University of West Bohemia
Faculty of Applied Sciences
Department of Mathematics

Bachelor Thesis

Priority Queueing Systems M/G/1

Pilsen, 2012 Hana Sedláková

Declaration

I hereby declare that this bachelor thesis is completely my own work and that I used only the cited sources.

In Pilsen May 28, 2012

signature

Acknowledgement

I would like to thank my supervisor, Ing. Jan Pospíšil Ph.D., for his support, suggestions and his patience that he devoted to me during the time of developing my thesis. Also I would like to thank my consultant from VUT Brno, Mgr. Jan Pavlík, Ph.D., for his pieces of advice. My one week internship at VUT in Brno was supported by the project A-Math-Net - knowledge transfer network in applied mathematics (project no. CZ.1.07/2.4.00/17.0100).

Preface

The subject of the bachelor thesis is queueing theory that means the mathematical study of queues. We introduce basic and necessary information about queueing systems. We especially focus on the systems that have a Poisson arrival process and general service time distribution that are called M/G/1 systems.

There is an option that systems have some type of priority. Priority queueing systems we study in the second part of the thesis.

Probably the most interesting part of this thesis could be the simulation of the process in the studied types of queueing systems. The simulations are created in MATLAB® and Simulink® that is a component of MATLAB.

Pilsen, May 28, 2012

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Queues are one of the most unpleasant part of our everyday lives. Unfortunately they occur everywhere, for example: at a doctor, in supermarket at a checkout counter, at a bank counter, at the school canteen... The entities that wait for service are called customers/users. Here the word customer is used in its generic sense, and thus maybe a job or a program in a computer system, a request in a database system... Customers who want service have to arrive at the service facility and ask the server for service demands. Typically queueing system has one service facility but there can be more than one server, and a waiting room of finite (or theoretically infinite) capacity. After arrival a customer waits in the waiting room/queue if all servers are busy. When some server becomes free, customer is chosen from the queue according to an order and when it is her turn, she is served. Then she leaves the queueing system. The basic queueing model is shown in Figure 1.1

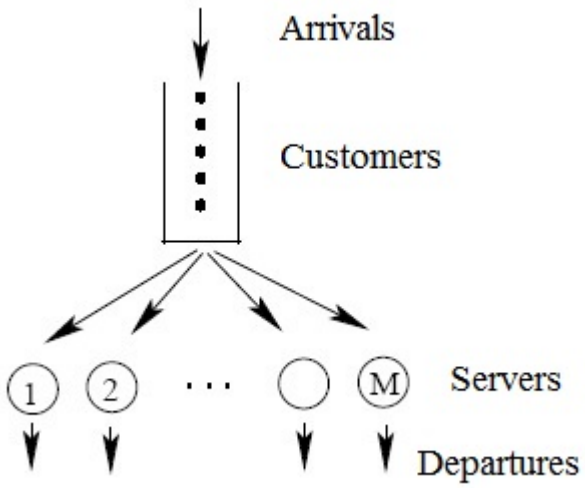

Figure 1.1: Basic queueing model

We have many possibilities how to illustrate behaviour in the queueing system. We can see one possibility in Figure 1.2 where τ_n is the time at which n^{th} customer arrives in the system, t_n is the interarrival time between the arrival of $(n-1)^{st}$ and n^{th} customer, x_n represents service time for n^{th} customer and w_n is waiting time for n^{th} customer.

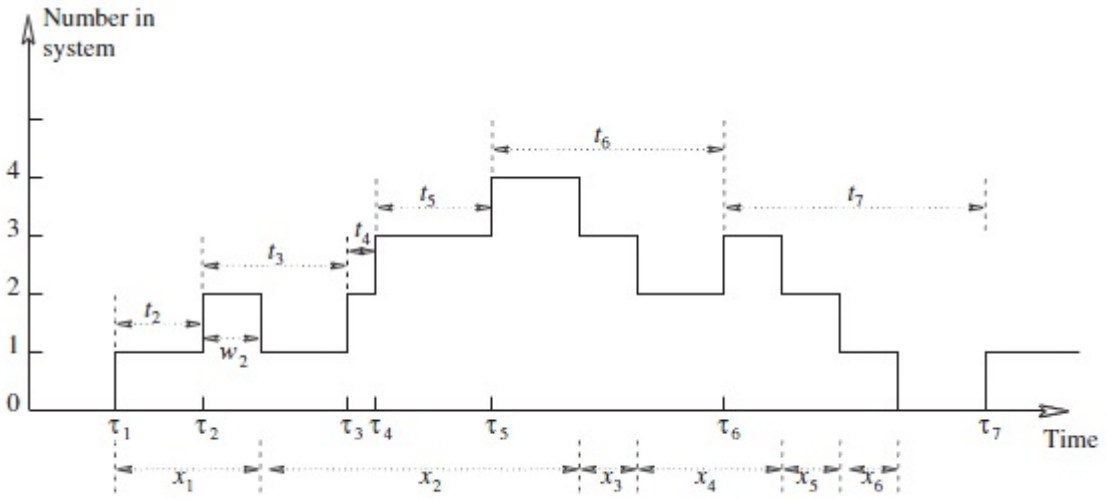

Figure 1.2: Graphical representation of behaviour in a single server queueing system

Our aim is to analyze the system in order to be able to make decisions such as how to optimize and upgrade the system. The analysis tell us about the expected time that a server will be in use, or the expected time that a customer must wait. At first we have to specify the manner in which arrivals occur, how the next customer who will be served is chosen from the queue, and so on.

The standard system used to describe and classify the queueing model is Kendall's notation. In 1951 D. G. Kendall (English mathematician) [5] suggested the first three-factor $A/B/C$ notation system. Later on A.M. Lee [8] extended the notation of D, E and H. A. Taha [12] added F. The meaning of these letters is in the Table 1.1 [1]

Table 1.1: Queueing system classification

A	The inter-arrival time distribution	**M** exponential inter-arrival distribution (Markovian); Poisson process $\mathbf{E}_k$ Erlang-k distribution **N** normal distribution **G** general distribution **D** deterministic, constant interarrival time
B	The service time distribution	the same as A
C	The number of servers	
D	The system capacity	the maximum number of customers allowed in the system including those in service
E	The size of calling source	the size of the population from which the customers come
F	The queue's discipline	FIFO, LIFO, SIRO, PS

More information about the queueing systems can be found in [1], [9] or [10]. Interesting reading about queueing system is also in [6] and [7]. In this thesis we mostly use

[1] 5th May 2012. http://en.wikipedia.org/wiki/Kendall%27s_notation

[2] and [11].

The aim of this thesis are Queueing Systems, specifically Priority Queueing Systems $M/G/1$. Necessary information about queueing systems we note in the first and second paragraphs of Introduction.

Before we study Priority Queueing Systems $M/G/1$, introduction of $M/G/1$ queueing system is needful. Chapter 3 is focused on this type of system and there are determinations of all important performance measures and information about system for example: the number of customer in the system, the time the customer spends waiting in the queue, residual service time of customer, etc. For easier orientation the summary of notation of these values is in Chapter 4.1.

Then we finally get to $M/G/1$ queueing systems with priorities. It means that there is a single-server system, customers arrive with rate λ, same as $M/G/1$ queueing system, but customers have some priority. The advantage is that the customer with high priority do not have to wait in the queue like the customers with lower priority. There are a lot of cases of priority policies. Two basic types, preemptive and nonpreemptive priority, are described in Chapter 4. In this chapter we also determine basic performance measures for the M/G/1 queueing system with priorities.

The last part of the thesis is focused on simulations of studied type of queueing system. We observe the processes in the $M/G/1$ system or we can see the behavior preferred and nonpreferred customers in the priority queueing system.

Chapter 2

Queueing Theory Notation of Performance Measures

"Roses are red;
Violets are blue
If λ is big
Then ρ is too."
(student's saying from [2])

In the following chapter we especially use [2].

Performance measure (or measure of effectiveness) is a term commonly used for a value of certain system property. Performance measures are all random variables. For example, we have:

- the number of customer in the system,
- the number of customers waiting in the queue,
- the time the customer spends waiting in the queue,
- the length of a busy period.

There is a summary of the basic queueing theory notation in the Table 2.1. Similar table supplemented of some other notations we can find in Chapter 5 in [2].

The most widely used formula in queueing theory is the Little's Law. It equates the number of customers in a system to the arrival rate multiplied by the time spend in the system. It has been written as

$$L = \lambda W, \tag{2.1}$$

where W is defined as a response time, λ is arrival rate and L is number of customers in the queueing system.

We can apply the Little's Law to the parts of queueing facilities, specifically to the queue and to the server. We have

$$L_q = \lambda W_q,$$
$$L_s = \lambda W_s,$$

and thus

$$L = L_q + L_s = \lambda W_q + \lambda W_s = \lambda(W_q + W_s) = \lambda W.$$

More details of the Little's Law can be found in [11], subsection 11.1.6.

PASTA (Poisson Arrivals See Time Averages) is an important property of the Poisson arrival process. Basically it means that the probability of the state as seen by an outside random observer is the same as the probability of the state seen by an arriving customer.[2] More about PASTA we can find for example in [11]. section 11.1.

