still normally distributed.

The focus of Section 3.3 is to examine the performance behaviour of the general model applying different situations when a busy server breaks down. Two types of operation behaviour are distinguished:

- Without blocking: requests are able to enter the system and these are forwarded to the orbit instantaneously (same as in Section 3.2).

- With blocking: the arriving customers cannot enter the system; they return to the source and a new request generation process starts.

The achieved results in connection with the two different operation behaviour are compared with each other. The steady-state distribution of the investigated cases still follows normal distribution even when blocking is incurred. As a result of blocking it ensures the expected behaviour resulting in lower mean waiting time. Different figures illustrated that the customers spend more and more time in the source if the server is more likely subjected to breakdowns in case of blocking and the mean spent time in the source of the customers is constant when blocking is not applied.

In Chapter 4 retrial queuing systems by the help of two-way communication are analyzed in which of the base is the depicted general model described in section 2.1. I modified the performance model of the general retrial queuing model to get a two-way communication system of type $M/G/1//N$ with a reliable service unit that is capable to produce outgoing calls to the customers residing in the orbit in Section 4.1. It is performed after an exponentially distributed time with parameter ν. The service time of these outgoing customers follows gamma distribution with parameters α_2 and β_2. I have stated the results in connection with the effect of different service time distributions of incoming customers where the mean and variance are equal. To do so my program is completed with random number generators and the parameters are selected for a valid comparison. The goal is to achieve the same mean and variance so in every distribution, the fitting process is accomplished. Firstly, when the squared coefficient of variation is more than one. I presented the mean waiting time of primary customers in the function of arrival intensity and regardless of the applied distribution, it has a maximum value. These results reassure that the distribution affects the performance measures using the gamma and Pareto distribution with the same mean value and variance bring about a noticeable difference. Figures also represent that the utilization of server with outgoing requests decrease with the increment of the arrival intensity. I studied the steady-state distribution which displays the probability of how many customers residing in the orbit. Exploring the curves in more detail they correspond to a normal distribution and it can be seen how they differ from each other. To emphasize the importance of two-way communication systems the investigated model is compared with the classical retrial queuing model where there is no outgoing call. The outcome demonstrates that customers spend less time in the system in such a way that the utilization of the service unit is a much higher meaning that our investigated model is more efficient than the classical queuing system. The explanation of this phenomenon can be explained by that the service unit works less in idle state. This is true for every chosen distribution of service time. Afterwards, I selected the parameters of the different distributions of service time that the squared coefficient of variation would be less than one. However, in this case, the obtained results nearly coincide with each other in every aspect of performance measures. From these results, we come to the conclusion that when the squared coefficient of variation is more than one then the applied distribution has a great effect on the performance measures while it is less than one the chosen distribution has minimal impact on the operation of the system.

In Section 4.2 our objective is to compare different scenarios in connection with server breakdown using various distributions of service and failure time on performance measures like mean waiting time of primary customers or utilization of the server. The idle server after some exponentially distributed period with parameter γ is able to produce outgoing calls towards the customers (secondary) from an infinite source. The service of these customers follows a gamma distribution with parameters α_2 and β_2 but this is realized if no primary customers arrive from the finite source or from the orbit and the server is not in a failed state upon their arrivals. I differentiate four scenarios in case of server failure:

- Scenario 1: Primary customers are forwarded immediately towards the orbit and the secondary customers leave the system without service.
- Scenario 2: Primary customers are forwarded immediately towards the orbit and the secondary customers remain at the service area during the recovery of the service unit.
- Scenario 3: Primary customers remain at the service area during the recovery of the service unit and the secondary customers leave the system without service.
- Scenario 4: Both the primary and the secondary customers remain in the service area during the recovery of the service unit.

First, the applied values of the input parameters are chosen that the squared coefficient of variation would be greater than one. Hyper-exponential, gamma, lognormal and Pareto distributions are used for comparison. I have calculated the mean waiting time of primary customers and the total utilization of the server. I examined the effect of various distribution for every scenario and the results clearly indicates when the squared coefficient of variation is greater than one then the disparity is huge among the performance measures having the same mean and variance especially in the case of Scenario 1 and 2, but very little when it is less than one. Interestingly, in all scenarios, the mean waiting time has a maximum which is a typical characteristic of a finite-source retrial queueing system. The same phenomenon is experienced when I analyze the results using various distributions of failure time.

