The International Journal of
TRANSPORT &*LOGISTICS*
Medzinárodný časopis
DOPRAVA A LOGISTIKA

ISSN 1451-107X

ERROR REDUCTION IN SIMULATION OF TRANSIENT BEHAVIOR OF QUEUEING SYSTEMS UNDER CRITICAL TRAFFIC CONDITIONS

Nebojsa Nikolic[1], Momcilo Milinovic[2], Olivera Jeremic[2], Radomir Jankovic[3]

[1]*Strategic Research Institute, MoD, Neznanog junaka 38, 11000 Belgrade,+381 63 8771756, nidzan@ptt.rs*

[2]*Faculty of Mechanical Engineering, Belgrade University, Kraljice Marije 16, Belgrade, Serbia, mmilinovic@mas.bg.ac.rs*

[3]*School of Computing, Union University, Knez Mihailova 6/VI, 11000 Belgrade, Serbia, rjankovic@raf.edu.rs*

***Abstract*:** *The paper presents capacity of Monte Carlo methodology to produce simulation results with respective level of accuracy and controllability. Critical traffic conditions assume saturation and overloading. To study them, we have to obtain the transient solutions for queuing system behaviour. Theoretical closed forms transient solutions are extremely complex and are subject of various attempts towards simplified approximations. Meanwhile queuing simulations suffer from problematic accuracy of simulation results. While theory gives absolute accuracy, simulation introduces some level of errors, and the question is how to control those errors. Practical interest for studying transient behaviour of queuing systems appears in some real applications, as in military missions, logistics, air traffic, etc.*

Key words: *Simulation, Queuing, Saturation, Overloading, Accuracy*

1 INTRODUCTION

Motivation for this paper comes from the need to investigate performances of one specific logistics process in the military branch: field maintenance of heavy equipment (trucks, tanks, personnel carriers, artillery items, etc.) in the conditions of high rates of demands for maintenance. Main goal of the field maintenance in military units is to sustain some level of the unit operational readiness (percentage of ready-to-use equipment). Field maintenance process in the brigade-size units (approximately: few hundreds vehicles, few thousands troops), could be modelled as multiphase, multi-channel queuing system with general types of distribution functions presenting stochastic nature of demands for maintenance actions, as well as for servicing itself. While maintenance resources are always limited, maintenance demands could grow up to very high values. All this makes corresponding queuing model to be very complex. It is hard to obtain analytical, closed-form solutions for such queuing models. Monte Carlo simulation supports modelling and analysis of complex queuing models, but also brings certain level of errors.

The main idea of this paper is to use theoretical solution for simple queuing model as kind of a benchmark for testing and calibration of the corresponding simulation model. However, the real goal is to test and verify simulation methodology, in order to apply it on more complex queuing models which for there are no theoretical solutions.

Queuing systems, or waiting line systems, are generalized concept which comprises client and server entities and theirs relations and rules. A client (customer or service demand) arrives to the service channel on a random basis, and his servicing in the server entity is stochastic as well. Due to its generality, queuing concept is applicable on many real systems and processes in various areas. Queuing theory deals with queuing systems. Theoretical support for regular (non-critical) traffic conditions is very well. However, investigation of queuing behaviour becomes very difficult for critical traffic conditions because it assumes need for transient solutions for queuing behaviour. Monte Carlo simulation is recognized as an effective method to overcome theoretical complexities.

Traffic intensity (ρ) in a queuing system is expressed as relation ($\rho=\lambda/\mu$) between intensity of clients arrivals (λ) and intensity of servicing in a service channel (μ). There can be three cases of traffic intensity: normal traffic ($\rho<1$); saturation ($\rho=1$); and overloading ($\rho>1$). Under "Critical traffic conditions", (a watchword proposed by Brandao and Porta Nova [1]), we assume cases of saturation and overloading. Intuitively, it is clear that after enough time there will be enormous queues in the cases of critical traffic conditions. The question is how long queuing system can operate holding reasonable queue length and waiting time. Finite operational time (the time when system is opened for clients) of some queuing system gives a hope that critical traffic conditions could be survived.

To investigate behaviour of queuing system during finite operational time, we need so called transient solutions. Research interest for getting insight in the transient behaviour of queuing systems appears in different areas. Applied probability community tries to find a way for such closed form transient solutions which are appropriate for practical calculations. This was a challenging task for a long period, but still actual. One representative older effort is given by Conolly and Langaris [2], and some of the newer papers come from Hlynka, Hurajt and Cylwa [3] and Leonenko [4].

