MATHEMATICA APPLICANDA
Vol. 40(2) 2012, p. 1–12
doi: 10.14708/ma.v40i2.358

Vladimir V. Mazalov (Petrozavodsk)
Andrei Gurtov (Espoo)

Queuing System with On-Demand Number of Servers

Abstract We consider a queuing system where the number of active servers changes depending on the length of the queue. As a practical example of such a system, we consider the security check queue at airports. The number of active servers increases when the queue grows by k customers and decreases accordingly. This allows to save server resources while maintaining acceptable performance (average queuing time and its variation) for customers. We obtain a closed-form solution for the serving time, queue length and average number of servers.

To validate the model we selected data from Dallas – Fort Worth International Airport, the eighth largest in the world in terms of passenger traffic. Our simulation model shows a close match with analytic results. Cost savings in the number of open servers are achievable while providing acceptable waiting time for the customers.

2010 Mathematics Subject Classification: 68M20; 93A30; 60K25.

Key words and phrases: Capacity planning, queuing theory, dynamic queue, airport security check.

1. Introduction

Traditionally, queuing theory considers models with a fixed number of servers [10, 15, 19]. The main performance metrics there are the queue length and waiting time. However, it makes sense to consider queuing systems with a changing number of servers depending on the queue length [4]. In this paper, we propose a new queuing discipline which allows to save server resources by deploying new servers only when needed. Having only the minimum number of servers sufficient to provide the required performance with high probability saves costs. To the best of our knowledge, this problem has not been considered yet in such a setting. Such queuing systems could be readily used for example at airports, where a passenger should have a known upper bound for the waiting time at a security check [24].

Related work can be divided to two groups. In the first group, servers have different service rates and researchers aim at allocating the incoming customers to optimize the load of the system [1,18]. In this case, some customers can move from a long queue to a shorter one [17]. In the paper [23] the author proposes that the optimal number of servers is of the form $\lambda+\gamma\sqrt{\lambda}$ depending on the total arrival rate λ for a given grade of service γ. For optimal control it is suitable to use multi-threshold strategies; when the queue to a given server exceeds a certain threshold, customers move to other servers [5,16,22] or leave the system [2].

In the second group of works, the servers are identical and the researchers aim to distribute customers among servers, which can become active or inactive [8,12,21], [11].

The closest to this paper is work by Solovyev [20] that considers the birth-death process with a system of parameters $\{\lambda_n\}$ – arrival intensity and $\{\mu_n\}$ the service rate. The arrival rate is fixed, but the service rate can be set arbitrarily. The goal is to find such service rates μ_n given the arrival rate, that the system costs, consisting of waiting time of all the customers in the queue and the average cost of the server's work, are minimal. The model that we are considering differs in the minimization criteria and service discipline.

Other relevant work to this paper [3,7] is where multi-level strategies are used for controlling queues in retrial models, and also work [13] where the number of active servers is determined periodically depending on the state of the system in the previous time interval.

Consider a queuing system with one server. The queue grows and at some point reaches k customers. If one more customer arrives, we deploy a second server. If the queue exceeds $2k$ at some point, we deploy a third server and so on. A practical example of such a system is security checks at airports, where the waiting time for customers cannot exceed a certain limit to avoid being late for the departing flights. Later in the paper, we validate our queuing model using the parameters from a report of Dallas – Fort Worth International Airport [9].

Such a queuing system is illustrated in Figure 1. All servers, the security checkpoints in our case, share the same queue and are shown as circles. The queue is organized in a series of lines, separated by barriers on each side. We can assume that each line can hold k customers, and a new server becomes active if a new line starts getting filled. In Figure 1, we see three active servers and one inactive, because three lines in the queue have customers. Such arrangements are often found at airport security, fast food restaurants, ticket vendors, and many other real life scenarios. The management goal is to minimize the number of active servers while providing an acceptable level of service depending on the current queue length. Due to the real-life importance of this scenario, any novel models studying such queues from new angles are of high interest.