5.2 Összefoglalás

A disszertációban elvégzett tevékenység célja a SimPack [24] programcsomagon alapuló szimulációs program kifejlesztése volt, amely lehetőséget biztosít új szimulációs modellek létrehozására. Többek között a diszkrét eseményű szimuláció megvalósításához szükséges algoritmusokat is tartalmazza, így visszatéréses sorbanállási rendszerek hatékonyságának vizsgálatára is alkalmas. Az 1. fejezetben egy rövid, tömör bevezetést adtam a sorbanállási rendszerekről. Néhány példával demonstráltam, hogy mekkora szerepet töltenek be számos alkalmazási területen. Az 1.1.1. bekezdésben a visszatéréses és kétirányú kommunikációs sorbanállási rendszerekről adtam rövid ismertetőt. Ezek speciális rendszerek és az elmúlt években nagyon sok publikáció foglalkozott működésükkel. Az ilyen típusú rendszerek fontos szerepet játszanak életünk számos területén a különféle protokolloktól a fő kommunikációs rendszerekig, mint például a TCP-től az ügyfélszolgálatokig. Nagyon sok szolgáltatást tud nyújtani egy cég köztük az egyik ilyen a telefonos ügyfélszolgálat. Egy nagyobb méretű ügyfélszolgálat napi több ezer hívást bonyolít le. Az ügyfelek elvárják, hogy az az ügyfélszolgálat munkatársai a másodperc töredéke alatt rendelkezésükre álljon. Emellett telefonhívásokat szoktak kezdeményezni üzleti termékek és szolgáltatások eladása, reklámozása és népszerűsítése érdekében az ügyfelek részére. A call center egyik legfontosabb jellemzője a munkatárs (szerver) kihasználtsága, azaz minél kevesebb legyen annak valószínűsége, hogy az ügyfél megszakítja a hívást.

A 2. fejezet egy általános visszatéréses sorbanállási rendszert vezet be, amely az általam vizsgált rendszerek alapját teszi ki, illetve átfogó áttekintést ad az implementált szimulációs modellel kapcsolatban. Egy olyan $M/G/1//N$ típusú visszatérési rendszert veszek alapul, ahol az igények ütközést idézhetnek elő és a szerver megbízhatatlan. N jelöli a források számát és minden forrásban lévő egyed λ/N intenzitással generálhat kérést. Várakozási sor hiányában tétlen kiszolgáló esetén az érkező ügyfél kiszolgálása azonnal elkezdődik. Ellenkező esetben ütközés történik és ezen igények az orbitba kerülnek, ahol exponenciális eloszlású ideig várnak, hogy újabb kísérletet indítsanak a szerver eléréséhez. A szerver gamma eloszlású kiszolgálási idővel rendelkezik α és β paraméterekkel.

3. fejezet a sztenderd visszatéréses sorbanállási rendszerekkel kapcsolatos eredményeket tartalmazza. A 3.1. bekezdésben reprodukáltam a [27]-ben numerikus eljárással kapott eredményeket, és kiszámoltam az igényátmenetek számának valószínűségű eloszlását az orbitban. A Kolmogorov-Smirnov távolságot használtuk annak mérésére, hogy milyen közel van egymáshoz a szimulációval és az aszimptotikus módszer alkalmazásával kapott valószínűségi eloszlás. Amint arra számítottunk N növekedésével a Kolmogorov-Smirnov távolság kisebb lett, bár ezen paraméter beállítás mellett szignifikáns csökkenést nem vettünk észre $N > 50$ esetén.