A novel impulse appeared at the Winter Simulation Conference 2010, where Kaczynski, Leemis and Drew [5], clearly confirmed existence and importance of the transient problem in queuing systems behaviour in the military branch. They suggested a use of both approaches: simulation modelling and probability theory, but preferred the last. Some earlier studies in the field of air traffic (in general, not only in the military) realized by Peterson, Bertsimas and Odoni [6], recognized the importance of transient behaviour of queuing systems.

In the military branch, high tempo and short but intensive and unpredictable dynamics of events in contemporary battlefield, raises specific issues in the simulation modelling not only for logistics processes, but also for other aspects for military engagements. In constellation with high precision weapons and new combat concepts like swarming (presented by Jankovic [7]), a request arises for investigation of conflicts of short durations and high intensity.

Here, we present a comparison of simulation results versus theoretical ones for selected time-dependent state probability for M/M/1 queuing system. We run the model for two different traffic intensities ($\rho=\lambda/\mu$): saturation ($\rho=1$); and overloading ($\rho>1$). The goal is to check potential of Monte Carlo simulation method for generating time-dependent state probabilities with high and controlled accuracy. The wider context of this effort is to get confidence on specific Monte Carlo method verified in a case of a simple model, for future applications in more complex models. Problem of accuracy of simulation results is well-known and still ongoing research issue in the simulation field.

Transient regime, as operational mode of queuing systems behaviour, has been in the shadow of the steady-state behaviour of queuing systems for a long time. Research interest has been dominantly oriented towards stationary regime. However, transient and stationary regimes are complementary. Transient phenomenon indicates behaviour of queuing system in

a period which precedes the steady-state. Transient behaviour appears at the beginning of working cycle of queuing system. And it appears in spite of constant intensities of arrival stream (λ) and service rate (μ).

Practical consequences of transient regime are different values of measures of performances from their steady-state values. This is particularly important in situations when time interval, characterized by transient regime, is a respective part of the whole period of engagement (working time) of queuing systems (QS).

Perception of the transient problem in the Monte Carlo simulation of queuing systems is quite different than it is in a pure theoretical approach. Simulated queuing system cannot jump into its steady-state, as it is easy in the theoretical approach. For example, a state's equations for the M/M/n queuing model are the first order differential equations. And, with a stroke of the pen you can let the argument (time) to tend to the infinity. Doing so, you skip the initial transient period immediately, and reach the steady-state, while state's equations become algebraic instead of differential ones. Simulated queuing model, on the other hand, really travel through its transient regime.

2 SYSTEM STATES PROBABILITIES

States probabilities are queuing system primary measures of performances. Mathematical model of a queuing system of type M/M/1/∞ is given by a system of differential equations (1). Every possible state of queuing system is presented with one differential equation (Erlang's equations, or Kolmogorov-Chapman equations). Those are the following differential equations of first order (1):

$$\begin{aligned}
&\dot{p}_0(t) = -\lambda p_0(t) + \mu p_1(t) \\
&\dot{p}_1(t) = \lambda p_0(t) - (\lambda + \mu) p_1(t) + \mu p_2(t) \\
&\qquad \cdots \\
&\dot{p}_{n-1}(t) = \lambda p_{n-2}(t) - (\lambda + \mu) p_{n-1}(t) + \mu p_n(t) \\
&\qquad \cdots
\end{aligned} \tag{1}$$

Also, there are: the normalization condition (2), and initial conditions (3). That means queuing system will be certainly in one of the possible states in any time moment.

$$\sum_{i=0}^{n} p_i(t) = 1 \tag{2}$$

$$p_0(0) = 1,\ p_1(0) = p_2(0) = \ldots = p_n(0) = 0 \tag{3}$$

Variables $\boldsymbol{p_i(t)}$ present time-dependent probabilities of the queuing system's states. Index $\boldsymbol{i}$ presents the number of clients in a system. The independent variable is time ($\boldsymbol{t}$). Intensity of input client's stream is $\boldsymbol{\lambda}$. Intensity of output client's stream (servicing) is $\boldsymbol{\mu}$.