Table 2.1: Summary of basic queueing theory notations

Symbol	Meaning	Relation
c	Number of identical servers	
L	Expected steady state number of customers in the system	$L = E[N] = \sum_{i=0}^{\infty} np_n$
L_q	Expected steady state number of customers in the queue	$L_q = E[N_q]$
L_s	Expected steady state number of customers receiving service	
μ	Mean service rate per server	
λ	Mean arrival rate of customers to the system	
N	Random variable describing the steady state number of customers in the system	
N_q	Random variable describing the steady state number of customers in the queue	
p_n	Steady state probability that there are n customers in the system	$p_n = Prob\{N = n\}$
ρ	Server utilization	$\rho = \frac{\lambda}{c\mu}$
S	Random variable describing the service time	$E[S] = \frac{1}{\mu}$
Q	Random variable describing the time a customer spends in the queue	
R	Random variable describing the total time a customer spends in the queueing system (response time)	$R = Q + S$
W	Response time (also sojourn time) is expected steady time that a customer spends in the system	$W = E[R] = W_q + W_s$
W_q	Expected steady state time that a customer spends in the queue	$W_q = E[W_q] = W - W_s$
W_s	Expected customer service time	$W_s = E[S]$

[2] 18th April 2012. http://en.wikipedia.org/wiki/Arrival_theorem

Chapter 3

Queueing Systems M/G/1

The main sources in this chapter are [2], [11], [10].
The $M/G/1$ queueing system is a single-server system where customers arrive according to a Poisson process with rate λ and its distribution function is $A(t) = 1 - e^{-\lambda t}, t \geq 0$ (Figure 3.1). The service times are independent and identically distributed with a general distribution function. The $M/G/1$ model is illustrated in Figure 3.2.

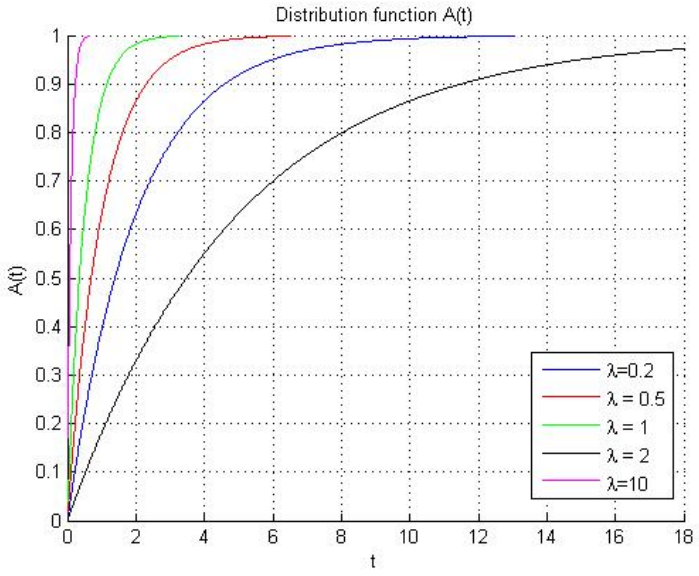

Figure 3.1: Simulation of $M/G/1$ queueing system with different λ

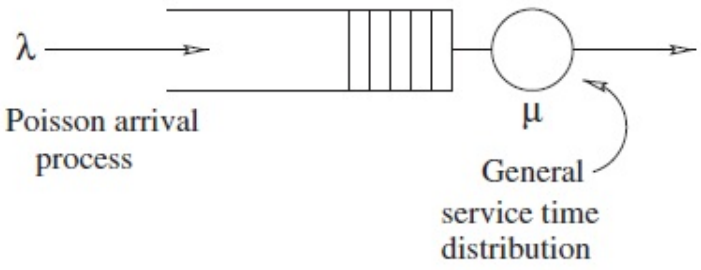

Figure 3.2: The $M/G/1$ queue

The mean service rate is denoted by μ, the service time distribution function is

$$B(x) = Prob\{x < S\},$$

where S is the random variable describing the service time and its density function is:

$$b(x)dx = Prob\{x < S \leq x + dx\}.$$

We use the notation from [11].

If $B(x)$ is the exponential distribution we have $M/M/1$ queueing system or if the service times are constant we obtain $M/D/1$ queueing system. These are the special cases of $M/G/1$ queueing system.

For this queueing system, the process $\{N(t), t \geq 0\}$, where $N(t)$ is the number of customers in the queue at time t, is not a Markov process since, when $N(t) \geq 1$, a customer is in service and the time already spent by that customer in service must be taken into account. It means that we must specify both:

(i) *N(t)*, the number of customers present at time t, and

(ii) *$S_0(t)$*, the service time already spent by the customer in service at time t.

Though $N(t)$ is not Markovian, $\{N(t), S_0(t)\}$ is a Markov process. The component $S_0(t)$ is called a supplementary variable. The embedded Markov chain approach permit us to substitute the two-dimensional state description $\{N(t), S_0(t)\}$ with a one-dimensional description N_k, where N_k is the number of customers that the k^{th} departing customer leaves behind.

Denote A_k the random variable describing the number of customers who arriving during the service time of the k^{th} customer. We modify a relationship in Chapter 14 in [11] for the number of customers left behind by the $(k+1)^{st}$ customer and we get:

$$N_{k+1} = \begin{cases} N_k - 1 + A_{k+1} & N_k = i > 0, \\ A_{k+1} & N_k = 0, \end{cases}$$

since there are N_k customers present in the system when the $(k+1)^{st}$ customer start the service. During serving this customer, A_{k+1} arrive. The number of customers in the system is reduced by 1 when this customer leaves.

When we define function $\delta(N_k)$ such that [11]

$$\delta(N_k) = \begin{cases} 1 & N_k > 0, \\ 0 & N_k = 0, \end{cases}$$

we can rewrite previous equation into single equation

$$N_{k+1} = N_k - \delta(N_k) + A. \tag{3.1}$$

Now we find the stochastic transition probability matrix F for the embedded Markov chain $\{N_k, k = 1, 2, 3, ...\}$. It is actually a system of matrices $F = f_{ij}(k)$ with

$$f_{ij}(k) = Prob(N_{k+1} = j | N_k = i).$$

It means that $f_{ij}(k)$ is the probability that the $(k+1)^{st}$ departing customer leaves behind j customers, given that the k^{th} departing customer leaves behind i customers.

Let p denote the stationary distribution of the Markov chain:

$$pF = p,$$

The j^{th} element of p represents the stationary probability of state j, it means the probability that a departing customer leaves j customers behind. Then the single-step transition probability matrix takes the form (matrix F is determined in [10]):

$$\mathbf{F} = \begin{pmatrix} \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & \dots \\ \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & \dots \\ 0 & \alpha_0 & \alpha_1 & \alpha_2 & \dots \\ 0 & 0 & \alpha_0 & \alpha_1 & \dots \\ \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix},$$

where α_i is the probability that i customers arrive during one service period. Since a departure cannot remove more than one customer, all elements in the matrix F for which $i > j+1$ must be zero (they lie below the subdiagonal). Since no customer can arrives during the service of the k^{th} customer, all elements lying above the diagonal are strictly positive. Therefore we may write:

$$\begin{aligned} Prob(N_{k+1} = j | N_k = i) &\equiv f_{ij}(k) = \alpha_{j-i+1} && \text{for } N_k = i > 0, j = i-1, i, i+1, i+2, \dots \\ &\equiv f_{0j}(k) = \alpha_j && \text{for } N_k = i = 0, j = 0, 1, 2, \dots \\ &\equiv 0 && \text{otherwise.} \end{aligned}$$

To calculate α_i, $i = 0, 1, 2 \dots$ we know that the number of customers that arrive during the service time is Poisson distributed with parameter λx. Hence, we have [11]

$$\alpha_i = \int_0^\infty \frac{1}{i!} (\lambda x)^i e^{-\lambda x} b(x) dx. \tag{3.2}$$

Unfortunately it does not tell us how to calculate the first row of F, it means the probabilities of transition from the state 0. The approach is the following: if there is no customer in the system when customer finishes service and leaves the system, then no state transition can occur until a new customer arrives; when this customer leaves the next transition occurs. It causes that transition probabilities are the same for $i = 0$ as for $i = 1$; first row and second row are identical.

The transition probability diagram is shown in Figure 3.3.

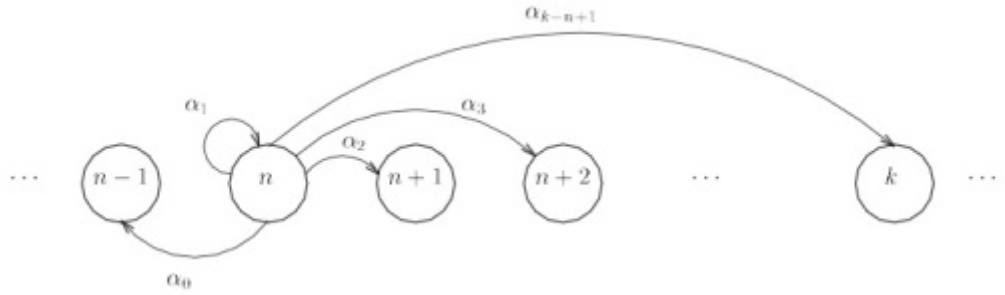

Figure 3.3: Transition probability diagram for the $M/G/1$ embedded Markov chain

3.1 Performance Measures

In this section we try to derive some important performance measures for the $M/G/1$ queueing system, for example number of customers in the system, time that customer spends waiting in the queue or total customer's time spent in the system.

3.1.1 The Pollaczek-Khintchine Mean Value Formula

```
"In a lobby in South Tennessee
Teenage Pollaczek gained his "esprit"
He watched as some guests
Made the lineups congest,
Then he left, humming Fi Fo, Fum Fee."
```
(Ben W. Lutek)

.....

We have a statement that the average number of arriving customers in a service period is equal to ρ, it can be written as

$$E[A] = \lim_{k\to\infty} Prob\{\text{server is busy}\} = E[\delta(N_k)] = \rho. \tag{3.3}$$

This equality will be use later and its complete deriving can be found in Chapter 14 in [11].