A 3.2. szakaszban a 2. fejezetben ismertetett általános modellt vizsgáltam azzal a kiegészítéssel, hogy szerverhiba esetén a kérések visszatérnek az orbitba. Érzékenységi elemzés elvégzéséhez számos egymástól eltérő eloszlást használtam a különböző vizsgált esetek rendszerjellemzőinek meghatározásához. Olyan metrikákat számoltattam ki, mint egy ügyfél átlagos várakozási ideje, egy ügyfél átlagos megszakítás nélküli kiszolgálási ideje vagy egy ügyfél átlagos megszakított kiszolgálási ideje. Ezeket összehasonlítottam az esetek között, hogy bemutassam a kiszolgálás, beérkezési és a visszatérési időközök eltérő eloszlásainak hatásait. A szimulációs program segítségével be tudtam mutatni minden szcenárió összes vizsgált eseténél, hogy a stacionárius eloszlások mindegyike normális eloszláshoz tart még akkor is, ha a visszatérési vagy beérkezési idő eloszlása nem exponenciális eloszlású. Ahogy azt vártam, az érkezési intenzitás növelésével az átlagos várakozási idő is növekszik, és az alkalmazott paraméter-beállítással a véges forrású visszatéréses sorbanállási rendszerek jellegzetes maximális érték tulajdonságát is tapasztaltam. Említésre méltó, hogy amikor a kiszolgálási idő exponenciális eloszlást követ, akkor az ügyfél átlagos megszakított kiszolgálási idejének és az ügyfél átlagos megszakítás nélküli kiszolgálás idejének összege megegyezik az 1 / kiszolgálási ráta paraméterrel. Ez a jelenség az ütközés jelenségéből adódik, mivel a szünetmentes kiszolgálási idők nagyon alacsony értékekből tevődnek össze, ami természetesen alacsony átlagot eredményez. Megfigyeltem, hogy a α és a β növekedésével minden egyes esetben a vizsgált rendszerjellemzők egyre nagyobb értékeket vesznek fel. Ezt a táblázatokban szereplő numerikus eredmények jól mutatják.

A következő szakaszban (lásd 3.3. szakasz) az előzőleg ismertetett modellt vettem alapul, de egy másik működési módot is megvizsgáltam, amennyiben egy igény kiszolgálása alatt szerver meghibásodás következne be:

- 1. üzemmód: a megszakított kérés azonnal az orbitba kerül (az előző szakaszban ez valósult meg).

- 2. üzemmód: a megszakított kérés kiszolgálása abbamarad, és a kiszolgáló egységnél fog tartózkodni. Annak helyrejövetele után a kiszolgálása folytatódik.

A két működési mód hatását vizsgáltam meg a legfontosabb rendszerjellemzők szempontjából. A 2. üzemmód alkalmazása mellett kedvezőbb eredményeket kaptam az 1. üzemmódnál kapott értékekhez képest, tehát alacsonyabb átlagos várakozás időt és nagyobb átlagos nem megszakított kiszolgálási időt. Több paraméter beállítás mellett vizsgáltam a rendszert és minden esetben az igények stacionárius eloszlása a rendszerben továbbra is normális eloszlást követett.

3.3. szakasz középpontjában az általános modell működésének további elemzése folytatódott. Itt a forrásban lévő egyedek eltérő viselkedésére került a hangsúly szerver meghibásodás alatt. Kétféle működési viselkedésmódot különböztettem meg:

- Blokkolás nélkül: az igények be tudnak jutni a rendszerbe, és ezek azonnal továbbítódnak az orbitba (teljesen megegyező 3.2. szakaszban leírtakkal).
- Blokkolással: az érkező igények nem tudnak bejutni a rendszerbe, így visszatérnek a forrásba, és új kérésgenerálási folyamat kezdődik el.

A blokkolás hatását tanulmányoztam néhány rendszerjellemző összehasonlításával. A vizsgált esetek stacionárius eloszlása továbbra is normál eloszlást követ, még blokkolás fennállása esetén is. A blokkolás a várt viselkedést eredményezi, ami alacsonyabb átlagos várakozási időben jelenik meg. Számos ábra érzékelteti, hogy az igények egyre több időt töltenek a forrásban, ha a kiszolgáló hajlamosabb a meghibásodásra blokkolás esetén. Nyilvánvalóan az igények forrásban eltöltött átlagos ideje állandó, amikor a rendszer nem blokkol.

