Models and Simulations of Queueing Systems

Miloš Šeda[1(✉)], Jindřiška Šedová[2], and Miroslav Horký[1]

[1] Faculty of Mechanical Engineering, Brno University of Technology, Technická 2, 616 69 Brno, Czech Republic
seda@fme.vutbr.cz
[2] Faculty of Economics and Administration, Masaryk University, Lipová 41a, 602 00 Brno, Czech Republic
jsedova@econ.muni.cz

Abstract. In the queueing theory, it is assumed that requirement arrivals correspond to the Poisson process and the service time has the exponential distribution. Using these assumptions, the behaviour of the queueing system can be described by means of the Markov chains and it is possible to derive characteristics of the system. In the paper, these theoretical approaches are presented and focused on systems with several service lines and the FIFO queue when the number of requirements exceeds the number of lines. Finally, it is also shown how to compute the characteristics in a situation when these assumptions are not satisfied.

Keywords: Queue · Markovian chain

1 Introduction

The fundamentals of the queueing theory were laid by the Danish mathematician A. K. Erlang who worked for a telecommunication company in Copenhagen and in 1909 described an application of the probabilistic theory in telephone practice. The further development of the theory is mainly associated with the Russian mathematician A. N. Kolmogorov. The classification of queueing systems, as we use it today, was introduced in the 1950's by the English mathematician D.G. Kendall. Today, the queueing theory belongs to the classic part of logistics, and it is described in several monographs such as [1–4, 6, 8, 9].

Generally, at random moments, customers (requirements) enter the system and require servicing. Service options may be limited, e.g., the number of service lines (or channel operator). If at least one serving line is empty, the demand entering the system is immediately processed. However, the service time is also random in nature because the performance requirements may vary. If all service lines are busy, then the requirements (customers) must wait for their turn in a queue for the processing of previous requirements. However, all requirements are not always handled or queued for later use. For example, a telephone call is not connected because the phone number is busy.

Service lines are frequently arranged in parallel, e.g., at the hairdresser's where customers waiting for a haircut are served by several stylists, or at a gas station, where motorists call at several stands of fuel. However, there is a serial configuration of the queue system.

R. Matoušek (ed.), *Recent Advances in Soft Computing*, Advances in Intelligent Systems and Computing 576, DOI 10.1007/978-3-319-58088-3_2

2 Classification of Queueing Systems

The queue is usually understood in the usual **FIFO** sense – (*first in*, *first out*), but a **LIFO** operation (*last in*, *first out*) is also possible, which is also referred to as a **LCFS** (*last come*, *first served*) strategy.

Besides the FIFO and LIFO service, we can also meet random selection of requirements from the queue to the service system (**SIRO** - selection in random order) and service managed by priority requirements (**PRI** - Priority).

The *queue length* may be *limited* by rejecting additional requirements if a certain (predefined) number of requirements is achieved, such as the number of reservations for the book in a library that is currently checked out or (virtually) *unlimited*.

The requirements in the queue may have *limited* or *unlimited patience*. In the case of unlimited patience, requirements wait for their turn while in, a system with limited patience entering the queue significantly depends on the queue length. Instead of the queue length the concept of system *capacity* may also be used, which means the maximum number of requirements that may be present in the system.

In 1951, Kendall proposed a classification based on three main aspects in the form A/B/C, where

A characterises the probability distribution of random variable period (interval) between the requirement arrivals to the system,

B characterises the probability distribution of random variable *service time of a requirement*, and

C is the number of parallel service lines (or channel number), in the case of "unrestricted" (i.e. very large) number of lines is usual to express the parameter C by ∞.

As already mentioned, the system can be characterised by a larger number of features, so Kendall classification was further extended to the form A/B/C/D/E/F, where the meanings of the symbols D, E and F are as follows:

D integer indicating the maximum number of requirements in the system (i.e. the capacity of the system), unless explicitly restricted, expressed by ∞,

E integer expressing the maximum number of requirements in the input stream (or in a resource requirements), if it is unlimited, ∞ is used,

F queue type (FIFO/LIFO/SIRO/PRI).

Parameter A can have the following values:

M intervals between the arrivals of requirements are mutually stochastically independent and have exponential distribution, this means that the input stream represents a Poisson (Markov) process, for details, see below,

E_k Erlang distribution with parameters λ and k,

K_n χ^2 distribution with n degrees of freedom,

N normal (Gaussian) distribution,

U uniform distribution,

G general case, the time between the arrivals of requirements is given by its distribution function,
D intervals between the arrivals of demands are constant (they are deterministic in nature).