The rest of the paper is organized as follows. In Section 2, we formulate the queuing system with a variable number of servers. In Section 3, we derive the main characteristics of the system, such as queuing time, analytically. In Section 4, we present simulation results. Finally, Section 5 presents conclusions and plans for future work.

2. Queuing model with a variable number of servers

Consider a classic queuing system model $M/M/N$ with incoming Poisson arrivals with intensity λ, exponential serving time μ and number of servers N. Assume that N is a random variable that can take natural values. It is fully defined by the following serving system, which we call the serving protocol.

Assume that the queue at each server cannot exceed k customers. This is a parameter in our protocol which is now fixed, but will be varied later. Denote by $x(t)$ the common queue length at time t. Note that if at moment t the following condition holds: $k(n-1) < x(t) \leqslant kn$ for some natural n, the system has n active servers. Denote by $P_i(t)$, $i = 0, 1, \ldots$ the probability that at time t the common queue has i customers. For these probabilities we can write down a system of Kolmogorov equations.

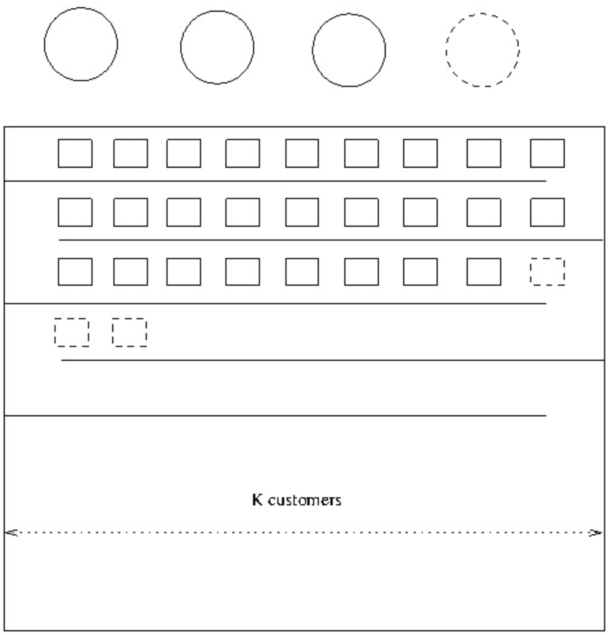

Figure 1: The airport queuing system with on-demand number of servers.

$$P'_{kn}(t) = -(\lambda + n\mu)P_{kn}(t) + \lambda P_{kn-1}(t) + (n+1)\mu P_{kn+1}(t)$$
$$P'_{kn+j}(t) = -(\lambda + (n+1)\mu)P_{kn+j}(t) + \lambda P_{kn+j-1}(t) + (n+1)\mu P_{kn+j+1}(t),$$

with $P_{-1}(t) = 0$ for $n = 0, 1, \ldots, j = 1, ..., k-1$.

This system of equations is a special case of a general system of equations for a birth-death process

$$P_l'(t) = \lambda_{l-1} P_{l-1}(t) - (\lambda_l + \mu_l) P_l(t) + \mu_{l+1} P_{l+1}(t),$$

for $l = 0, 1, \ldots$, where $\lambda_{-1} = \mu_0 = 0$, $\lambda_l = \lambda$, $l = 1, 2, \ldots$, $\mu_{kn+j} = (n+1)\mu$, for $n = 0, 1, \ldots, j = 1, \ldots, k-1$ and $\mu_{kn} = n\mu$ for $n = 1, 2, \ldots$.

It is known [14], that such a process has a steady state if and only if the following holds

$$\sum_{l=0}^{\infty} \theta_l < \infty, \quad \sum_{l=0}^{\infty} \frac{1}{\lambda_l \theta_l} = \infty, \tag{1}$$

where

$$\theta_0 = 1, \quad \theta_l = \frac{\lambda_0 \ldots \lambda_{l-1}}{\mu_1 \ldots \mu_l}, \quad l \geqslant 1.$$