4. fejezetben kétirányú kommunikációs visszatéréses sorbanállási rendszereket analizáltam, amelyek alapját 2.1. szakaszban bemutatott modell alkotja. Olyan szimulációs modellt alakítottam ki, hogy egy $M/G/1//N$ típusú kétirányú kommunikációs rendszert kapjak megbízható szerverrel, amely képes kimenő hívásokat kezdeményezni az orbitban tartózkodó ügyfelek felé a 4.1. szakaszban. Ez egy exponenciális eloszlású idő után következik be ν paraméterrel. Ezen kimenő ügyfelek kiszolgálási ideje gamma eloszlást követ α_2 és β_2 paraméterekkel. Kiszámítattam a bejövő ügyfelek különböző eloszlású kiszolgálási idejének hatásával kapcsolatos eredményeket, amikor az átlag és a szórásnégyzet megegyezik mindegyiknél. Ennek eléréshez a programom véletlenszám-generátorokkal lett kiegészítve, hogy a lehetséges összehasonlításhoz úgy tudjam megválasztani a paramétereket, hogy az első két momentum megegyezzen az összes eloszlás esetén. Egy illesztési folyamatot is végrehajtottam, mivel az egyes eloszlásoknál található paraméterek különböznek. Először azt az esetet vizsgáltam amikor

a relatív szórásnégyzet egynél nagyobb. Megmutattam az elsődleges ügyfelek átlagos várakozási idejét az érkezési intenzitás függvényében és az alkalmazott eloszlástól függetlenül maximális értéke lett. Ezek az eredmények azt is jól megmutatták, hogy az alkalmazott eloszlás nagyban befolyásolja a rendszerjellemzőket, a gamma- és Pareto-eloszlást összevetve látványos különbséget eredményezve. Az ábrák azt is szemléltetik, hogy a kiszolgáló kihasználtsága a kimenő igényekkel az érkezési intenzitás növekedésével csökken. Tanulmányoztam az orbitban tartózkodó igények stacionárius eloszlását, amely megmutatja annak valószínűségét, hogy hány ügyfél tartózkodik egyidejűleg az orbitban. A görbék részletesebb elemzése után azt találtam, hogy ezen eloszlások a normális eloszlásnak felelnek meg, és az is látható, hogy mennyire különböznek egymástól. A kétirányú kommunikációs rendszerek fontosságának hangsúlyozására a vizsgált modellt összehasonlítottam a klasszikus visszatéréses sorbanállási modellel, ahol nincs kimenő hívás. Ezen paraméterek mellett az ügyfelek kevesebb időt töltenek a rendszerben oly módon, hogy a kiszolgálási egység kihasználtsága sokkal magasabb, ami azt jelenti, hogy a vizsgált modellünk hatékonyabb, mint a klasszikus sorbanállási rendszer. Ez a jelenség azzal magyarázható, hogy a kiszolgálási egység kevesebb időt tölt el tétlen állapotban. Ez fennállt minden választott eloszlás esetén. Ezután úgy választottam meg a különböző kiszolgálási eloszlások paramétereit, hogy a relatív szórásnégyzet egynél kisebb legyen. Ebben az esetben azonban a kapott eredmények szinte megegyeztek egymással bármely rendszerjellemzőt nézve. Így arra a következtetésre jutottam, hogy ha a relatív szórásnégyzet egynél több, akkor a választott eloszlás jelentősen megváltoztatja a rendszer működését, míg amikor egynél kisebb, akkor minimális hatással van a teljesítményjellemzőkre.

A 4.2. szakaszban négy szcenáriót találtam ki a különböző típusú igények viselkedésére nézve. A célom az volt, hogy néhány eltérő kiszolgálási és meghibásodási idő eloszlás alkalmazásával összehasonlítsak fontosabb rendszerjellemzőket, mint az elsődleges igények átlagos várakozási idejét vagy a kiszolgálási egység kihasználtságát. A tétlen kiszolgáló egy exponenciális eloszlású időszak után γ paraméterrel képes kimenő hívásokat létesíteni egy (másodlagos) végtelen forrásból származó igények felé. Ezen ügyfelek kiszolgálása gamma-eloszlást követnek α_2 és β_2 paraméterekkel. Azonban ez akkor valósul meg, ha a másodlagos igény beérkezésének pillanatában nem érkezik elsődleges ügyfél a véges forrásból vagy az orbitból, és a szerver nem kerül meghibásodott állapotba. A korábban említett négy szcenárió a következőket tartalmazza szerver meghibásodás esetén:

- Első szcenárió: Az elsődleges ügyfelek azonnal az orbitba kerülnek, a másodlagos ügyfelek pedig nem jönnek be a rendszerbe.