Parameter B can have the same value as parameter A, but these values here refer to a requirement-service-time random variable.

Since most of the queueing systems assume that the input current requirements are a *Poisson* (*Markov*) process, we will further describe it. A Poisson process is a stream of events that satisfies the following properties:

1. *Stationarity* (*homogeneity over time*) - the number of events in equally long time intervals is constant.
2. *Regularity* - the probability of more than one event at a sufficiently small interval of length Δt is negligibly small. This means that in, the interval $(t, t + \Delta t)$, there is either exactly one event with probability $\lambda\,\Delta t$ or no event with probability $1 - \lambda\,\Delta t$. In other words, in a Poisson process, the only system transition to the next "higher" state is possible or the system remains in the same condition.
3. *Independence of increases* - the number of events that occur in one time interval does not depend on the number of events in other intervals.

3 The M/M/n/nN/∞/FIFO System

In [7], we studied the M/M/1/∞/∞/FIFO system, here we focus on M/M/*n*/*n*/∞/FIFO system, which is more frequent in practice.

To derive the characteristics of the system, it is convenient to describe the system activity by a *graph of system transitions*. The nodes of the graph represent the states and the directed edges transitions from one state to another. The evaluation of these edges is described by the probability of transition from one state to another. State S_n or more specifically, $S_n(t)$ for fixed $t \in \langle 0, \infty)$, is a random variable and indicates that, at time *t, n* requirements are in the system. If *m* requirements, $m > n > 1$, are in the system M/M/*n*/*n*/∞/FIFO, then each requirement is operating in a service channel and the remaining *m*-*n* are waiting in the queue. Transitions between states that differ in the number of requirements in a system can be understood as a process of birth and death where the requirement birth represents the requirement entry into the system and death corresponds to a requirement leaving from the system after being processed.

We assume a Poisson stream of requirements with a parameter λ and an exponential distribution of service time with parameter μ, generally $\mu \neq \lambda$, and the queueing system behaviour described by the *Markov processes* (Fig. 1).

Due to the regularity, only transition probabilities $P(S_i \rightarrow S_j)$, where either $i = j$ or i and j differ by 1 have sense.

Using the regularity property and the method of calculating the total probability and neglecting the powers of the interval length Δt, from the partial probabilities of

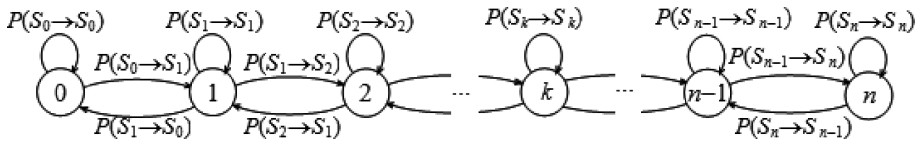

Fig. 1. Graph of M/M/n/n/∞ system transitions.

conjunction and disjunction of independent events, we get the transition probabilities as follows:

For example, transition probability $P(S_0 \rightarrow S_0)$ corresponds to the probability of the event that, during the time interval of length Δt, no requirement enters the system, transition probability $P(S_k \rightarrow S_{k-1})$, $n > k \geq 1$, is the probability of the event that, during the time interval of length Δt, no requirement enters the system and at the same time one requirement will be served and leaves the system, transition probability $P(S_k \rightarrow S_k)$, $k \geq 1$, is equal to the probability of the event that, during the time interval of length Δt, no requirement enters the system and no requirement leaves the system, or, during this interval, one requirement enters and one requirement will be served, transition probability $P(S_{k+1} \rightarrow S_k)$, $n > k \geq 1$, is equal to the probability of the event that, during the time interval of length Δt, one requirement was served in one of the service channels, i.e. $P(S_{k+1} \rightarrow S_k) = \mu\ \Delta t + \mu\ \Delta t + \ldots + \mu\ \Delta t = (k+1)\ \mu\ \Delta t$.

$P(S_n \rightarrow S_n)$ would differ in the M/M/n/n/∞/ (i.e. without the FIFO queue), because no requirement was served and no requirement entered a channel as all were occupied $P(S_n \rightarrow S_n) = 1 - (\mu\ \Delta t + \mu\ \Delta t + \ldots + \mu\ \Delta t) = 1 - n\ \mu\ \Delta t$.