Complete solution of above system of differential equations assumes obtaining states probabilities as time-dependent variables. This solution (usually termed as the transient solution) actually exists and could be found in many queuing theory books, while Kleinrock [8] gives some interesting comments about it. In their essence, transient solutions are complete solutions which are valid for any traffic intensity, and for both regimes: initial (transient, warm-up, start-up, relaxation) and steady-state (equilibrium). Transient solutions assume time-dependent variables. Division of queuing system behaviour on a transient and steady-state regime is artificial in some sense. Practically, there is no clear and definite

"switching moment" from the transient to the steady-state regime, at least for the models with unlimited queues. Instead, that is a continual change across time, more or less long.

The problem with transient solution lies in the complexity of that solution particularly when someone tries to use it in practical calculations. Because of that, transient solutions are subject of research aimed towards finding more simplified approximation which will be appropriate for practical calculations. Instead of pure but complex theoretical approach, we can use numerical methods to solve a system of differential equations. However, numerical methods approach becomes cumbersome in case of queuing systems with many possible states of queuing system. Besides, this is not the only problem with application of numerical methods. In case of other types of queuing systems (non-exponential distributions, queuing networks, etc.) it is even difficult to establish system of differential equations. In short, regardless do we have, or, we do not have system of differential equations for analytical description of behaviour of queuing system, we want to get solutions: time-dependent probabilities of possible states of the queuing system under study. Monte Carlo simulation methodology has capacity to produce time-dependent solutions under such limitations.

3 SIMULATING STATES PROBABILITIES

Complexity of pure analytical approach or numerical methods application to this task could be avoided, by use of Monte Carlo simulation modelling methodology. Nikolic proposed a concrete simulation method for simulating states probabilities as time-dependent variables [9]. That method ("Automated Independent Replications with Gathering Statistics of Stochastic Processes", shorten as: AIRGSSP) is used here. Practically, we can get numerical solutions for time-dependent states probabilities by use of Monte Carlo simulation modelling, and without dealing with the system of differential equations itself. This methodological capacity could be termed as "Statistical integration of differential equations", or, which is already known in a literature, as "Monte Carlo integration".

The goal is to make numerous and independent simulation experiments (designated as IR – Independent Replications) and to collect statistics of dynamic variables under study (states probabilities). Functional connection among variables of interest is given by formula (4), which comes from the basics of the probability theory and mathematical statistics. Accuracy of simulation results we perceived through complementary term: maximal error of estimation in percents - $\boldsymbol{\varepsilon}$; Confidence level on simulation results (confidence coefficient for Normal distribution - $\boldsymbol{Z_c}$); Number of IR of simulation experiment ($\boldsymbol{n}$, sample size); Probability (proportion - $\boldsymbol{p}$) is the entity under study; Complementary variable of the probability under study is: $\boldsymbol{q = 1-p}$.

$$n = \frac{q}{p}\left(\frac{100}{\varepsilon}\right)^2 \boldsymbol{Z}_c^2 \tag{4}$$

Depending on a desired accuracy, or confidence level, or the order of magnitude of the state probability under study, we can choose various numbers of IR of simulation experiment. For example, one thousand IR of simulation run, permit a maximum 24,5 % discrepancy of estimation, for probability level of 0,1, and with level of statistical confidence at 0,99: $\boldsymbol{Z_c}$ (0,99) = 2,58.

In our example we did next steps (detailed description is given by Nikolic, [9]):

- Take some fixed number of IR.
- Make simulation runs with chosen number of IR.
- Chose high and fixed level of confidence.
- Calculate maximally permitted error of estimation for a given value of states probability.

- Obtain theoretical values for time-dependent state probabilities.
- Compare and analyze correspond values of simulated and theoretical time-dependent state probability.
- Make conclusion based on comparison results: does or does not simulation methodological approach is able to produce output with declared accuracy and confidence.

4 MODEL EXAMPLE

A single channel queuing system of a type M/M/1 with infinite queue has been modelled. The system operates for some finite time. Initial condition is: "queuing system is empty" (that is: no clients in queue, no clients in service channel). After reaching the end of the operational time, system closes. The goal is to obtain time-dependent response of the state $p_0(t)$: "system is empty". The principle is the same for all other states probabilities. As a referential theoretical values for state probability $p_0(t)$, we used results calculated by Conolly and Langaris, which are presented in their paper [2].

Described conceptual queuing model has been developed further as a simulation model and prepared for Monte Carlo simulation on a personal computer. A simulation model development is crucial step and general methodology is well known. We used here general simulation methodology well described by Law and Kelton in their famous book [10]. Being applied in different branches, general simulation approaches have been further developed and accommodated, according to specific needs of every branch of application. In example, Malindzak et al [11], proposed a systematic procedure for simulation modelling of large scale logistics systems in an specific real application (mining and metallurgy manufacturing).