In our case, the values θ_l are of the form

$$\theta_{kn+i} = \frac{\rho^{kn+i}}{(n!)^k (n+1)^i}, \quad i = 1, \ldots, k, \; n = 0, \; \ldots,$$

where $\rho = \lambda/\mu$ is the system load. It is easy to see that the conditions (1) for the existence of steady-state hold, since the series

$$\sum_{n=0}^{\infty} \sum_{i=1}^{k} \theta_{kn+i} = \sum_{n=0}^{\infty} \left(\frac{\rho^n}{n!} \right)^k \sum_{i=1}^{k} \left(\frac{\rho}{n+1} \right)^i$$

converges, and

$$\frac{1}{\lambda} \sum_{n=0}^{\infty} \sum_{i=1}^{k} \frac{1}{\theta_{kn+i}} = \frac{1}{\lambda} \sum_{n=0}^{\infty} \left(\frac{n!}{\rho^n} \right)^k \sum_{i=1}^{k} \left(\frac{n+1}{\rho} \right)^i = \infty$$

since, according to Stirling's formula $n! = (n/e)^n \sqrt{2\pi n}(1 + O(1/n))$. We obtain a system of equations for the steady state ($P_i(t) = P_i = \text{const}$), which contains several **groups** of similar recurrent equations.

The first group of k equations:

$$P_1 = \rho P_0, \quad P_{i+1} = (1+\rho) P_i - \rho P_{i-1}, \quad i = 1, \; \ldots, \; k-1.$$

The second group of equations:

$$P_{k+1} = \frac{1+\rho}{2} P_k - \frac{\rho}{2} P_{k-1}, \; P_{k+i+1} = \left(1 + \frac{\rho}{2}\right) P_{k+i} - \frac{\rho}{2} P_{k+i-1}, \; i = 1, \; \ldots, \; k-1.$$

The general formula for P_i:

$$P_{kn+i} = \frac{\rho^{kn+i}}{(n!)^k (n+1)^i} P_0, \; i = 1, \; \ldots, \; k, \quad n = 0, \; 1, \; \ldots. \tag{2}$$

We find P_0 from the condition

$$\sum_{i=0}^{\infty} P_i = 1.$$

From the general formula (2) we obtain the equality

$$P_0 + \sum_{n=0}^{\infty} \sum_{i=1}^{k} P_{kn+i} = 1$$

or

$$P_0 \left(1 + \sum_{n=0}^{\infty} \sum_{i=1}^{k} \frac{\rho^{kn+i}}{(n!)^k (n+1)^i} \right) = 1.$$

Simplifying, we obtain

$$P_0 = \left(1 + \rho \sum_{j=1}^{\infty} \frac{\rho^{(j-1)k}}{(j!)^k} \frac{j^k - \rho^k}{j - \rho} \right)^{-1}.$$

Now we can calculate all the P_js and consequently all the main characteristics of the queuing system.

3. Main characteristics of the queuing system

An important parameter of the system is the average number of active servers

$$E\{N\} = \sum_{n=0}^{\infty} \sum_{i=1}^{k} (n+1) P_{kn+i}.$$

Simplifying, we obtain

$$E\{N\} = \sum_{n=0}^{\infty} (n+1) \sum_{i=1}^{k} P_{kn+i} = \rho \sum_{n=0}^{\infty} \sum_{i=0}^{k-1} P_{kn+i} =$$
$$= \rho \left(1 - P_0 - \sum_{n=0}^{\infty} P_{k(n+1)} + \sum_{n=0}^{\infty} P_{kn} \right) = \rho,$$

i.e. the average number of active servers for any k will be equal to the system load $E\{N\} = \rho$. That does not depend on the maximal queue length of any server.

However, the variance of N depends on k. First we calculate

$$E\{N^2\} = \sum_{n=0}^{\infty} (n+1)^2 \sum_{i=1}^{k} P_{kn+i} =$$
$$= \sum_{n=0}^{\infty} \frac{\rho^{kn+1}(n+1)}{(n!)^k} P_0 + \rho^2 \sum_{n=0}^{\infty} \sum_{i=0}^{k-2} P_{kn+i} =$$
$$= \sum_{n=0}^{\infty} \frac{\rho^{kn+1}(n+1)}{(n!)^k} P_0 + \rho^2 \left(1 - \sum_{n=0}^{\infty} P_{k(n+1)-1} \right).$$