To get the mean number of customers in the system $M/G/1$ we should proceed as follows [11]. First if we square both side of (3.1). we get

$$\begin{aligned} N_{k+1}^2 &= N_k^2 + \delta(N_k)^2 + A^2 - 2N_k\delta(N_k) - 2\delta(N_k)A + 2N_kA \\ &= N_k^2 + \delta(N_k) + A^2 - 2N_k - 2\delta(N_k)A + 2N_kA. \end{aligned}$$

Then we take the expectation of each side and take the limit $k \to \infty$ where $N = \lim_{k\to\infty} N_k$. We make use of the relationship in equation (3.3) and we obtain

$$\begin{aligned} E[N_{k+1}^2] &= E[N_k^2] + E[\delta(N_k)] + E[A^2] - 2E[N_k] - 2E[A\delta(N_k)] + 2E[AN_k] \\ 0 &= E[\delta(N)] + E[A^2] - 2E[N] - 2E[A\delta(N)] + 2E[AN] \\ &= \rho + E[A^2] - 2E[N] - 2E[A]E[\delta(N)] + 2E[A]E[N] \\ &= \rho + E[A^2] - 2E[N] - 2\rho^2 + 2\rho E[N]. \end{aligned}$$

By rearranging the last equation we get

$$E[N](2-2\rho) = \rho + E[A^2] - 2\rho^2$$

which means

$$L = E[N] = \frac{\rho - 2\rho^2 + E[A^2]}{2(1-\rho)}. \tag{3.4}$$

Finally it only remains to find $E[A^2]$. We can use a statement from [11], section 14.3, that $E[A^2] = \rho + \lambda^2 E[S^2]$ where $E[S^2]$ is the second moment of service time distribution and we know that $E[S^2] = \sigma_s^2 + E[S]^2$, where σ_s^2 is the variance of the service time. When we use these relationship in the equation (3.4) we obtain

$$L = E[N] = \frac{\rho - 2\rho^2 + \rho + \lambda^2 E[S^2]}{2(1-\rho)} = \frac{2\rho(1-\rho) + \lambda^2 E[S^2]}{2(1-\rho)} =$$

$$= \rho + \frac{\lambda^2 E[S^2]}{2(1-\rho)} = \rho + \frac{\lambda^2(\sigma_s^2) + 1/\mu^2}{2(1-\rho)} = \rho + \rho^2 \frac{C_s^2 + 1}{2(1-\rho)} \tag{3.5}$$

where $C_s^2 = \mu^2 \sigma_s^2$. The equation (3.5) in any of the forms is called the Pollaczek-Khintchine mean value formula. Thanks to this formula we can get the average number of customers in the $M/G/1$ queueing system.

Using Little's formula (2.1), we can compute W, the expected time a customer spends in the system (response time). Thus

$$L = \lambda W$$
$$W = \frac{\rho}{\lambda} + \frac{\lambda^2 E[S^2]}{2\lambda(1-\rho)} = \frac{1}{\mu} + \frac{\lambda E[S^2]}{2(1-\rho)} = \frac{1}{\mu} + \frac{\lambda[(1/\mu)^2 + \sigma_s^2]}{2(1-\lambda/\mu)}.$$

We can also compute W_q the time a customer spends in the queue and L_q the number of customers in the queue using $W = W_q + W_s$ and $L_q = L - \rho$. We obtain

$$W_q = \frac{\lambda E[S^2]}{2(1-\rho)} = \frac{\lambda[(1/\mu)^2 + \sigma_s^2]}{2(1-\lambda/\mu)},$$
$$L_q = \frac{\lambda^2 E[S^2]}{2(1-\rho)} = \frac{\lambda^2[(1/\mu)^2 + \sigma_s^2]}{2(1-\lambda/\mu)}.$$

These equations are also known as the Pollaczek-Khintchine mean value formulae.

Figures 3.4 and 3.5 show the steady-state expected waiting time in an $M/G/1$ queueing system for a range of arrival rates λ. We can determine different mean service times $E[S]$ and the variances of the service time σ_s^2 and then we can observe the changing value of waiting time. These demonstration were taken from Wolfram Demonstrations Projects - Expected Time in System for M/G/1 Queue [3]

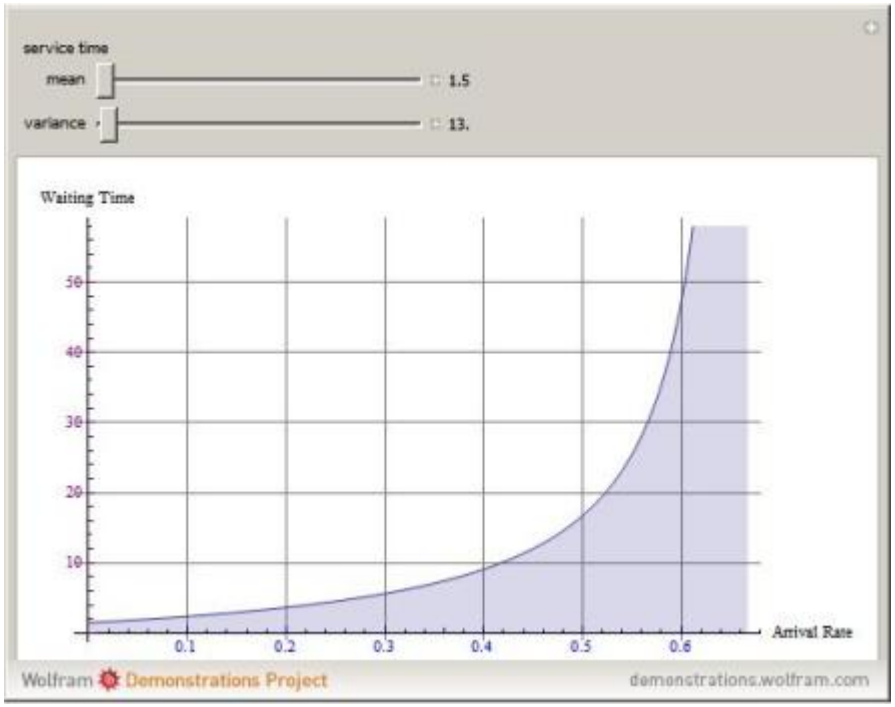

Figure 3.4: Expected waiting time for $E[S] = 0.15$, $\sigma_s^2 = 13$

[3] 6th April 2012. http://www.demonstrations.wolfram.com/ExpectedTimeInSystemForMG1Queue/

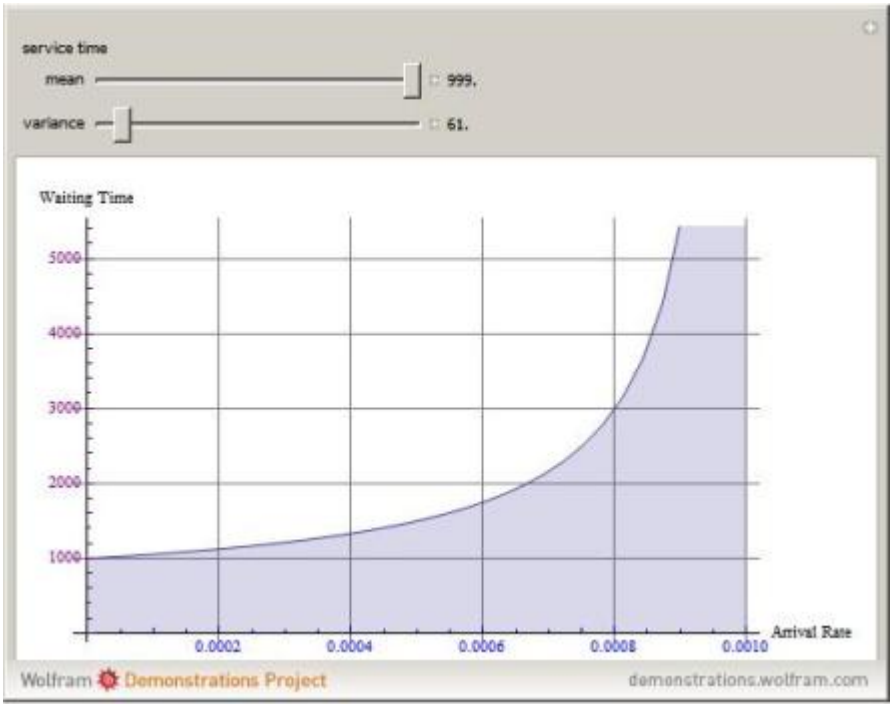

Figure 3.5: Expected waiting time for $E[S] = 999$, $\sigma_s^2 = 61$

3.1.2 The Pollaczek-Khintchine Transform Equations

We will show that queueing system has a steady state distribution of number of customers in the system and the distribution of response time. First we focus on the distribution of number of customers. To get a relationship for this distribution we will need the equation we mentioned at the beginning

$$p = pF$$

$$(p_0, p_1, p_2, ...) = (p_0, p_1, p_2, ...) \begin{pmatrix} \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & \dots \\ \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & \dots \\ 0 & \alpha_0 & \alpha_1 & \alpha_2 & \dots \\ 0 & 0 & \alpha_0 & \alpha_1 & \dots \\ \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix},$$

where p is a stationary distribution and p_j is for $j = 0, 1, 2...$ given by

$$p_j = p_0\alpha_j + \sum_{i=1}^{j+1} p_i\alpha_{j-i+1}.$$

If we multiply this equation by z^j we get

$$p_j z^j = p_0\alpha_j z^j + \frac{1}{z}\sum_{i=0}^{j+1} p_i\alpha_{j-i+1}z^{j+1} - \frac{p_0\alpha_{j+1}z^{j+1}}{z}$$

for $j = 0, 1, 2....$ Summing over j we have

$$\sum_{j=0}^{\infty} p_j z^j = \sum_{j=0}^{\infty} p_0\alpha_j z^j = \sum_{j=0}^{\infty}\sum_{i=1}^{j+1} p_i\alpha_{j-i+1}z^j. \tag{3.6}$$