The idea was to repeat execution of the model, that is, to make numerous IR of the simulation run. The purpose of numerous IR is to collect many data in order to obtain good statistical sample for estimation of desired measures of performances. In this task we make three sets of experiments for three different numbers of IR: 1.000; 10.000; and 100.000 IR. Doing so, we got selected state probability, $p_0(t)$, as time-dependent variable.

Computational time on a typical PC (2,2GHz, 2RAM) varied due to the experimental conditions: it takes few minutes for examples with 1.000 IR; about 15 minutes for 10.000 IR; and about two hours for 100.000 IR. For computer implementation of simulation model we used student version of GPSS simulation language [12].

5 RESULTS AND ANALYSIS

After six simulation experiments, we got simulation results. Graphical presentations of time-dependent state probability $p_0(t)$ is given in Figure 1. All three cases of different number of IR (1.000; 10.000; 100.000) were executed for both cases of traffic intensity (saturation and overloading). Then, theoretical values for state probability $p_0(t)$ were associated for both cases of traffic intensities. All this is presented in Figure 1.

On the basis of three classes of simulation results (in a table form), and corresponding theoretical results for $p_0(t)$, for a set of six time points (0; $\mathbf{T}_\mu$; $2\mathbf{T}_\mu$;…$5\mathbf{T}_\mu$), we calculated percentages for realized discrepancy of the simulation results versus theoretical results. Permitted discrepancies are calculated from formula (4) using: theoretical values for $p_0(t)$ and its corresponding counterpart q(t), a given number of IR, and a level of confidence at "3σ". Realized discrepancies for two traffic intensities are as follows (Table 1 and Table 2).

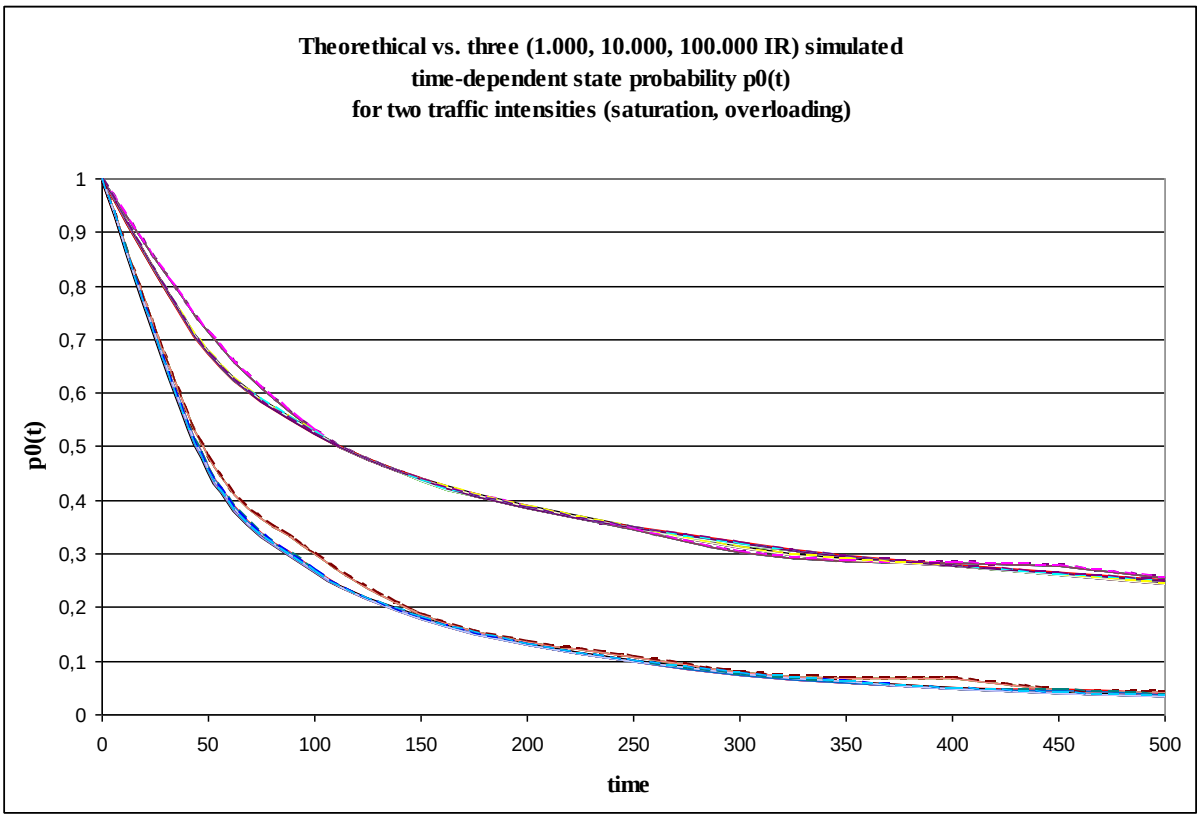

Figure 1: Time-dependent state probability $p_0(t)$ for two different traffic intensities