Simplifying, we obtain

$$E\{N^2\} = \rho^2 + P_0 \sum_{n=0}^{\infty} \frac{\rho^{kn+1}}{(n!)^k}.$$

Hence, the variance of the number of active servers is

$$\mathrm{Var}\{N\} = E\{N^2\} - (E\{N\})^2 = P_0\rho \sum_{n=0}^{\infty} \left(\frac{\rho^n}{n!}\right)^k.$$

Hence, the variance is a decreasing function of k. It is an important characteristic in practice, since it shows the variance of the necessary number of servers around the mean value. If this variance is large, the system might need many additional servers at some point of time for serving customers.

In Figure 2, the solid line shows the standard deviation $\sigma(k)$ (square root of variance) of the number of active servers for different values of parameter k. To aid the reliable functioning of the system, we can construct confidence intervals $[0, M]$ for the number of active servers with high significance δ such that

$$P\{N \leqslant M\} \geqslant 1 - \delta.$$

Then the proportion of time for which at most M servers are needed is at least $1 - \delta$.

In Figure 2, the black circles show the value of M such that at most M servers are required at least 99 % of the time. We see a good match of the graphs of this level and the standard deviation. To choose an optimal value k, we have to compute other important characteristics of the queuing system, such as the average queue length and more importantly average waiting time in the queue.

The average number of customers in the system will be less than $k\rho$.

$$E\{q\} = \sum_{n=0}^{\infty}\sum_{i=1}^{k}(kn+i)P_{kn+i} = k\rho - \sum_{n=0}^{\infty}\sum_{i=1}^{k}(k-i)P_{kn+i} \leqslant k\rho.$$

Let us find an upper bound on the mean waiting time. It is important that the number of active servers can change depending on the current queue length. We can definitely say that the worst case for a customer entering the queue at position $nk + i$ (i.e. when $n + 1$ servers are active) is when no customers arrive afterwards. Then, during the service the number of active servers will gradually decrease and the expected waiting time for that customer will be

$$\frac{i}{(n+1)\mu} + \frac{k}{n\mu} + \frac{k}{(n-1)\mu} + \ldots + \frac{k}{\mu} = \frac{i}{(n+1)\mu} + \frac{k}{\mu}\left(1 + \frac{1}{2} + \ldots + \frac{1}{n}\right).$$

Table 1: Optimal value of k^* for different criteria weights.

C	1	2	3	4	5	6	7	8	9	10	20	30	40	50	60	70	80	90	100
k^*	2	3	4	5	5	6	7	7	8	8	13	18	20	23	26	29	31	33	35

Hence, the following is an upper bound for the average waiting time in the queue

$$E\{t\} = \sum_{n=0}^{\infty}\sum_{i=1}^{k}\left(\frac{i}{(n+1)\mu} + \frac{k}{\mu}s_n\right)P_{kn+i},$$

where $s_n = 1 + \frac{1}{2} + ... + \frac{1}{n}$ is a harmonic series.

4. Simulations

To validate the model we have selected Dallas – Fort Worth International Airport. According to Wikipedia, it is the fourth busiest airport in the world in terms of aircraft movements; in terms of passenger traffic, it is the eighth busiest airport in the world. The choice of this particular airport is motivated by the availability of public data on its security checking process.

According to the airport authorities, it serves on average 54 passengers per minute. The security checks are performed by a maximum of 18 lines each with a peak throughput of 260 passengers per hour (for computer simulation we chose parameters $\lambda = 54$ and $\mu = 4.25$).

Figure 2 and Figure 3 show simulation results for the queuing system using the Mathematica software package. Figure 2 shows the necessary number of servers versus the parameter k (maximum queue length per server). Figure 3 shows the average waiting time in minutes versus the parameter k.

As we can see in the figures, if we fix the parameter k e.g. equal to 15 (the maximum queue per server) then for Dallas – Fort Worth International Airport the maximum number of active servers will be approximately 14. Then, the average waiting time in the queue for a passenger will be 11 minutes. That matches the real world data from the airport report. However, our results permit keeping 4 lines out of 18 closed while providing the necessary performance, therefore saving costs.