Let us summarise the previous considerations:

$$P(S_{k-1} \rightarrow S_k) = \lambda \Delta t, k = 1, \ldots, n \tag{1}$$

$$\begin{aligned} P(S_k \rightarrow S_k) &= (1 - \lambda\,\Delta t)\,(1 - k\,\mu\,\Delta t) = 1 - (\lambda + k\,\mu)\Delta t + k\,\lambda\,\mu\,\Delta t^2 \\ &1 - (\lambda + k\,\mu)\,\Delta t, k = 0, \ldots \end{aligned} \tag{2}$$

$$P(S_{k+1} \rightarrow S_k) = (k+1)\mu\Delta t, k = 0, \ldots, n-1 \tag{3}$$

Let $p_k(t)$ denote the probability that, at time t, just k requirements are in the system. Using the previous equations, we can calculate $p_0(t)$, $p_1(t)$, …, $p_k(t)$, …, $p_n(t)$.

$$p_0(t + \Delta t) = P(S_0 \rightarrow S_0) + P(S_1 \rightarrow S_0) = p_0(t).(1 - \lambda\Delta t) + p_1(t).\mu\Delta t \tag{4}$$

$$\begin{aligned} p_1(t + \Delta t) &= P(S_0 \rightarrow S_1) + P(S_1 \rightarrow S_1) + P(S_2 \rightarrow S_1) \\ &= p_0(t).\lambda\,\Delta t + p_1(t).[1 - (\lambda + \mu)\Delta t] + p_2(t).2\mu\Delta t \ldots \end{aligned} \tag{5}$$

$$\begin{aligned} p_k(t + \Delta t) &= P(S_{k-1} \rightarrow S_k) + P(S_k \rightarrow S_k) + P(S_{k+1} \rightarrow S_k) \\ = p_{k-1}(t).\lambda\,\Delta t + p_k(t).[1 - (\lambda + k\mu)\Delta t] &+ p_{k+1}(t).(k+1)\mu\Delta t, \quad k = 2, \ldots, n-1 \end{aligned} \tag{6}$$

However, if all channels are occupied and the queue is nonempty, the last equation changes to (7).

$$p_k(t+\Delta t) = p_{k-1}(t).\lambda\,\Delta t + p_k(t).[1-(\lambda+n\mu)\Delta t] + p_{k+1}(t).n\mu\Delta t, k \geq n \quad (7)$$

After easy simplification of Eqs. (4), (6) and (7), we obtain Eqs. (8), (9) and (10).

$$\frac{p_0(t+\Delta t)-p_0(t)}{\Delta t} = -\lambda p_0(t) + \mu p_1(t) \quad (8)$$

$$\frac{p_k(t+\Delta t)-p_k(t)}{\Delta t} = \lambda p_{k-1}(t) - (\lambda+k\mu)\,p_k(t) + (k+1)\mu p_{k+1}(t), \quad k = 1,2,\ldots,n-1 \quad (9)$$

$$\frac{p_k(t+\Delta t)-p_k(t)}{\Delta t} = \lambda p_{k-1}(t) - (\lambda+n\mu)\,p_k(t) + n\mu p_{k+1}(t), \quad k \geq n \quad (10)$$

Let us now consider a limit transition for $\Delta t \to 0$ in Eqs. (22) and (23). We get:

$$\lim_{\Delta t\to 0}\frac{p_0(t+\Delta t)-p_0(t)}{\Delta t} = \lim_{\Delta t\to 0}[-\lambda p_0(t) + \mu p_1(t)]$$

$$\lim_{\Delta t\to 0}\frac{p_k(t+\Delta t)-p_k(t)}{\Delta t} = \lim_{\Delta t\to 0}[\lambda p_{k-1}(t) - (\lambda+k\mu)\,p_k(t) + (k+1)\mu p_{k+1}(t)], k = 1,2,\ldots,n-1$$

$$\lim_{\Delta t\to 0}\frac{p_k(t+\Delta t)-p_k(t)}{\Delta t} = \lim_{\Delta t\to 0}[\lambda p_{k-1}(t) - (\lambda+n\mu)\,p_k(t) + n\mu p_{k+1}(t)], \quad k \geq n$$