(1) Saturation, Table 1, is a case when intensity of input client's stream (a stream of demands for service) is equal to intensity of output client's stream (nominal capacity of service channel to process demands). Realized and permitted discrepancies, Table 1, are in good agreement, so, declared accuracy is achieved at a chosen level of confidence in the case of saturation.

Table 1: Simulated vs. theoretical $p_0(t)$, with percentage error, case of the saturation

Queuing model: M/M/1/∞ Average service time: $T_\mu=1/\mu = 100$ Average inter-arrival time: $T_\lambda=1/\lambda = 100$										
Time	Time-dependent state probability - $p_0(t)$				Error - ε (% of the theoretical value)					
t (in $\mathbf{T_\mu}$, as a relative t. unit)	Simulation results for 3 experiments with different numbers of IR			Theoretical results (Conoly & Langaris, 1993)	Realized error (%)			Permitted error for "3 Sigma" (%)		
	10^3 IR	10^4 IR	10^5 IR		10^3 IR	10^4 IR	10^5 IR	10^3 IR	10^4 IR	10^5 IR
0 T_μ	1	1	1	1	0,0	0,0	0,0	0,0	0,0	0,0
1 T_μ	0,531	0,5255	0,5259	0,523778	1,4	0,3	0,4	9,0	2,9	0,9
2 T_μ	0,388	0,3869	0,3854	0,385753	0,6	0,3	0,1	12,0	3,8	1,2
3 T_μ	0,303	0,3124	0,3174	0,318709	4,9	2,0	0,4	13,9	4,4	1,4
4 T_μ	0,284	0,2768	0,2771	0,277574	2,3	0,3	0,2	15,3	4,8	1,5
5 T_μ	0,255	0,2446	0,2497	0,249096	2,4	1,8	0,2	16,5	5,2	1,6

(2) Overloading, Table 2, is a case when intensity of input client streams (λ) is greater then intensity of servicing (μ). Then, traffic intensity (ρ) is greater than 1. In our experiments traffic intensity has a value 2. As time passes the queue becomes longer, and service channel should be continually engaged (sized). That means, probability of the state $p_0(t)$: "system is

empty of clients", should be zero. The question is when it happens, and how fast this probability approach to zero? To answer the question we can look (Figure 1) at the curve presenting time-dependent probability $p_0(t)$: there are obvious a smaller values for the case of overloading comparing it with the other (saturation). Realized and permitted discrepancies are in good agreement as it could be noticed from the Table 2 and perceived from Figure 1.

Table 2: Simulated vs. theoretical $p_0(t)$, with percentage error, case of the overloading

Queuing model: M/M/1/∞ Average service time: $T_\mu=1/\mu = 100$ Average inter-arrival time: $T_\lambda=1/\lambda = 50$										
Time	Time-dependent state probability - $p_0(t)$				Error - ε (% of theoretical values)					
t (in T_μ, as a relative time units)	Simulation results for three experiments with different numbers of IR			Theoretical results (Conoly & Langaris, 1993)	Realized error (%)			Permitted error for "3 Sigma" (%)		
	10^3 IR	10^4 IR	10^5 IR		10^3 IR	10^4 IR	10^5 IR	10^3 IR	10^4 IR	10^5 IR
0 T_μ	1	1	1	1	0,0	0,0	0,0	0,0	0,0	0,0
1 T_μ	0,299	0,2694	0,26922	0,2676	11,7	0,7	0,6	15,7	5,0	1,6
2 T_μ	0,137	0,1316	0,13051	0,1303	5,1	1,0	0,2	24,5	7,8	2,5
3 T_μ	0,08	0,0743	0,07676	0,0764	4,7	2,7	0,5	33,0	10,4	3,3
4 T_μ	0,067	0,0483	0,04916	0,0489	37,1	1,2	0,6	41,9	13,2	4,2
5 T_μ	0,043	0,0377	0,03365	0,0329	30,7	14,6	2,3	51,4	16,3	5,1

6 CONCLUSIONS

Simulation results show good concordance with exact theoretical results. Simulation errors decrease with increase of the number of independent replications of simulation experiment. Proposed functional relation among relevant measures for error control of simulation results works satisfactorily. Agreement is evident for both traffic intensities: saturation and overloading. This contributes to the robustness of simulation approach.