- Második szcenárió: Az elsődleges ügyfelek azonnal az orbitba kerülnek, a másodlagos ügyfelek pedig a kiszolgálónál maradnak a kiszolgáló egység megjavítása során.

- Harmadik szcenárió: Az elsődleges ügyfelek a kiszolgálónál maradnak a kiszolgáló egység megjavítása során, a másodlagos ügyfelek pedig nem jönnek be a rendszerbe.
- Negyedik szcenárió: Mind az elsődleges és másodlagos ügyfelek a kiszolgálónál maradnak a kiszolgáló egység megjavítása során.

Első körben az alkalmazott bemeneti paraméterek értékeit úgy választottam meg, hogy relatív szórásnégyzet egynél nagyobb legyen. Az összehasonlításhoz hiperexponenciális, gamma, lognormális és Pareto eloszlást használtam. Kiszámítottam az elsődleges ügyfelek átlagos várakozási idejét és a szerver teljes kihasználtságát. Megvizsgáltam a különböző eloszlások hatását az egyes forgatókönyvekre és az eredmények egyértelműen jelezték, hogy amennyiben a relatív szórásnégyzet egynél nagyobb, akkor hatalmas az eltérés a rendszerjellemzők között annak ellenére, hogy az első két momentum megegyezik. Ezzel szemben nagyon kevés, ha a relatív szórásnégyzet egynél kisebb. Érdekes módon az összes forgatókönyv esetén jelentkezett az átlagos várakozási idő maximum tulajdonsága. Ugyanez a jelenség volt tapasztalható azokon az eredményeken is, amikor az érzékenységi vizsgálat a különféle meghibásodási idő eloszlások alapján történt.

6

Acknowledgements

Hereby I would like to thank all the people who contributed to the preparation of my dissertation directly or indirectly.

Special thanks go to my family for always being close to me and got me through in all the difficult moments that university life may have held. To my father and my mother for being a model and for raising me up and to my two sisters for always being supportive.

I am grateful to my supervisor, Sztrik János, for supporting me all the way from my Bachelor's work. Without his many helpful advice, guidance and encouragement this thesis could not have been completed.

I am thankful to Bérczes Tamás for his collaboration throughout the years. His guidance and instructions helped me a lot to advance and to participate in many projects.

Finally, I am forever beholden to my girlfriend Ágnes for all of her love, support, and patience even when I was mentally distracted or worked many hours at home.

Bibliography

[1] Ivo Adan and Jacques Resing. Queueing systems, 2015. URL: `https://www.win.tue.nl/~iadan/queueing.pdf`.

[2] Salah Aguir, Fikri Karaesmen, O Zeynep Akşin, and Fabrice Chauvet. The impact of retrials on call center performance. *OR Spectrum*, 26(3):353–376, 2004.

[3] A Aissani. A retrial queue with redundancy and unreliable server. *Queueing Systems*, 17(3-4):431–449, 1994.

[4] Zeynep Aksin, Mor Armony, and Vijay Mehrotra. The modern call center: A multi-disciplinary perspective on operations management research. *Production and operations management*, 16(6):665–688, 2007.

[5] Ahsan-Abbas Ali and Shuangqing Wei. Modeling of coupled collision and congestion in finite source wireless access systems. In *Wireless Communications and Networking Conference (WCNC)*, pages 1113–1118. IEEE, 2015.

[6] B. Almási, J. Roszik, and J. Sztrik. Homogeneous finite-source retrial queues with server subject to breakdowns and repairs. *Math. Comput. Modelling*, 42(5-6):673–682, 2005.

[7] J. R. Artalejo. New results in retrial queueing systems with breakdown of the servers. *Statistica Neerlandica*, 48(1):23–36, 1994.

[8] Jesus R Artalejo and Tuan Phung-Duc. Markovian retrial queues with two way communication. *Journal of industrial and management optimization*, 8(4):781–806, 2012.

[9] J.R. Artalejo and A. Gomez Corral. *Retrial Queueing Systems: A Computational Approach*. Springer, 2008.

[10] JR Artalejo and T Phung-Duc. Single server retrial queues with two way communication. *Applied Mathematical Modelling*, 37(4):1811–1822, 2013.