We define the generating function of p and $\alpha = (\alpha_0, \alpha_1, \alpha_2...)$ by

$$P(z) = \sum_{j=0}^{\infty} p_j z^j$$

and

$$\alpha(z) = \sum_{j=0}^{\infty} \alpha_j z^j = G_A(z).$$

If we replace the double summation $\sum_{j=0}^{\infty}\sum_{i=1}^{j+1}$ with $\sum_{i=1}^{\infty}\sum_{j=i-1}^{\infty}$ and substitute generating function in the equation (3.6) we find

$$P(z) = p_0 G_A(z) + \frac{1}{z}[P(z) - p_0]G_A(z) = \frac{(z-1)p_0 G_A(z)}{z - G_A(z)}. \tag{3.7}$$

There are two unknowns in this equation ρ_0 $G_A(z)$. First we find ρ_0. Note that

$$P(1) = \sum_{j=0}^{\infty} p_j = 1 = \sum_{j=0}^{\infty} \alpha_j = G_A(1)$$

and when we use Theorem 2.9.2(Properties of the Generating function or z-transform)(c) from [2] and equation (3.3) we get

$$G'_A(1) = E[A] = \rho.$$

We find $\lim_{z\to 1} P(z)$ with applying L'Hôpital's rule and substitution of $G'_A(1) = \rho$. Then we obtain

$$1 = P(1) = \lim_{z\to 1} P(z) = \lim_{z\to 1}\left[p_0 \frac{(z-1)G'_A(z) + G_A(z)}{1 - G'_A(z)}\right] = p_0 \frac{1}{1 - G'_A(1)} = p_0 \frac{1}{1-\rho}$$

It means $p_0 = 1 - \rho$. It remains to derive the second unknown $G_A(z)$. For finding it we use the equation (3.2) so that

$$\begin{aligned} G_A(z) = \sum_{j=0}^{\infty} \alpha_j z^j &= \sum_{j=0}^{\infty} \int_0^{\infty} \frac{1}{j!}(\lambda x)^j z^j e^{-\lambda x} b(x)\,\mathrm{d}x \\ &= \int_0^{\infty} e^{-\lambda x} \sum_{j=0}^{\infty} \frac{1}{j!}(\lambda x z)^j b(x)\,\mathrm{d}x \end{aligned}$$

$$G_A(z) = \int_0^{\infty} e^{-\lambda x(1-z)} b(x)\mathrm{d}x = B^*[\lambda(1-z)], \tag{3.8}$$

where $B^*[\lambda(1-z)]$ is the Laplace transform of the service time distribution, $s = \lambda(1-z)$. Now we know p_0 and $G_A(z)$ so we can substitute into equation (3.7)

$$P(z) = \frac{(1-\rho)(z-1)B^*[\lambda(1-z)]}{z - B^*[\lambda(1-z)]}. \tag{3.9}$$

This equation is called Pollaczek-Khintchine transform equation No. 1. If we want to get Pollaczek-Khintchine transform equation No. 2 (we know the Laplace transform of

the distribution of response time of customer) we have to change our concern it means we focus on the number of arrivals during the response time of customers instead of the number of arrivals during the service time. Replacing α_j (the probability of j arrivals during service) with p_j (the probability of j arrivals during response time of customer) in equation (3.8) we get

$$\sum_{j=0}^{\infty} p_j z^j = P(z) = \sum_{j=0}^{\infty} \int_0^{\infty} \frac{1}{j!}(\lambda x z)^j e^{-\lambda x} w(x)\,\mathrm{d}x = \int_0^{\infty} e^{-\lambda x(1-z)} w(x)\mathrm{d}x = W^*[\lambda(1-z)], \tag{3.10}$$

where $w(x)$ is the probability density function of R and W^* is the Laplace transform of customer response time evaluated at $s = \lambda(1-z)$ in other words $z = 1 - \frac{s}{\lambda}$. Applying equation (3.10) into the equation (3.9) and substitution of z yields

$$W^*[\lambda(1-z)] = \frac{(1-\rho)(z-1)B^*[\lambda(1-z)]}{z - B^*[\lambda(1-z)]}$$
$$W^*(s) = B^*(s)\frac{s(1-\rho)}{s - \lambda + \lambda B^*(s)}.$$

This expression is known as the Pollaczek-Khintchine transform equation No.2.

3.2 Residual Time: Remaining Service Time

The residual service time (also forward recurrence time) is the time that remains until finishing the service. In other words it is the time that arriving customer has to wait if there is at least one customer in the process of being served. The time that has elapsed from the beginning service until current time is called backward recurrence time. The random variable describing residual service time we denote by $\mathcal{R}$, its probability density function is $f_{\mathcal{R}}(x) = \mu e^{-\lambda x}, x > 0$. If there is no customer in the system, $\mathcal{R} = 0$.

The mean residual service time can be found by using the Pollaczek-Khintchine mean value formula for W_q and L_q. Then the mean residual service we obtain from this relationship

$$E[\mathcal{R}] = W_q - \frac{1}{\mu}L_q = \frac{\lambda E[S^2]}{2(1-\rho)} - \frac{1}{\mu}\frac{\lambda^2 E[S^2]}{2(1-\rho)} = \frac{\lambda E[S^2]}{2(1-\rho)}(1-\rho) = \frac{\lambda E[S^2]}{2}.$$

Now we come to an interesting relationship between the mean residual service time and the expected time an arriving customer must wait in the queue i.e. $E[\mathcal{R}] = (1-\rho)W_q$, where ρ is the probability that server is busy, $(1-\rho)$ is probability that server is idle. $\mathcal{R}(t)$ is shown in Figure 3.6.

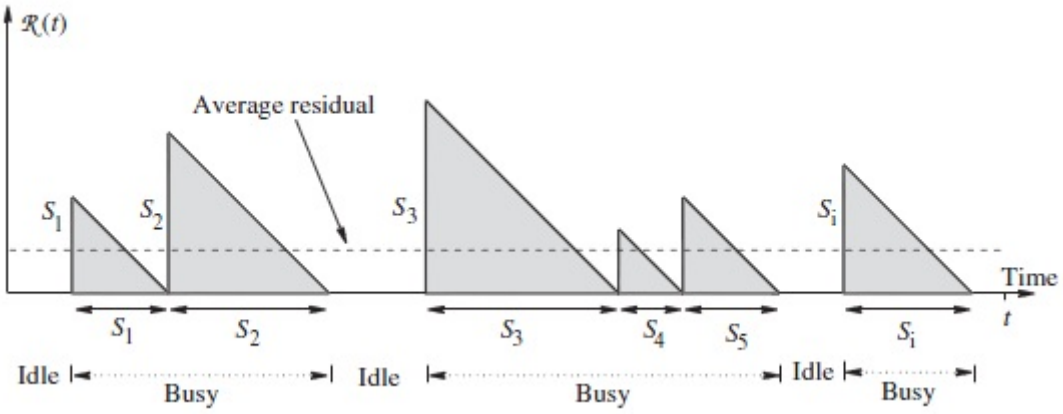

Figure 3.6: Residual service time in $M/G/1$ system

Let $\mathcal{R}_b$ denote the random variable that describes residual time that is conditioned on the server is busy. A number of approaches in section 14.4 from [11] may be used for finding the relationship

$$E[\mathcal{R}] = \rho E[\mathcal{R}_b]$$

where ρ is the probability that server is busy and

$$E[\mathcal{R}_b] = \frac{\mu E[S^2]}{2} = \frac{E[S^2]}{2E[S]}. \tag{3.11}$$

This argument applies also to the backward recurrence and it means that the mean backward recurrence time must also be equal to $E[S^2]/2E[S]$. Paradox of residual time is that the sum of forward and backward recurrence time isn't equal to the expected customer service time W_s.

Chapter 4

Priority Queueing Systems

In this chapter we follow these books: [11], [2].
Queueing systems in which some customers have preferential treatment are called priority queueing systems. We assume that customers in queues where they have no priority are served in first-come, first-served (FCFS) or first-in, first-out (FIFO) order. Other queueing disciplines in nonpriority systems are last-come, first served (LCFS or LIFO) and random-selection-for service (RSS) or service-in-random-order (SIRO). Biggest-in, first out (BIFO), first-in, still here (FISH) and whatever-in, never out (WINO) are part of the queueing theory folklore.

In the priority queueing systems customers are distinguished by priority classes, which are numbered from 1 to n. The lower the priority class number, the higher the priority. In other words, customer in priority class i is prefered over customer in priority class j if $i < j$. But the question is how a customer with priority j in service should be treated when a higher-priority customer with priority i arrives. To resolve this situation we have two basic cases of priority policies: preemptive and nonpreemptive priority.

In case of preemptive priority there is a rule: by the time a higher-priority customer arrives, the service with the lower-priority customer is interrupted. The new customer begins to be served and customer whose service was interrupted returns to the head of the j-th class. The interruption of service can cause a loss of progress and customer have to start her service from the beginning once again. This is called preemptive-repeat. In a preemptive-resume scenario a customer can continue at the point of interruption. In nonpreemtive priority system the arrived customer with higher priority may not interrupt the service time of a lower priority customer, she has to wait until the customer in service has been completed. A single server system with priority classes is illustrated in Figure 4.1.

More about the priority queueing system you can find in [3] or [4].