The expressions on the left-hand side of the previous two equations are derivatives of the functions $p_0(t)$ and $p_k(t)$ at point t, i.e. $p_0'(t)$ and $p_k'(t)$, while, on their right-hand sides, the limit transition does not have any effect. Hence, we get recurrence Eqs. (11), (12) and (13) as follows:

$$p_0'(t) = -\lambda p_0(t) + \mu p_1(t) \quad (11)$$

$$p_k'(t) = \lambda p_{k-1}(t) - (\lambda+k\mu)\,p_k(t) + (k+1)\mu p_{k+1}(t), \quad k = 1,2,\ldots,n-1 \quad (12)$$

$$p_k'(t) = \lambda p_{k-1}(t) - (\lambda+n\mu)\,p_k(t) + n\mu p_{k+1}(t), \quad k \geq n \quad (13)$$

These recurrence equations are a set of infinitely many first-order ordinary differential equations. To address them, we need to know the initial conditions, which are given by the state of a system at time $t_0 = 0$. If there are k_0 requirements in a system at time $t_0 = 0$, then the initial conditions are given by (14) and (15)

$$p_{k_0}(0) = 1 \quad (14)$$

$$p_k(0) = 0, \quad k \geq 1,\ k \neq k_0 \quad (15)$$

In the sequel, we assume that $\lambda < \mu$, i.e. $\lambda/\mu < 1$. Denote the ratio λ/μ by the ψ symbol. We call it the *intensity of the system load*. The condition

$$\frac{\lambda}{n\mu} < 1 \tag{16}$$

is a necessary and sufficient condition for a queue not to grow beyond all bounds. This condition also ensures that, after a sufficiently long time from the opening of a queueing system, its situation stabilizes, i.e., there are limits

$$\lim_{t\to\infty} p_k(t) = p_k, \quad k = 0, 1, \ldots, \tag{17}$$

and then, after a sufficiently long time from the opening of a queueing system, the probabilities $p_k(t)$ can be regarded as constant, i.e.,

$$p_k(t) = p_k = \text{const} \tag{18}$$

Since the derivatives of constants are zeros, by Eqs. (11), (12) and (13), we get an infinite set of linear algebraic equations determined by (19), (20) and (21).

$$0 = -\lambda p_0 + \mu p_1 \tag{19}$$

$$0 = \lambda p_{k-1} - (\lambda + k\mu) p_k + \mu p_{k+1}, \quad k = 1, 2, \ldots, n-1 \tag{20}$$

$$0 = \lambda p_{k-1} - (\lambda + n\mu) p_k + n\mu p_{k+1}, \quad k \geq n \tag{21}$$

From the above system of equations, we get:

$$p_k = \frac{\psi^k}{k!} p_0, \quad k = 1, \ldots, n-1 \tag{22}$$

$$p_k = \frac{\psi^k}{n!} \frac{n^n}{n^k} p_0, \quad k \geq n \tag{23}$$

Obviously, we have

$$\sum_{k=0}^{\infty} p_k = 1 \tag{24}$$

Now p_0 remains to be found. To do this, we use Eqs. (19), (20) and (21).

$$p_0 \left[\sum_{k=0}^{n-1} \frac{\psi^k}{k!} + \sum_{k=n}^{\infty} \frac{\psi^k}{n! n^{k-n}} \right] = 1 \tag{25}$$

The second expression in brackets may be simplified using a geometric series with quotient ψ/n as follows

$$\sum_{k=n}^{\infty}\frac{\psi^k}{n!n^{k-n}}=\frac{\psi^n}{n!}\sum_{k=n}^{\infty}\frac{\psi^{k-n}}{n^{k-n}}=\frac{\psi^n}{n!}\sum_{k=n}^{\infty}\left(\frac{\psi}{n}\right)^{k-n}=\frac{\psi^n}{n!}\sum_{i=0}^{\infty}\left(\frac{\psi}{n}\right)^{i}=\frac{\psi^n}{n!}\frac{1}{1-\frac{\psi}{n}}$$

Therefore, by (25), we get

$$p_0=\left[\sum_{k=0}^{n-1}\frac{\psi^k}{k!}+\frac{\psi^n}{n!}\frac{1}{1-\frac{\psi}{n}}\right]^{-1} \tag{26}$$

These equations may now be used to derive other important characteristics of the M/M/*n*/*n*/∞/FIFO system, which include:

1. *Mean number of requirements in the system*:

$$E(N_s)=\overline{n_s}=\sum_{k=0}^{\infty}k\,p_k \tag{27}$$

2. *Mean number of requirements in the queue* (*mean queue length*):

$$E(N_f)=\overline{n_f}=\sum_{k=n}^{\infty}(k-n)p_k \tag{28}$$

3. *Mean number of free service channels*:

$$E(N_c)=\overline{n_c}=\sum_{k=0}^{n-1}(n-k)p_k \tag{29}$$

4. *Mean time spent by a requirement in the system*:

$$E(T_s)=\overline{t_s}=\frac{\overline{n_s}}{\lambda} \tag{30}$$

5. *Mean waiting time of a requirement in the queue*:

$$E(T_f)=\overline{t_f}=\frac{\overline{n_f}}{\lambda} \tag{31}$$

6. *Factor of service channel idle time*

$$K_0=p_0 \tag{32}$$

7. *Factor of service channel load*

$$K_1=1-p_0 \tag{33}$$

4 Simulation of Queueing Processes

As, in practice, some assumptions may not be satisfied, particularly the Poisson (Markov) process properties of stationarity and the independence of increases, such as the number of clients in shops and railway stations substantially changing during the daytime, the formulas that we have derived, may not be entirely accurate. However, queueing systems can also be studied by Monte Carlo simulations, which generate random numbers representing the moment of the requirements entering into the system and the service time.

If the values of these random variables should have a certain probability distribution, then it must be provided.

To do this there are many methods, such as the *elimination method* and *inverse function method*. The elimination method may be employed to generate the values of continuous random variables whose probability density f is bounded in an interval $\langle a, b\rangle$ and zero outside this interval. The principle of this method is based on the fact that we generate random points with coordinates (x, y) with uniform distribution in the rectangle $\langle a, b\rangle \times \langle 0, c\rangle$, where c is the maximum value of the probability density f in the interval $\langle a, b\rangle$.

If the point generated is under the graph of the function f, i.e., $y \leq f(x)$, then x is regarded as the generated value of a random variable with the given distribution, otherwise it is not, that it, is discarded from the calculations. In the inverse function method, we first determine the probability distribution function F from the density function f by Eq. (34).

$$F(x) = \int_{-\infty}^{x} f(t)\, dt \tag{34}$$

We generate a random number r with uniform distribution on the interval $\langle 0, 1\rangle$, that we consider as the value of the distribution function at a still unknown point x, i.e., $F(x) = r$. The point x here is obtained by the inverse function (35):

$$x = F^{-1}(r) \tag{35}$$

During the simulation experiments, it is necessary to decide how to express the dynamic properties of the model, i.e., what strategy you choose for recording time. There are two options - a *fixed time step method* and the *method of variable time step*.

In the first case, at fixed intervals of time, changes are monitored. In the variable time step method, the time step bounds are exactly those times at which the system changes such as a new requirement arrives or the required service is terminated and the requirement leaves the system.

In [5], the M/M/n/n/∞/FIFO system was implemented in MATLAB using simulation data from a supermarket.

It makes it possible to enter λ (*mean intensity of the input*), μ (*mean service intensity*), the number of service lines n (then it is checked if $\lambda/n\mu < 1$), and the number

of requirements. For these data, the probabilities $p_k(t)$ and the above-mentioned characteristics are computed.

To understand the behaviour of the system, the program also offers graphical output of simulations. In Fig. 2, four graphs are shown for a system with three lines, which show requirement arrival times, requirement service times, requirement waiting times for service and finishing times of services for these requirements.