[11] K. Avrachenkov and Uri Yechiali. On tandem blocking queues with a common retrial queue. *Computers & OR*, 37:1174–1180, 07 2010.

[12] Kostia Avrachenkov and Uri Yechiali. Retrial networks with finite buffers and their application to internet data traffic. *Probability in the Engineering and Informational Sciences*, 22(4):519–536, 2008.

[13] Lawrence Brown, Noah Gans, Avishai Mandelbaum, Anat Sakov, Haipeng Shen, Sergey Zeltyn, and Linda Zhao. Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American statistical association*, 100(469):36–50, 2005.

[14] Edward Carlstein. The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *Ann. Statist.*, 14(3):1171–1179, 09 1986.

[15] E. Jack Chen and W. David Kelton. A procedure for generating batch-means confidence intervals for simulation: Checking independence and normality. *SIMULATION*, 83(10):683–694, 2007.

[16] Bong Dae Choi, Yang Woo Shin, and Wi Chong Ahn. Retrial queues with collision arising from unslotted CSMA/CD protocol. *Queueing Syst.*, 11(4):335–356, 1992.

[17] Ioannis Dimitriou. A retrial queue to model a two-relay cooperative wireless system with simultaneous packet reception. In *International Conference on Analytical and Stochastic Modeling Techniques and Applications*, pages 123–139. Springer, 2016.

[18] Velika Dragieva and Tuan Phung-Duc. Two-way communication M/M/1 retrial queue with server-orbit interaction. In *Proceedings of the 11th International Conference on Queueing Theory and Network Applications*, QTNA '16, New York, NY, USA, 2016. Association for Computing Machinery.

[19] Velika Dragieva and Tuan Phung-Duc. Two-way communication $M/M/1//N$ retrial queue. In *International Conference on Analytical and Stochastic Modeling Techniques and Applications*, pages 81–94. Springer, 2017.

[20] Velika I. Dragieva. Number of retrials in a finite source retrial queue with unreliable server. *Asia-Pac. J. Oper. Res.*, 31(2):23, 2014.

[21] Gennadi Falin. Model of coupled switching in presence of recurrent calls. 1979.

[22] G.I. Falin and J.R. Artalejo. A finite source retrial queue. *European Journal of Operational Research*, 108:409–424, 1998.

[23] Dieter Fiems and Tuan Phung-Duc. Light-traffic analysis of random access systems without collisions. *Annals of Operations Research*, pages 1–17, 2017.

[24] Paul A. Fishwick. Simpack: Getting started with simulation programming in c and c++. In *1992 Winter Simulation Conference*, pages 154–162, 1992.

[25] A. Francini and F. Neri. A comparison of methodologies for the stationary analysis of data gathered in the simulation of telecommunication networks. In *Proceedings of MASCOTS '96 - 4th International Workshop on Modeling, Analysis and Simulation of Computer and Telecommunication Systems*, pages 116–122, Feb 1996.

[26] Noah Gans, Ger Koole, and Avishai Mandelbaum. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5(2):79–141, 2003.

[27] Nawel Gharbi and Claude Dutheillet. An algorithmic approach for analysis of finite-source retrial systems with unreliable servers. *Computers & Mathematics with Applications*, 62(6):2535–2546, 2011.

[28] Nawel Gharbi and Malika Ioualalen. GSPN analysis of retrial systems with servers breakdowns and repairs. *Applied Mathematics and Computation*, 174(2):1151–1168, 2006.

[29] Nawel Gharbi, Lynda Mokdad, and Jalel Ben-Othman. A performance study of next generation cellular networks with base stations channels vacations. In *Global Communications Conference (GLOBECOM)*, pages 1–6. IEEE, 2015.

[30] Nawel Gharbi, Besma Nemmouchi, Lynda Mokdad, and Jalel Ben-Othman. The impact of breakdowns disciplines and repeated attempts on performances of small cell networks. *Journal of Computational Science*, 5(4):633–644, 2014.

[31] Antonio Gómez-Corral and Tuan Phung-Duc. Retrial queues and related models. *Annals of Operations Research*, 247(1):1–2, 2016.

[32] Frederick S. Hillier and Gerald J. Lieberman. *Introduction to operations research.* McGraw Hill Higher Education, 2014.