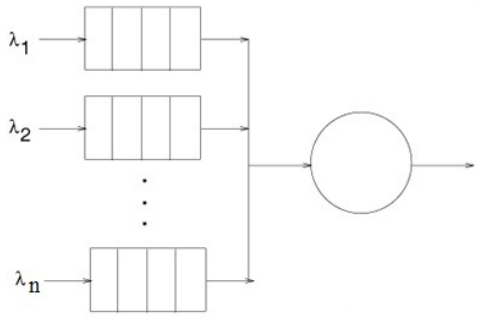

Figure 4.1: A single server system with priority classes

4.1 M/G/1 Nonpreemptive Priority Scheduling

Now we consider $M/G/1$ queueing system in which there are $J \geq 2$ different priority classes of customers where the classes can have different service requirements. We assume that the first priority class customer have higher priority than the customer of the second class, etc. Then we also assume that customers from class $j, j = 1, 2, ..., J$. arrive in a Poisson pattern with parameter λ_j, each class have the general service time distribution with probability density function $b_j(x), x \geq 0$ and $E[S_j] = 1/\mu_j$. Then $\rho_j = \lambda_j/\mu_j$ and we assume that $\lambda = \sum_{j=1}^{J} \lambda_j$ and $\rho = \sum_{j=1}^{J} \rho_j$. Now shal we denote (notation is used from Appendix C. in [2])

L_j	Mean number of a class j customer in the system
L_j^q	Mean number of a class j customer waiting in the queue
$E[R_j]$	Mean response time of a class j customer
W_j^q	Mean time of a class j customer spent waiting in the queue
$E[\mathcal{R}_j]$	Expected residual service time of a class j customer.

For computing the time of an arriving class j customer (tagged customer) that spends waiting in the queue we need to sum these three time period:

- the residual service time of the customer who is in service,
- the sum of all service times of customers of class 1 to j that are present at the moment when the tagged customer arrives,
- the sum of all service times of customers with higher priority (than tagged customer) who arrive during the tagged customer's waiting time in the queue.

Finding the first period is quite simple. We know that ρ_j is the probability that customer in service is of class j and next we know that the residual service time of any customer in service as seen by arriving customer is $E[\mathcal{R}_j]$. How to compute residual time we show in the equation (3.11) thus the searched expected residual service time of the customer in service as seen by tagged customer is

$$E[\mathcal{R}] = \sum_{i=1}^{J} \rho_i E[\mathcal{R}_i]. \tag{4.1}$$

For finding the second time period we need to use PASTA property that is mentioned in Chapter . We denote the mean number of customers waiting in the queue as L_i^q. Then the sum of service times of customers of 1 to j class found by the tagged customer is

$$\sum_{i=1}^{j} L_i^q E[S_i] = \sum_{i=1}^{j} \frac{L_i^q}{\mu_i}. \tag{4.2}$$

If we sum equation (4.1) and (4.2) we get expected residual service time of the customer in service but only if the tagged customer has the highest priority 1. Thus the result for this case is

$$W_1^q = L_i^q E[S_i] + \sum_{i=1}^{J} \rho_i E[\mathcal{R}_i]$$

After applying the Little's law i.e. $L_1^q = \lambda_1 W_1^q$ we obtain

$$W_1^q = \lambda_1 W_1^q E[S_i] + \sum_{i=1}^{J} \rho_i E[\mathcal{R}_i] = \rho_1 W_1^q + \sum_{i=1}^{J} \rho_i E[\mathcal{R}_i] = \frac{\sum_{i=1}^{J} \rho_i E[\mathcal{R}_i]}{1-\rho_1}. \tag{4.3}$$

From this equation we can compute L_1^q the mean number of customers waiting in the queue

$$L_1^q = \lambda_1 W_1^q$$
$$L_1^q = \lambda_1 \frac{\sum_{i=1}^{J} \rho_i E[\mathcal{R}_i]}{1-\rho_1}.$$

Now if the customer does not have the highest priority we have to determine the third time period, the sum of all service times of higher-priority customer who arrive during the time that tagged customer waiting in the queue. If W_j^q is the time of tagged customer of class j that spends in the queue then the sum of times spent by serving customers with higher priority who arrive during tagged customer's waiting is

$$\sum_{i=1}^{j-1} \frac{\lambda_i W_j^q}{\mu_i} = W_j^q \sum_{i=1}^{j-1} \rho_i.$$

Our sum of three time periods that represents the time that arriving class j customer spends waiting in the queue is the following

$$W_j^q = \sum_{i=1}^{J} \rho_i E[\mathcal{R}_i] + \sum_{i=1}^{j} \frac{L_i^q}{\mu_i} + W_j^q \sum_{i=1}^{j-1} \rho_i$$

We apply the Little's law i.e. $L_i^q = \lambda_i W_i^q$ and we obtain

$$\begin{aligned} W_j^q \left(1 - \sum_{i=1}^{j-1} \rho_i\right) &= \sum_{i=1}^{J} \rho_i E[\mathcal{R}_i] + \sum_{i=1}^{j} \frac{\lambda_i W_i^q}{\mu_i} \\ &= \sum_{i=1}^{J} \rho_i E[\mathcal{R}_i] + \sum_{i=1}^{j} \rho_i W_i^q \\ &= \sum_{i=1}^{J} \rho_i E[\mathcal{R}_i] + \sum_{i=1}^{j-1} \rho_i W_i^q + \rho_j W_j^q \\ W_j^q \left(1 - \sum_{i=1}^{j} \rho_i\right) &= \sum_{i=1}^{J} \rho_i E[\mathcal{R}_i] + \sum_{i=1}^{j-1} \rho_i W_i^q \end{aligned}$$

After comparing these equations we see that

$$W_j^q \left(1 - \sum_{i=1}^{j-1} \rho_i\right) = W_{j-1}^q \left(1 - \sum_{i=1}^{j-2} \rho_i\right). \tag{4.4}$$

If we multiply both sides of equation (4.4) with $\left(1 - \sum_{i=1}^{j-1} \rho_i\right)$ we get this recurrence

$$W_j^q \left(1 - \sum_{i=1}^{j} \rho_i\right) \left(1 - \sum_{i=1}^{j-1} \rho_i\right) = W_{j-1}^q \left(1 - \sum_{i=1}^{j-1} \rho_i\right) \left(1 - \sum_{i=1}^{j-2} \rho_i\right).$$

The repetition of application of this recursive relationship and using equation (4.3) yields

$$W_j^q \left(1 - \sum_{i=1}^{j} \rho_i\right) \left(1 - \sum_{i=1}^{j-1} \rho_i\right) = W_1^q (1 - \rho_1)$$

$$W_j^q = \frac{\sum_{i=1}^{J} \rho_i E[\mathcal{R}_i]}{1 - \sum_{i=1}^{j} \rho_i (1 - \sum_{i=1}^{j-1} \rho_i)}, j = 1, 2, ..., J.$$

This process of finding W_j^q, the mean time of a class j customer spent waiting in the queue, is from [11], subsection 14.6.1.

The mean response time of class j customer W_j we compute as

$$\begin{aligned} W_j &= W_j^q + W_j^s \\ &= \frac{\sum_{i=1}^{J} \rho_i E[\mathcal{R}_i]}{1 - \sum_{i=1}^{j} \rho_i \left(1 - \sum_{i=1}^{j-1} \rho_i\right)} + \frac{1}{\mu_j}, j = 1, 2, ..., J. \end{aligned}$$

Using the Little's law we determine the mean number of class j customers waiting in the queue L_j^q or the mean number of class j customers in the system L_j:

$$L_j^q = \lambda_j \frac{\sum_{i=1}^{J} \rho_i E[\mathcal{R}_i]}{1 - \sum_{i=1}^{j} \rho_i \left(1 - \sum_{i=1}^{j-1} \rho_i\right)}$$

$$L_j = \lambda_j \frac{\sum_{i=1}^{J} \rho_i E[\mathcal{R}_i]}{1 - \sum_{i=1}^{j} \rho_i \left(1 - \sum_{i=1}^{j-1} \rho_i\right)} + \frac{\lambda_j}{\mu_j} = \lambda_j \frac{\sum_{i=1}^{J} \rho_i E[\mathcal{R}_i]}{1 - \sum_{i=1}^{j} \rho_i \left(1 - \sum_{i=1}^{j-1} \rho_i\right)} + \rho_j.$$

4.2 M/G/1 Preemptive-Resume Priority Schedulling

Now we consider the preemptive-resume priority scheduling policy that means that a low-priority customer in service is interrupted by arriving customer of higher priority. Service of arriving customer begin immediately and the interrupted customer returns to the head of the j-th class and later she continues her service from the point at which that was interrupted.

We have two options, A and B, how to determine W_j^q, the time that a class j customer spends waiting in the queue. We begin with approach A.