We can combine the penalty of the waiting time of passengers and cost of servers into one weighted criterion

$$H(k) = E\{t\} + C(E\{N\} + \sigma(k)) = E\{t\} + C(\rho + \sigma(k)),$$

where C is some weight. We can now look for the optimal value of k giving the minimum average weighted costs.

For Dallas – Fort Worth International Airport, we computed the values given in Table 1. We can see that if the costs of servers exceed the costs of customers waiting by an order of magnitude, then the parameter k should be made equal to 8; if by two orders of magnitude then 35.

5. Conclusion

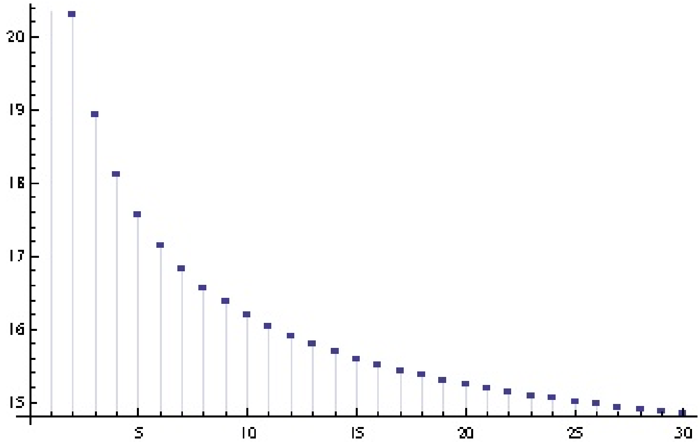

Figure 2: The necessary number of servers versus the parameter k (max queue length per server).

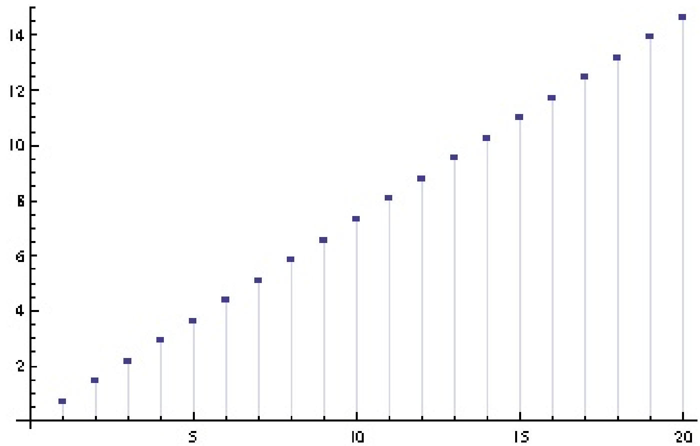

Figure 3: Average waiting time in minutes versus the parameter k (max queue length per server).

We proposed a queuing system where the number of active servers is dynamically adjusted based on the queue length. We analyzed the system using the Kolmogorov differential equations and obtained upper bounds on the average waiting time, queue length and number of active servers.

The proposed system can model the security check procedure at airports. Our simulation model shows a close match with analytic results using real-world data from Dallas – Fort Worth International Airport. Using our model permits costs savings by limiting the number of active servers (security checkpoints) while providing an acceptable waiting time for the customers with high probability.

A more generic threshold queuing discipline was considered by Solovjev [20] and in a more specific variant, close to the one considered in our paper, but for a more general arrival process (MAP) by Chakravarthy [6]. In this paper, we considered a particular queuing mechanism in use by airport security checkpoints. It allowed us to simplify the problem and formulate it using just one parameter k, the queue segment length. This allowed us to find a solution analytically in closed form that facilitates the application of results in practice.

In future work, we plan to consider a queuing system in supermarkets, where each server has their own queue. A cashier closes when their queue becomes empty. When the queue to any cashier reaches k, a new cashier opens, which attracts customers from different queues.

Acknowledgement. We thank the Editor and anonymous reviewers for comments that helped to improve the paper.