[33] Lyes Ikhlef, Ouiza Lekadir, and Djamil Aïssani. MRSPN analysis of Semi-Markovian finite source retrial queues. *Annals of Operations Research*, 247(1):141–167, 2016.

[34] Wang Jinting. Reliability analysis M/G/1 queues with general retrial times and server breakdowns. *Progress in Natural Science*, 16(5):464–473, 2006.

[35] Jeong Sim Kim. Retrial queueing system with collision and impatience. *Communications of the Korean Mathematical Society*, 25(4):647–653, 2010.

[36] Jeongsim Kim and Bara Kim. A survey of retrial queueing systems. *Annals of Operations Research*, 247(1):3–36, 2016.

[37] A. Krishnamoorthy, P. K. Pramod, and S. R. Chakravarthy. Queues with interruptions: a survey. *TOP*, 22(1):290–320, 2014.

[38] Attila Kuki, János Sztrik, Ádám Tóth, and Tamás Bérczes. A Contribution to Modeling Two-Way Communication with Retrial Queueing Systems . In *Information Technologies and Mathematical Modelling. Queueing Theory and Applications*, pages 236–247. Springer, 2018.

[39] V G Kulkarni and Bong Dae Choi. Retrial queues with server subject to breakdowns and repairs. *Queueing Systems*, 7(2):191–208, 1990.

[40] B Krishna Kumar, G Vijayalakshmi, A Krishnamoorthy, and S Sadiq Basha. A single server feedback retrial queue with collisions. *Computers & Operations Research*, 37(7):1247–1255, 2010.

[41] Anna Kvach and Anatoly Nazarov. *Sojourn Time Analysis of Finite Source Markov Retrial Queuing System with Collision*, chapter 8, pages 64–72. Springer International Publishing, Cham, 2015.

[42] A.S. Kvach. Numerical research of a Markov closed retrial queueing system without collisions and with the collision of the customers. In *Proceedings of Tomsk State University. A series of physics and mathematics. Tomsk*, volume 295 of *Materials of the II All-Russian Scientific Conference*, pages 105–112. TSU Publishing House, 2014. (In Russian).

[43] A.S. Kvach and A. Nazarov. Numerical research of a closed retrial queueing system M/GI/1//N with collision of the customers . In *Proceedings of Tomsk State University. A series of physics and mathematics. Tomsk*, volume 297 of *Materials of the III All-Russian Scientific Conference*, pages 65–70. TSU Publishing House, 2015. (In Russian).

[44] A.S. Kvach and A.A. Nazarov. The research of a closed RQ-system M/GI/1//N with collision of the customers in the condition of an unlimited increasing number of sources. In *Probability Theory, Random Processes, Mathematical Statistics and Applications: Materials of the International Scientific Conference Devoted to the 80th Anniversary of Professor Gennady Medvedev, Doctor of Physical and Mathematical Sciences,*, pages 65–70, 2015. (In Russian).

[45] Lamia Lakaour, Djamil Aïssani, Karima Adel-Aissanou, and Kamel Barkaoui. $M/M/1$ retrial queue with collisions and transmission errors. *Methodology and Computing in Applied Probability*, pages 1–12, 2018.

[46] Averill M. Law and W. David Kelton. *Simulation Modeling and Analysis*. McGraw-Hill Education, 1991.

[47] A. A. Nazarov and E. A. Sudyko. Method of asymptotic semiinvariants for studying a mathematical model of a random access network. *Problems of Information Transmission*, 46(1):86–102, Mar 2010.

[48] Anatoly Nazarov, Anna Kvach, and Vladimir Yampolsky. *Asymptotic Analysis of Closed Markov Retrial Queuing System with Collision*, chapter 1, pages 334–341. Springer International Publishing, Cham, 2014.

[49] Anatoly Nazarov, Janos Sztrik, and Anna Kvach. Comparative Analysis of Methods of Residual and Elapsed Service Time in the Study of the Closed Retrial Queuing System $M/GI/1//N$ with Collision of the Customers and Unreliable Server . In *International Conference on Information Technologies and Mathematical Modelling*, pages 97–110. Springer, 2017.