First we compute T_1^A, the average time we need to serve all customers of equal or higher priority that are present at the moment when the tagged customer arrives to the system. Then we need to compute T_2^A, the time spent serving all customers of higher-priority who arrive during the total time that the tagged customer spends in the system, i.e., during the mean response time of customer j, $E[R_j]$. T_1^A is equal to

$$T_1^A = \sum_{i=1}^{j} \rho_i E[\mathcal{R}_i] + \sum_{i=1}^{j} E[S_i] L_i^q, \tag{4.5}$$

where$E[\mathcal{R}_i]$ is residual service time of a class i customer, $E[S_i]$ mean service time of a class i customer and L_i^q is the mean number of a class j customer found waiting in the queue at the equilibrium. We know that number of class i customers arriving during the time period T_1^A is $\lambda_i T_1^A$ and since the number of customers who are in the queue at a departure instant is the same as at an arrival instant we get $L_i^q = \lambda_i T_1^A$. We modify equation 4.5 and we obtain

$$\begin{aligned} T_1^A &= \sum_{i=1}^{j} \rho_i E[\mathcal{R}_i] + \sum_{i=1}^{j} \rho_i T_1^A \\ &= \frac{\sum_{i=1}^{j} \rho_i E[\mathcal{R}_i]}{1 - \sum_{i=1}^{j} \rho_i} \end{aligned}$$

Secondly we need to compute T_2^A. The number of class i customers arriving during the time period $E[R_j]$ is $\lambda_i E[R_j]$. Then the time to serve customers of higher-priority who arrive during $E[R_j]$ is the following

$$T_2^A = \sum_{i=1}^{j-1} \rho_i E[R_j] = E[R_j] \sum_{i=1}^{j-1} \rho_i = (W_j^q + 1/\mu_j) \sum_{i=1}^{j-1} \rho_i.$$

W_j^q the total waiting time of class j customer in the queue we get if we sum T_1^A, T_2^A

$$\begin{aligned} W_j^q = T_1^A + T_2^A &= \frac{\sum_{i=1}^{j} \rho_i E[\mathcal{R}_i]}{1 - \sum_{i=1}^{j} \rho_i} + (W_j^q + 1/\mu_j) \sum_{i=1}^{j-1} \rho_i \\ &= \frac{\sum_{i=1}^{j} \rho_i E[\mathcal{R}_i]}{\left(1 - \sum_{i=1}^{j} \rho_i\right)\left(1 - \sum_{i=1}^{j-1} \rho_i\right)} + \frac{1/\mu_j \sum_{i=1}^{j-1} \rho_i}{\left(1 - \sum_{i=1}^{j-1} \rho_i\right)}. \end{aligned}$$

Now we can express the mean response time of a class j customer

$$E[R_j] = W_j^q + W_j^s = W_j^q + \frac{1}{\mu_j} = \frac{\sum_{i=1}^{j} \rho_i E[\mathcal{R}_i]}{\left(1 - \sum_{i=1}^{j} \rho_i\right)\left(1 - \sum_{i=1}^{j-1} \rho_i\right)} + \frac{1/\mu_j}{\left(1 - \sum_{i=1}^{j-1} \rho_i\right)}.$$

Using the Little's law we obtain L_j, the mean number of class j customers that are in the system, and L_j^q, the mean number of class j customers waiting in the queue

$$\begin{aligned} L_j &= \frac{\lambda_j \sum_{i=1}^{j} \rho_i E[\mathcal{R}_i]}{\left(1 - \sum_{i=1}^{j} \rho_i\right)\left(1 - \sum_{i=1}^{j-1} \rho_i\right)} + \frac{\rho_j}{\left(1 - \sum_{i=1}^{j-1} \rho_i\right)} \\ L_j^q &= \frac{\lambda_j \sum_{i=1}^{j} \rho_i E[\mathcal{R}_i]}{\left(1 - \sum_{i=1}^{j} \rho_i\right)\left(1 - \sum_{i=1}^{j-1} \rho_i\right)} + \frac{\rho_j \sum_{i=1}^{j-1} \rho_i}{\left(1 - \sum_{i=1}^{j-1} \rho_i\right)}. \end{aligned}$$

Now we focus on the approach B. We have to compute T_1^B, the time spent waiting until the tagged class j customer go into a service for the first time, and T_2^B, the time spent in service and in interrupted period periods caused by higher-priority customers who arrive after the tagged customer first enters service.

The time spent waiting by a class j customer prior to entering service fort the first time is given by

$$T_1^B = \frac{\sum_{i=1}^{j} \rho_i E[\mathcal{R}_i]}{\left(1 - \sum_{i=1}^{j} \rho_i\right)\left(1 - \sum_{i=1}^{j-1} \rho_i\right)},$$

and then T_2^B has to be equal to

$$T_2^B = \frac{1/\mu_j \sum_{i=1}^{j-1} \rho_i}{\left(1 - \sum_{i=1}^{j-1} \rho_i\right)}.$$

If we sum T_1^B and T_2^B we get the desired result W_j^q that is the same as in approach A. Both of approaches A, B are described in subsection 14.6.2 from [11].

We compute and compare W_j^q, the mean time of class j customer spent waiting in the queue for both cases of priority policies in the following example. The example is taken from [11], the values are changed.

Example: Consider a queueing system which caters to three different classes of customers whose arrival processes are all Poisson. The most important customers require $E[S_1] = 1$ time unit of service and have a mean interarrival period of $1/\lambda_1 = 5$ time units. The corresponding values for classes 2 and 3 are $E[S_2] = 4$, $1/\lambda_2 = 16$ and $E[S_3] = 30$, $1/\lambda_3 = 60$.
First we need to compute $\rho_i, i = 1, 2, 3$. Thus $\rho_1 = 1/5$, $\rho_2 = 1/4$, $\rho_3 = 1/2$ and $\rho = \rho_1 + \rho_2 + \rho_3 = 0.95 < 1$.
To facilitate the computation of the residual service times, we shall assume that all service time distributions are deterministic. Thus $\mathcal{R}_1 = 0.5$, $\mathcal{R}_2 = 2$, and $\mathcal{R}_3 = 15$.
Then the times spent waiting in the queue by a customer of each classes are as follows:
nonpreemptive priority policy

$$W_1^q = \frac{\rho_1\mathcal{R}_1 + \rho_2\mathcal{R}_2 + \rho_3\mathcal{R}_3}{(1-\rho_1)} = \frac{8.1}{0.8} = 10.125$$

$$W_2^q = \frac{\rho_1\mathcal{R}_1 + \rho_2\mathcal{R}_2 + \rho_3\mathcal{R}_3}{(1-\rho_1-\rho_2)(1-\rho_1)} = \frac{8.1}{0.44} = 18.409$$

$$W_3^q = \frac{\rho_1\mathcal{R}_1 + \rho_2\mathcal{R}_2 + \rho_3\mathcal{R}_3}{(1-\rho_1-\rho_2-\rho_3)(1-\rho_1-\rho_2)} = \frac{8.1}{0.0275} = 294.5454$$

preempt-resume policy

$$W_1^q = \frac{\rho_1\mathcal{R}_1}{(1-\rho_1)} = \frac{0.1}{0.8} = 0.125$$

$$W_2^q = \frac{\rho_1\mathcal{R}_1 + \rho_2\mathcal{R}_2}{(1-\rho_1-\rho_2)(1-\rho_1)} + \frac{\rho_1/\mu_2}{1-\rho_1} = \frac{0.6}{0.44} + \frac{0.8}{0.8} = 2.3636$$

$$W_3^q = \frac{\rho_1\mathcal{R}_1 + \rho_2\mathcal{R}_2 + \rho_3\mathcal{R}_3}{(1-\rho_1-\rho_2-\rho_3)(1-\rho_1-\rho_2)} + \frac{(\rho_1+\rho_2)/\mu_3}{1-\rho_1-\rho_2} = \frac{8.1}{0.0275} + \frac{13.5}{0.55} = 319.0909.$$

Chapter 5

Simulations

In this chapter we simulate the processes in the queueing system. The simulations are created in MATLAB or Simulink (component of MATLAB).

5.1 M/G/1 Queueing Systems Simulation

First we simulate the process in an $M/G/1$ system. We illustrate the dependence of the number of customers waiting in the queue on time. System has a Poisson arrival process and general service time distribution. In Figure 5.1 this dependence is shown with different arrival rate λ of customers to the system, in particular for $\lambda = 0.2$, $\lambda = 0.5$, $\lambda = 0.9$. We determine the time $tmax = 60$ time units.

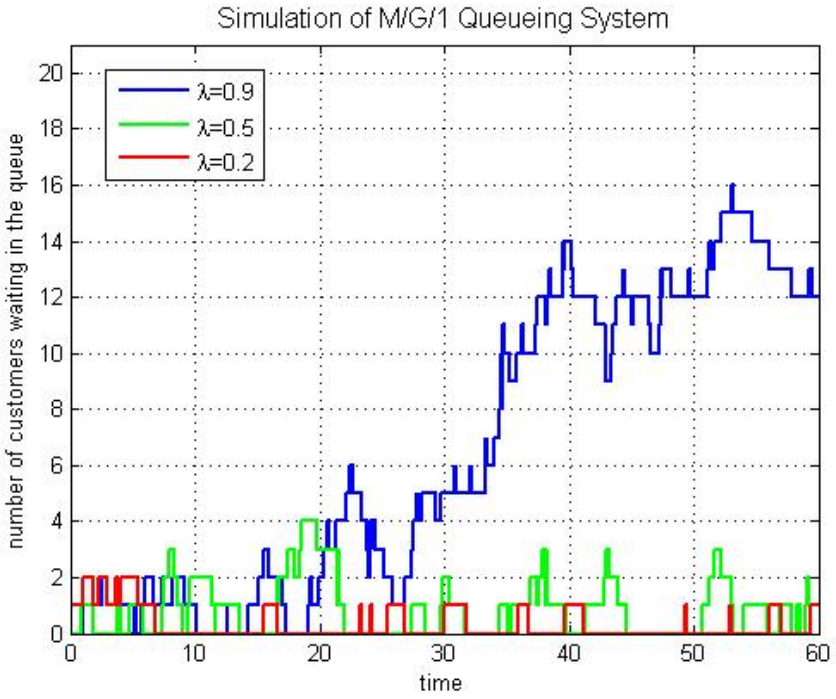

Figure 5.1: Simulation of $M/G/1$ queueing system with different λ

In Figure 5.2 we can see the dependence of the number of customers waiting in the queue on time with arrival rate $\lambda = 0.87$ for four different processes together. $Tmax = 60$ time units again.