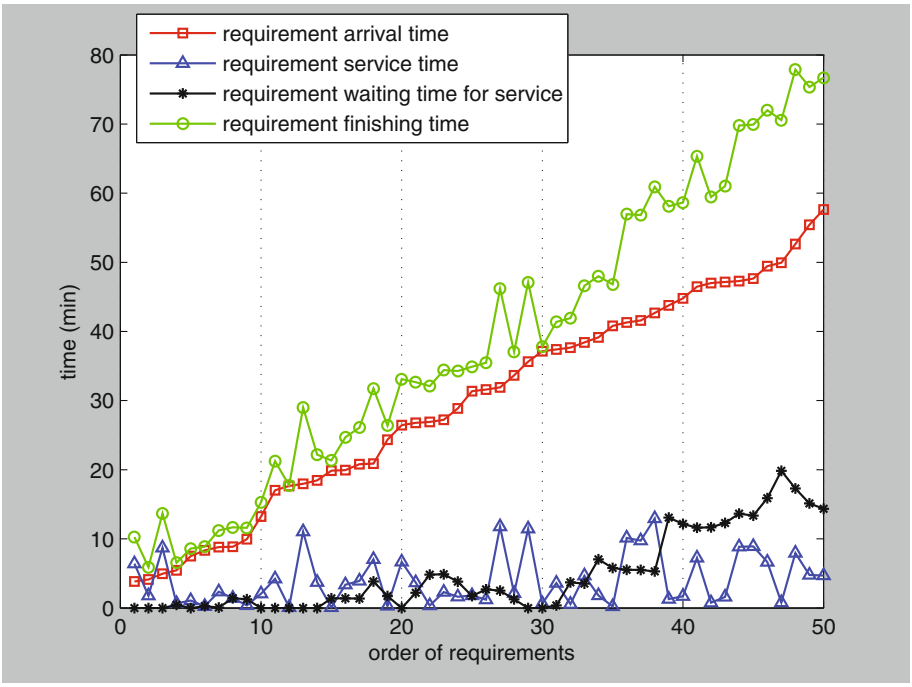

Fig. 2. Simulation of the M/M/n/n/∞ system for $\lambda = 45$, $\mu = 18$, $n = 3$, and 45 requirements.

Now we compare the analytical solution with the values obtained by simulation for different numbers of requirements. In the analytical part, we use formulas (22), (23) and (26), derived in Sect. 3, and the corresponding characteristics (27), etc. Simulations were run for 50, 200, and 500 requirements (clients). Table 1 sums up the results of the analytical formulas and simulations.

We can see that, if the number of requirements increases, then the difference between the analytical and the simulation results decreases. For 50 requirements, the difference is about 12%, but for 500 requirements, only 5%. Based on these achievements, we can conclude that the computer implementation of the simulation model reasonably approximates the M/M/n/n/∞/FIFO system.

Table 1. A comparison of analytical and simulation results

	Analytical evaluation mean values	Simulation results number of requirements		
		50	200	500
$E(T_s)$–$E(T_f)$	3.33333	3.64353	3.5118	3.45478
$E(T_f)$	4.68165	3.06941	4.17567	5.21177
$E(T_s)$	8.01498	6.71293	7.68747	8.66655
$E(N_s)$	6.01124	5.09125	5.62803	6.23332
$E(N_f)$	3.51124	2.30994	3.05703	3.74851
$E(N_s)$–$E(T_f)$	2.5	2.78131	2.57101	2.48481

5 Conclusions

This paper describes an approach to modelling of queuing system based on the use of Markov processes and, for a $M/M/n/n/\infty$/FIFO system, derives its characteristics in detail.

As some of the assumptions of the theoretical derivations may not be satisfied, such as the stationarity and the independence of increases, the simulation approach was also presented and implemented in MATLAB. This makes it possible to get all the statistics for any time intervals of real processes given the number of service lines, the times of requirement arrivals, and their service times. However, the approximation of theoretical model by implemented model, which computes with real data, depends on the number of requirements (and, therefore, on the time interval length). The higher the number of requirements, the better precision the simulation model has.

References

1. Bolch, G., Greiner, S., Meer, H., Trivedi, K.S.: Queueing Networks and Markov Chains. Wiley, New York (2006)
2. Bose, S.K.: An Introduction to Queueing Systems. Springer-Verlag, Berlin (2001)
3. Cooper, R.B.: Introduction to Queueing Theory. North Holland, New York (1981)
4. Gross, D., Shortle, J.F., Thompson, J.M., Harris, C.M.: Fundamentals of Queueing Theory. Wiley, New York (2008)
5. Horký, M.: Models of Queueing Systems. Diploma Thesis, Brno University of Technology, Brno (2015)
6. Hrubina, K., Jadlovská, A., Hrehová, S.: Optimisation Algorithms Using Programme Systems. Lecture Notes. Technical University in Košice, Prešov-Košice (2005)
7. Šeda, M.: Models of queueing systems. Acta Logistica Moravica **1**, 16–33 (2011)
8. Virtamo, J.: Queueing Theory. Lecture Notes. Helsinki University of Technology, Espoo (2005)
9. Willig, A.: A Short Introduction to Queueing Theory. Lecture Notes. Technical University, Berlin (1999). 42 pp.

http://www.springer.com/978-3-319-58087-6

Recent Advances in Soft Computing
Proceedings of the 22nd International Conference on Soft Computing (MENDEL 2016) held in Brno, Czech Republic, at June 8-10, 2016
Matousek, R. (Ed.)
2017, XI, 278 p. 90 illus., Softcover
ISBN: 978-3-319-58087-6