This work was supported in part by the Academy of Finland project SEMOHealth and the Division of Mathematical Science of the Russian Academy of Sciences and the Program of Strategic Development of Petrozavodsk State University.

References

[1] E. Altman, K. Avrachenkov, and U. Ayesta. A survey on discriminatory processor sharing. *Queueing Syst. Theory Appl.*, 53:53–63, June 2006. doi: 10.1007/s11134-006-7586-8

[2] E. Altman and P. Nain. Optimality of a threshold policy in the m/m/1 queue with repeated vacations. *Mathematical Methods of Operations Research*, 44:75–96, 1996. doi: 10.1007/BF01246330

[3] J. R. Artalejo, D. S. Orlovsky, and A. N. Dudin. Multi-server retrial model with variable number of active servers. *Comput. Ind. Eng.*, 48:273–288, March 2005. doi: 10.1016/j.cie.2005.01.013

[4] A. Bhandari, A. Scheller-Wolf, and M. Harchol-Balter. An exact and efficient algorithm for the constrained dynamic operator staffing problem for call centers. *Management Science*, 54(2):339–353, Feb. 2008. doi: 10.1287/mnsc.1070.0819

[5] S. Bhulai and G. Koole. On the structure of value functions for threshold policies in queuing models. *Journal of Appl. Probab.*, 40(3):613–622, 2003. doi: 10.1239/jap/1059060891

[6] S. R. Chakravarthy A multi-server queueing model with Markovian arrivals and multiple thresholds. *Asia Pac. J. Oper. Res.*, 24(2):223–243, (2007). MR 2320106 Zbl 1122.90020

[7] S. R. Chakravarthy and A. N. Dudin. A multi-server retrial queue with bmap arrivals and group services. *Queueing Syst. Theory Appl.*, 42:5–31, September 2002. doi: 10.1023/A:1019989127190

[8] T. B. Crabill, D. Gross, and M. J. Magazine. A classified bibliography of research on optimal design and control of queues. *Operations Research*, 25(2):219–232, Mar. 1977. doi: 10.1287/opre.25.2.219

[9] Security checkpoints. *Tiger team 2005. Improving throughput*, Dallas (Forth Worth)(DFW) International Airport, July 2006.

[10] R. Hassin and M. Haviv. *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*. Springer, 2002. Zbl 1064.60002 MR 2006433

[11] B. Hu and S. Benjaafar. Partitioning of servers in queueing systems during rush hour. *Manufacturing & Service Operations Management*, 11:416–428, July 2009. doi: 10.1287/msom.1080.0225

[12] J. J. Teghem. Control of the service process in a queueing system. *European Journal of Operational Research*, 23(2):141 – 158, 1986.

[13] M. A. Kaboudan. A dynamic-server queuing simulation. *Comput. Oper. Res.*, 25:431–439, June 1998. doi: 10.1016/S0305-0548(97)00090-7

[14] S. Karlin and J. McGregor. The differential equations of birth-and-death processes and the stieltjes moment problem. *Trans. Amer. Math. Soc.*, 85(2):489–546, July 1957.

[15] L. Kleinrock. *Queueing Systems*, volume I: Theory. Wiley Interscience, 1975. Zbl 0334.60045

[16] G. Koole. A simple proof of the optimality of a threshold policy in a two-server queueing system. *Syst. Control Lett.*, 26:301–303, December 1995.

[17] A. Levine and D. Finkel. Load balancing in a multi-server queuing system. *Computers and Operations Research*, 17(1):17 – 25, 1990. doi: 10.1016/0305-0548(90)90024-2

[18] R. Nelson and D. Towsley. Approximating the mean time in system in a multiple-server queue that uses threshold scheduling. *Operations Research*, 35(3):419–427, May 1987. doi: 10.1287/opre.35.3.419

[19] S. M. Ross. *Stochastic Processes*. John Wiley and Sons, New York, 1996. MR 1373653

[20] A. D. Solovjev. A problem of optimal queueing. *Technical Cybernetics*, 5:40–49, 1970.