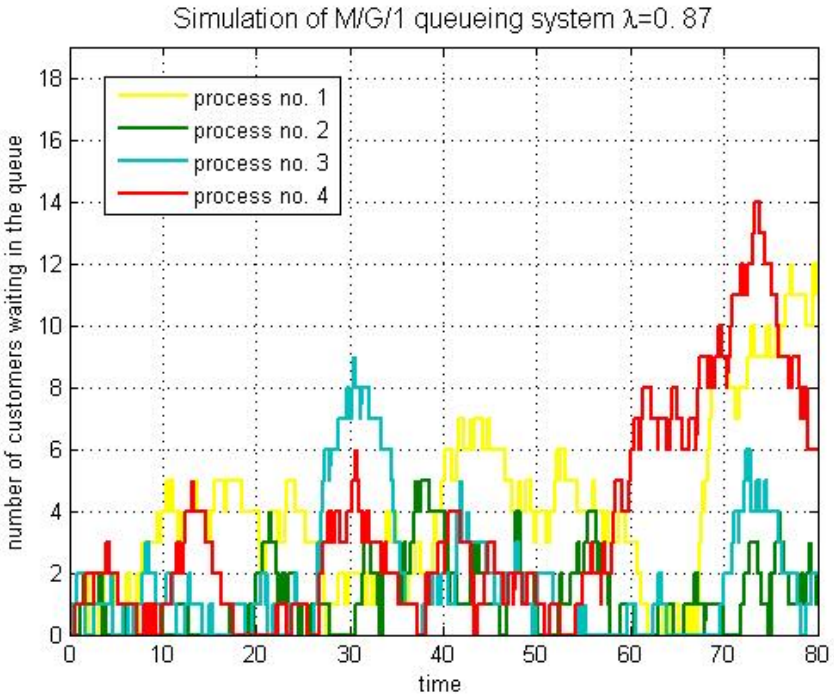

Figure 5.2: Simulation of 4 processes in $M/G/1$ queueing system with $\lambda = 0.87$

Matlab code used for these simulations was downloaded from MATLAB CENTRAL[4] and sligthly modified for our purposes.

```
%USAGE
function [jumptimes, systsize, systtime] = simmg1(tmax, lambda)
% SIMMG1 simulate a M/G/1 queueing system. Poisson arrivals
% of intensity lambda, uniform service times.
%
% Inputs: tmax - simulation interval
%         lambda - arrival intensity
%
% Outputs: jumptimes - time points of arrivals or departures
%          systsize - system size in M/G/1 queue
%          systtime - system times
% Original Authors: R.Gaigalas, I.Kaj

%THE CORE SIMULATION:
arrtime=-log(rand)/lambda;  % Poisson arrivals
i=1;
  while (min(arrtime(i,:))<=tmax)
    arrtime = [arrtime; arrtime(i, :)-log(rand)/lambda];
    i=i+1;
  end
n=length(arrtime);          % arrival times t_1,...,t_n

```

[4] 9th March 2012. http://www.mathworks.com/matlabcentral/fileexchange/?term=tag%3A%22mg1%22

```
servtime=2.*rand(1,n);       % service times s_1,...,s_k
cumservtime=cumsum(servtime);

arrsubtr=arrtime-[0 cumservtime(:,1:n-1)]';  % t_k-(k-1)
arrmatrix=arrsubtr*ones(1,n);
deptime=cumservtime+max(triu(arrmatrix));    % departure times
                                             % u_k=k+max(t_1,..,
                                                t_k-k+1)

% Output is system size process N and system waiting times W.
B=[ones(n,1) arrtime ; -ones(n,1) deptime'];
Bsort=sortrows(B,2);                 % sort jumps in order
jumps=Bsort(:,1);
jumptimes=[0;Bsort(:,2)];
systsize=[0;cumsum(jumps)];          % size of M/G/1 queue
systtime=deptime-arrtime';           % system times

% GRAPH:
figure(1)
title('Simulation of M/G/1 queueing system','color','k','
   fontsize',12)
xlabel('time','color','k','fontsize',10)
ylabel('number of customers waiting in the queue','color','k','
   fontsize',10)
stairs(jumptimes,systsize,'b');
xmax=max(systsize)+5;
axis([0 tmax 0 xmax]);
grid
```

5.2 M/G/1 Priority Queueing Systems Simulation

First we consider nonpriority system with gueueing disciplines LIFO (last-in, first-out) and FIFO (first-in, first-out). In this system customer also have preferential treatment. In Figure 5.3 we can see the model in Simulink for LIFO and FIFO queueing system. At the start of the simulation each model generate 19 entities and time of the simulation is 20 time units. In Figures 5.4 and 5.5 are shown graphs that represent the dependence of time and attribute Count, whose values are the entity's arrival sequence. In Figure 5.4 we can see an increasing sequence of Count values and in Figure 5.5 we can see a descending sequence of Count values. In this simulation, the servers do not permit preemption (preemptive servers would behave differently). This model was downloaded from MathWorks Product Documentation[5] and sligthly modified for our purposes.

[5] 10th March 2012. http://www.mathworks.com/help/toolbox/simevents/ug/a1076690284b1.html

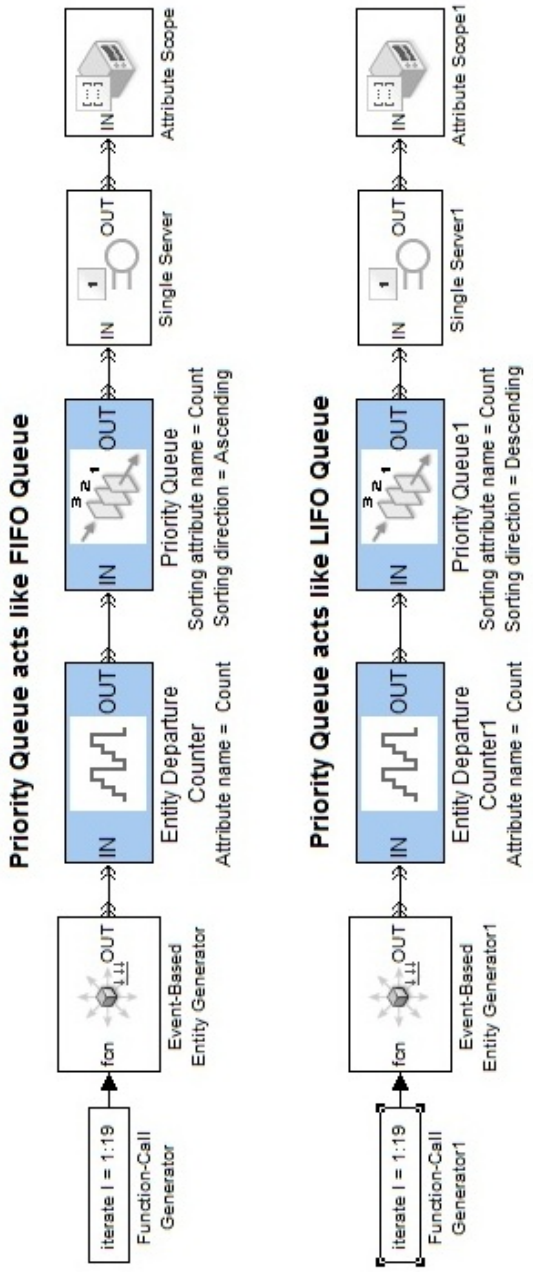

Figure 5.3: Model of priority queueing system act like LIFO and FIFO

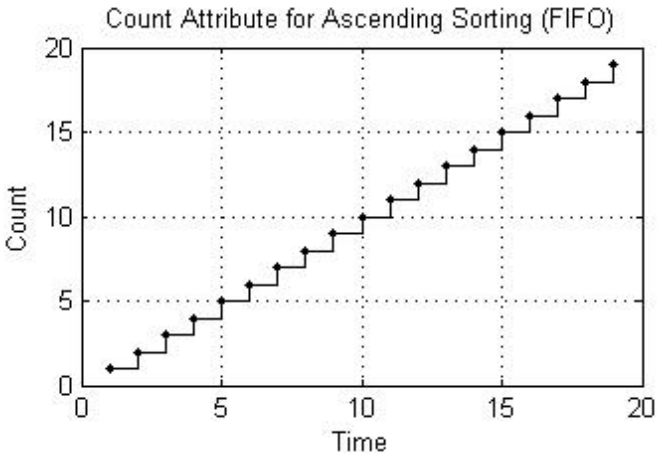

Figure 5.4: The FIFO plot

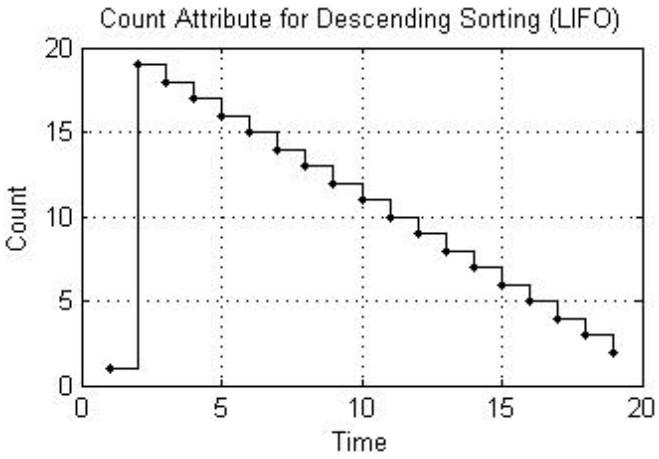

Figure 5.5: The LIFO plot

In the second simulation we have two types of customers: preferred and nonpreferred. Preferred customers are less common but they require longer service. Preferred customers are placed ahead of nonpreferred customers. Figure 5.6 represents the model of Serving Preferred Customers First in Simulink. In Figures 5.7 and 5.8 we can see the average system time for the set of preferred customers and for the set of nonpreferred customers. This model is also based on the MathWorks Product Documentation[5].