[50] Anatoly Nazarov, János Sztrik, and Anna Kvach. Some Features of a Finite-Source $M/GI/1$ Retrial Queuing System with Collisions of Customers. In *International Conference on Distributed Computer and Communication Networks*, pages 186–200. Springer, 2017.

[51] Anatoly Nazarov, János Sztrik, and Anna Kvach. A survey of recent results in finite-source retrial queues with collisions. In *Information Technologies and Mathematical Modelling. Queueing Theory and Applications*, pages 1–15. Springer, 2018.

[52] Anatoly Nazarov, János Sztrik, Anna Kvach, and Tamás Bérczes. Asymptotic analysis of finite-source M/M/1 retrial queueing system with collisions and server subject to breakdowns and repairs. *Annals of Operations Research*, 277(2):213–229, Jun 2019.

[53] Yi Peng, Zaiming Liu, and Jinbiao Wu. An M/G/1 retrial G-queue with preemptive resume priority and collisions subject to the server breakdowns and delayed repairs. *J. Appl. Math. Comput.*, 44(1-2):187–213, 2014.

[54] Tuan Phung-Duc and Wouter Rogiest. Two way communication retrial queues with balanced call blending. In *International Conference on Analytical and Stochastic Modeling Techniques and Applications*, pages 16–31. Springer, 2012.

[55] SV Pustova. Investigation of call centers as retrial queuing systems. *Cybernetics and Systems Analysis*, 46(3):494–499, 2010.

[56] J. Roszik. Homogeneous finite-source retrial queues with server and sources subject to breakdowns and repairs. *Ann. Univ. Sci. Budap. Rolando Eötvös, Sect. Comput.*, 23:213–227, 2004.

[57] Hiroyuki Sakurai and Tuan Phung-Duc. Two-way communication retrial queues with multiple types of outgoing calls. *Top*, 23(2):466–492, 2015.

[58] J. Sztrik. Tool supported performance modelling of finite-source retrial queues with breakdowns. *Publicationes Mathematicae*, 66:197–211, 2005.

[59] J. Sztrik, B. Almási, and J. Roszik. Heterogeneous finite-source retrial queues with server subject to breakdowns and repairs. *Journal of Mathematical Sciences*, 132:677–685, 2006.

[60] Toshiki Takeda and Takuya Yoshihiro. A distributed scheduling through queue-length exchange in CSMA-based wireless mesh networks. *Journal of Information Processing*, 25:174–181, 2017.

[61] Jinting Wang. Analysis of the finite source retrial queues with server breakdowns and repairs, 2011. URL: `http://aimsciences.org//article/id/d3e0b757-9d37-4555-99a9-c9b881c053dd`.

[62] Jinting Wang, Jinhua Cao, and Quanlin Li. Reliability analysis of the retrial queue with server breakdowns and repairs. *Queueing Systems*, 38(4):363–380, 1990.

[63] Jinting Wang, Linfei Zhao, and Feng Zhang. Performance analysis of the finite source retrial queue with server breakdowns and repairs. In *Proceedings of the 5th International Conference on Queueing Theory and Network Applications*, pages 169–176. ACM, 2010.

[64] Jinting Wang, Linfei Zhao, and Feng Zhang. Analysis of the finite source retrial queues with server breakdowns and repairs. *Journal of Industrial and Management Optimization*, 7(3):655–676, 2011.

[65] Andreas Willig. Performance evaluation techniques. 2005. URL: `https://pdfs.semanticscholar.org/6bf7/ecd9c2ba86ad4bb59bc6f33f9fe011b6efe3.pdf`.

[66] Tom Wolf. System and method for improving call center communications, November 30 2017. US Patent App. 15/604,068.

[67] Patrick Wüchner, János Sztrik, and Hermann de Meer. Finite-source retrial queues with applications. In *Proceedings of 8th International Conference on Applied Informatics, Eger, Hungary*, volume 2, pages 275–285, 2010.

[68] Tao Yang and Hui Li. The M/G/1 retrial queue with the server subject to starting failures. *Queueing Systems*, 16(1–2):83–96, 1994.

[69] Feng Zhang and Jinting Wang. Performance analysis of the retrial queues with finite number of sources and service interruptions. *Journal of the Korean Statistical Society*, 42(1):117–131, 2013.