[21] D. S. Szarkowicz and T. W. Knowles. Optimal control of an M/M/S queueing system. *Operations Research*, 33(3):644–660, May 1985. doi: 10.1287/opre.33.3.644

[22] N. Tian and Z. Zhang. A two threshold vacation policy in multiserver queuing systems. *European Journal of Operational Res*, 168:153–163, 2006. doi: 10.1016/j.ejor.2004.01.053

[23] W. Whitt. Understanding the efficiency of multi-server service systems. *Management Science*, 38(5):708–723, May 1992. doi: 10.1287/mnsc.38.5.708

[24] Z. G. Zhang, H. P. Luh, and C.-H. Wang. Modeling security-check queues. *Management Science*, 57(11):1979–1995, Nov. 2011. doi: 10.1287/mnsc.1110.1399

System kolejkowy ze zmienną liczba serwerów

Streszczenie. W pracy rozważany jest system kolejkowy ze zmienną liczbą serwerów zależną od długości kolejki. Przykładem takiego systemu jest system kontroli bezpieczeństwa na lotniskach. Liczba aktywnych serwerów zwiększa się, gdy kolejka pasażerów rośnie i zmniejsza się, gdy zgłoszenia do odprawy maleją. Pozwala to zaoszczędzić zasoby przy zachowaniu odpowiedniej wydajności (średnim czasie przebywania w kolejce) dla klientów. Otrzymano w zamkniętej formie czas obsługi, długość kolejki i średnią liczbę wykorzystanych serwerów. Dla sprawdzenia poprawności modeli posłużono się danymi z portu lotniczego Dallas - Fort Worth International, ósmego na świecie pod względem wielkości ruchu pasażerskiego. Badania symulacyjne potwierdziły rezultaty analityczne. Pozwala to zmniejszenie liczby otwartych serwerów przy jednoczesnej kontroli dopuszczalnego czasu oczekiwania na odprawę przez pasażera.

Słowa kluczowe: teoria kolejek, dynamiczna kolejka, kontrola bezpieczeństwa na lotniskach, planowanie przepustowości

Vladimir V. Mazalov finished his PhD studies at the Faculty of Applied Mathematics, Leningrad University in 1979. After that he has mainly worked in research projects funded by the Russian Academy of Sciences, in 1980-1998 in Chita Institute of Natural Resources, East Siberia and, currently, in the Institute of Applied Mathematical Research, Karelian Research Center. His research interests are related to game theory and stochastic analysis and applications in behavioral biology, networking and economic systems. He is Professor of the Chair of Probability Theory in Petrozavodsk State University and Director of the Institute of Applied Mathematical Research, Karelian Research Center, Russian Academy of Sciences.

Andrei Gurtov received his M.Sc (2000) and Ph.D. (2004) degrees in Computer Science from the University of Helsinki, Finland and M.Sc. (2001) in Applied Mathematics from Russia. He was appointed a Professor at the University of Oulu in the area of Wireless Internet in December 2009. He is also a Principal Scientist (on leave currently) leading the Networking Research group at the Helsinki Institute for Information Technology. He is an adjunct professor at Aalto University and the University of Helsinki. In 2000–2004, he was a senior researcher at Sonera Finland. In 2003–2005, he was a visiting researcher in the International Computer Science Institute at Berkeley, USA. In 2004, he was a consultant at the Ericsson NomadicLab.

In the Internet Engineering Task Force, Dr. Gurtov co-chaired the Host Identity Protocol Research Group (2005–2012) and co-authored five RFCs. He has supervised four PhD and 20 Master's theses. He is a senior member of IEEE and received the best paper award at IEEE Globecom'11. Dr. Gurtov is a co-author of over 120 publications including two books, research papers, and patents. His publications have received more than 1000 citations according to Google Scholar. His research interests include network security, peer-to-peer systems, transport protocols, mobile communication systems, and game theory.

Vladimir V. Mazalov
Institute of Applied Mathematical Research, Karelian Research Center
Russian Academy of Sciences, Petrozavodsk, Russia
E-mail: vmazalov@krc.karelia.ru

Andrei Gurtov
Aalto University
Department of Computer Science and Engineering, Finland
E-mail: gurtov@hiit.fi

(Received: 25th October 2012)