[5]10th March 2012. http://www.mathworks.com/help/toolbox/simevents/ug/a1076690284b1.html

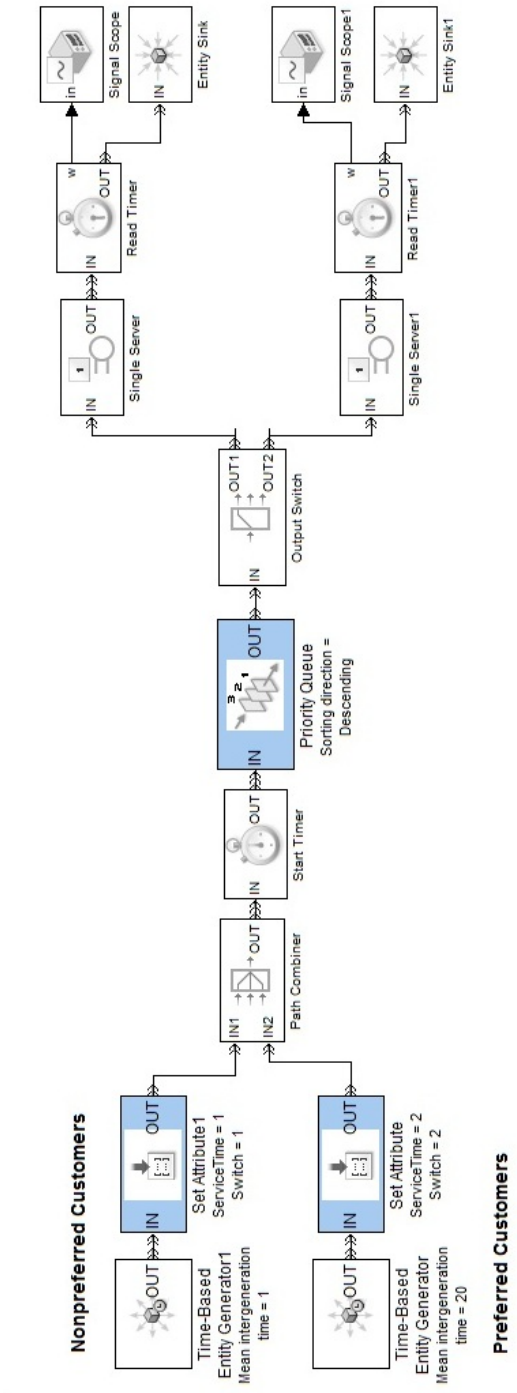

Figure 5.6: Serving Preferred Customers First

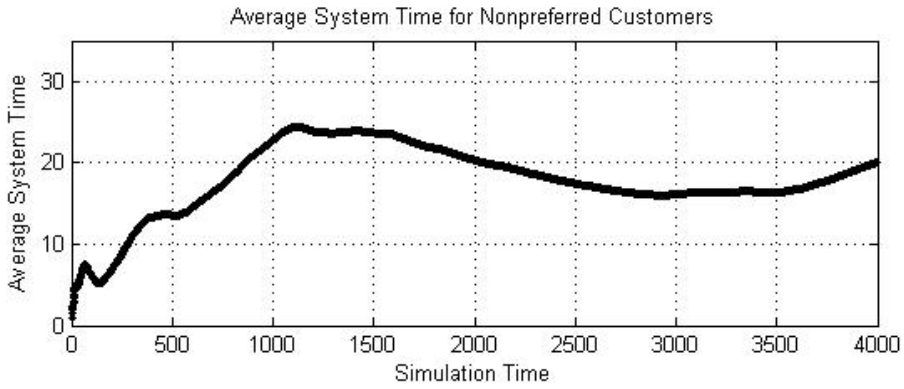

Figure 5.7: Average System Time for Nonpreferred Customers Sorted by Priority

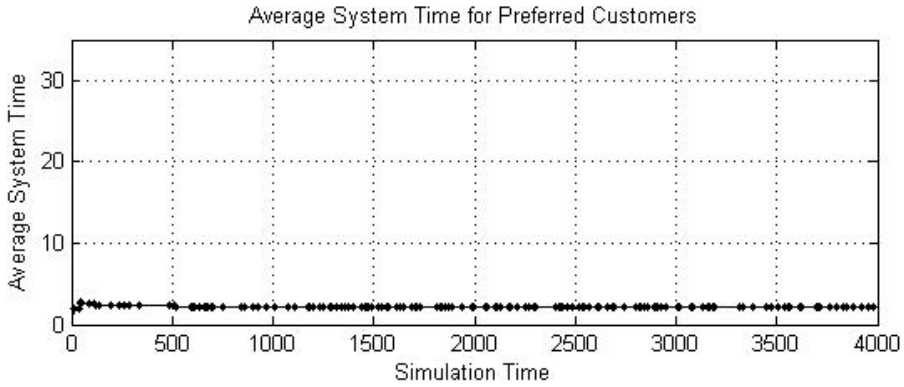

Figure 5.8: Average System Time for Preferred Customers Sorted by Priority

Chapter 6

Conclusion

This thesis describes the Priority Queueing Systems $M/G/1$. At first we introduced the queueing system in general for easier understanding of this topic. We mentioned a little of history of Queueing system classification and we recommended literature for more information about queueing systems.

In Chapter 4.1 we introduced the basic queueing theory notation and important property and formula in queueing theory. It should help with orientation between relationship later.

Before we started to focus on the $M/G/1$ queueing system with priority we described basic queueing system $M/G/1$ with no priority in Chapter 3. We determined well-known and probably the most important The Pollaczek-Khintchine Mean Value Formula that shows us how to compute the mean number of customers waiting in the queue, the mean number of customers in the system or expected steady state time that a customer spends in the queue. We found The Pollaczek-Khintchine Transform Equations no. 1 and no. 2 and we also determined the relationship for residual service time.

The aim of Chapter 4 is the topic of this thesis i.e., The Priority Queueing Systems $M/G/1$. After introducing basic information about these type of queueing system we focused on nonpreemptive and preemptive-resume priority scheduling. In both cases we determined the most important relationship such that: the mean number of a class j customer spent waiting in the queue, the mean number of a class j customer in the system or in the queue. At the end we compared W_j^q of nonpreemptive and preemptive-resume priority policies.

Simulations in Chapter 5 illustrate the processes in M/G/1 queueing systems. We can observe the dependences of the number of customers in the queue on time or in section 5.2 where we simulated priority queueing systems we can see how the FIFO (LIFO) queue behave. In the last simulation we have preferred and nonpreferred customers and we can compare the average system time for both of them.

In priority queuing system we can further study A Conservation Law and SPTF Scheduling. Basically it means when some classes of customers are privileged and have short waiting times, it is at the expense of other customers who pay for this by having longer waiting times. Under certain conditions, it may be shown that a weighted sum of the mean time spent waiting by all customer classes is constant, so that it becomes possible to quantify the penalty paid by low priority customers. More information about this scheduling we can find in [11], subsection 14.6.4.

Also The $M/G/1/K$ Queueing System is worth to note. It is special case of $M/G/1$ with maximum K number of customers in the system at any one time. For more details see [11], section 14.7.

Appendix A

Description of the thesis attachment

The attached CD contains:

- /Source_TEX/ - LATEX files used for generating the thesis, all the settings, graphics and bibliography files included
- /Matlab/queueing_system_simulation/ - includes Matlab scripts for simulation of M/G/1 Queueing Systems Simulations
- /Simulink/priority_queueing_system_simulation/ - includes Simulink models for simulation of Priority Queueing Systems Simulations
- thesis.pdf - the Bachelor thesis
- README.txt - the text file that includes information about the structure of CD

Bibliography

[1] Ivo Adan and Jacques Resing, *Queueing Theory*, Eindhoven University of Technology, Eindhoven, The Netherlands, 2002
URL http://www.win.tue.nl/ iadan/queueing.pdf.

[2] Arnold O. Allen, *Probability, Statistics and Queueing Theory with Computer Science Applications*, Academic Press, London, 1990, ISBN 0-12-051051-0.

[3] Francois Baccelli and Pierre Bremaud, *Elements of Queueing Theory: Palm Martingale Calculus and Stochastic Recurrences (Stochastic Modelling and Applied Probability)*, Springer Verlag, Berlin, 2010, ISBN 978-3-642-08537-6.

[4] John N. Daigle, *Queueing Theory with Applications to Packet Telecommunication*, Springer Science+Business Media, New York, 2005, ISBN 0-387-22857-8.

[5] David G. Kendall, *Stochastic Processes Occurring in the Theory of Queues and their Analysis by the Method of the Imbedded Markov Chain*, Annals of Mathematical Statistics **24(3)** (1953), 338–354.

[6] Leonard Kleinrock, *Queueing Systems. Volume 1: Theory*, John Wiley & Sons, Inc., Canada, 1975, ISBN 0-471-49110-1.

[7] Leonard Kleinrock and Richard Gail, *Queueing Systems: Problems and Solutions*, John Wiley & Sons, Inc., Canada, 1996, ISBN 0-471-55568-1.

[8] Alec Miller Lee, *Applied Queueing Theory*, MacMillan, New York, 1996, ISBN 0-333-04079-1.

[9] Philippe Nain, *Basic Elements of Queueing theory. Application to the Modelling of Computer Systems*, Lecture notes from the University of Massachusetts, Amherst, Massachusetts, United States, 1994,
URL http://www.cs.columbia.edu/ misra/COMS6180/nain.pdf.

[10] Zuzana Prášková and Petr Lachout, *Fundamental of random processes (Základy náhodných procesů)*, Karolinum, Prague, 2001.

[11] William I. Stewart, *Probability, Markov Chains, Queues, and Simulation*, Princeton University Press, New Jersey, 2009, ISBN 978-0-691-14062-9.

[12] Hamdy A. Taha, *Operations research: an introduction (Preliminary ed.)*, 1968.