Results for cases of saturation and overloading are particularly interesting for queuing systems which operate for some finite portion of time. Such systems simply do not have time enough to reach steady-state because their mission ends before their steady-state happens. Results obtained from simulation can and should be used in or for the real system or process which from we actually started simulation endeavour. It is a kind of a circle, as it is presented by Malindzak [13]. According to that, simulation results in this example support one relaxing conclusion, in the sense that queues and waiting times will not explode if queuing system is exposed to the "critical traffic conditions". However, this conclusion stands only for some limited time period.

Field maintenance was, and still is important logistics process in military units in various armies and it seems to be the same in the future. First-hand experience in the field maintenance and military logistics as a whole, presented by Tilzey, Kasavicha, and Rote [14], confirms this conclusion. Besides other logistics aspect, they emphasized importance of maintenance and particularly capabilities on the field for recovery and evacuation of heavy military equipment (heavy, armoured vehicles).

Future research could be oriented toward investigation of other measures of performances. Also, it is worth to check practical capacity of theoretical approach to be applied for larger set of possible states of queuing system and longer duration of operational time. In regard to a kind of a real system, future research will be directed to the simulation of more complex models.

ACKNOWLEDGEMENT

This work is supported in part by the Ministry of Science and Technical Development of the Republic of Serbia under interdisciplinary Project No. III-47029.

References

[1] Brandao R.M, Porta Nova A.M: Non-stationary queue simulation analysis using time series. Proceedings Winter Simulation Conference 2003. Editors: S. Chick, P. J. Sanchez, D. Ferrin, and D. J. Morrice. 408-413.
[2] Conolly B.W., Langaris C.: On a New Formula for the Transient State Probabilities for M/M/1 Queues and Computational Implications. Journal of Applied Probability. 1993. 30(1). 237-246.
[3] Hlynka M., Hurajt L.M., Cylwa M.: Transient results for M/M/1/c queues via path counting. Int. J. Mathematics in Operational Research. 2009. 1(1/2). 20-36.
[4] Leonenko G.M.: A new formula for the transient solution of the Erlang queueing model. Statistics and Probability Letters. 2009. 79. 400-406.
[5] Kaczynski W., Leemis L., Drew J.: Modeling and Analyzing Transient Military Air Traffic Control. Proceedings Winter Simulation Conference 2010. Editors: B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, E. Yucesan. 1395-1406.
[6] Peterson M. D., Bertsimas D.J., Odoni A.R.: Decomposition algorithms for analyzing transient phenomena in multi-class queueing networks in air transportation. Operations Research. 1995. 43(6). 995-1011.
[7] Jankovic R.: Computer simulation of an armoured battalion swarming. Defence Science Journal. 2011. 61(1). 36-43.
[8] Kleinrock L.: Queueing systems – Volume I: Theory. New York: John Wiley & Sons. 1975. 416p. ISBN 0-471-49110-1.
[9] Nikolic N.: Statistical integration of Erlang's equations. European Journal of Operational Research. 2008. 187(3). 1487-1493.
[10] Law A., Kelton D., Simulation modeling and analysis. New York: McGraw Hill, 1982. 400 p. ISBN 0070366969
[11] Malindžak D., Straka M., Helo P., Takala J.: The methodology for the logistics system simulation model design. Metalurgija. 2010. 49(4). 348-352.
[12] Radenkovic B, Stanojevic M, Markovic A.: Racunarska simulacija. Beograd. FON (in Serbian, 4th edition). 2009. 313p. ISBN 978-86-7395-102-7.
[13] Malindžak D.: Modely a simulacia v logistike. Acta Montanistica Slovaca. Ročnik 15(2010), mimoriadne číslo 1, 1-3.
[14] Tilzey F.D., Kasavicha G., Rote X.C.: Stryker Brigade Combat Teams Need Forward Support Companies. Army Logistician. July-August 2008. 40(4). 26-32. ISSN 0004-2528. [cit. 2012-03-03] <http://www.almc.army.mil/alog/ issues/JulAug08/pdf/alog_jul_aug08.pdf >