Analytical Call Center Model with Voice Response Unit and Wrap-Up Time

Petr HAMPL

Department of Telecommunication Engineering, Faculty of Electrical Engineering,
Czech Technical University in Prague, Technicka 2, 166 36 Prague, Czech Republic

petr.hampl@fel.cvut.cz

DOI: 10.15598/aeee.v13i4.1486

Abstract. *The last twenty years of computer integration significantly changed the process of service in a call center service systems. Basic building modules of classical call centers – a switching system and a group of humans agents – was extended with other special modules such as skills-based routing module, automatic call distribution module, interactive voice response module and others to minimize the customer waiting time and wage costs. A calling customer of a modern call center is served in the first stage by the interactive voice response module without any human interaction. If the customer requirements are not satisfied in the first stage, the service continues to the second stage realized by the group of human agents. The service time of second stage – the average handle time – is divided into a conversation time and wrap-up time. During the conversation time, the agent answers customer questions and collects its requirements and during the wrap-up time (administrative time) the agent completes the task without any customer interaction. The analytical model presented in this contribution is solved under the condition of statistical equilibrium and takes into account the interactive voice response module service time, the conversation time and the wrap-up time.*

Keywords

Administrative time, call center, interactive voice response, handle time, queueing systems, wrap-up time.

1. Introduction

The proposed analytical model belongs to the category of Markovian models which means that the flow of incoming calls is described by homogeneous Poisson process and all service times are modeled by exponentially distributed random variables. Figure 1 shows a principal queuing model of an inbound call center with N incoming trunk lines, interactive voice response (IVR) module with a queue and a group of S agents ($S \leq N$). The incoming calls are described by i.i.d. exponential random variable with average arrival rate λ and CDF

$$F(t) = \begin{cases} 1 - e^{-\lambda t} = 1 - e^{-\frac{t}{t_p}} & \text{if } t \geq 0, \\ 0 & \text{if } t < 0, \end{cases} \tag{1}$$

parameter t_p represents the mean interarrival time. The incoming calls are routed through N trunk lines and the switching matrix to the IVR module, where the service times are modeled by an i.i.d. exponential random variables with the parameter mean service rate θ and CDF

$$F_I(t) = \begin{cases} 1 - e^{-\theta t} = 1 - e^{-\frac{t}{t_I}} & \text{if } t \geq 0, \\ 0 & \text{if } t < 0. \end{cases} \tag{2}$$

Parameter t_I is the mean value of service times in IVR module. An incoming call can be rejected with the probability of loss B in case the all N trunk lines are busy at arrival time. After the first phase of service in the IVR module is completed, the call may leave the system with the probability $1 - p$ or it may request a human assistance from a free agent with the probability p. The average output rate of served calls that finish its service in the IVR module without an agent interaction is

$$\lambda_I = (1 - B)(1 - p)\lambda. \tag{3}$$

The presented model supposes that these calls have completed its service and will no longer interact with the call center. The complementary part of calls that decide to continue to the second phase of service has average rate

$$\lambda_{IA} = (1 - B)p\lambda. \tag{4}$$

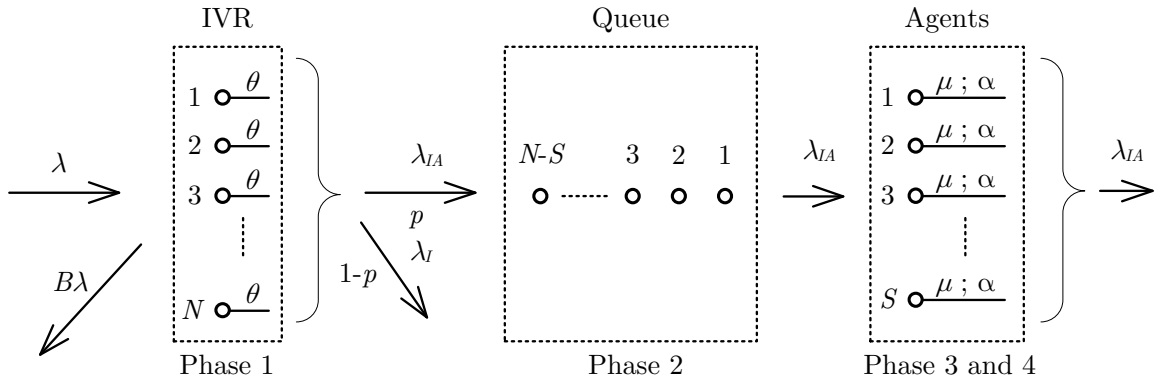

Fig. 1: Call center model as a service system.

In the second phase, a call is either assigned to an available agent or wait in the queue until an agent becomes free. The third phase of service – the conversation with an agent – is represented by i.i.d. exponential random variables with mean service rate μ and CDF

$$F_C(t) = \begin{cases} 1 - e^{-\mu t} = 1 - e^{-\frac{t}{t_C}} & \text{if } t \geq 0, \\ 0 & \text{if } t < 0, \end{cases} \tag{5}$$

where t_C is the mean conversation time of calls. Once a customer completes its conversation with the assigned agent, the trunk line is released and the service continues to the last phase – the after call work. In this phase, the agent completes the tasks related to the call. Which is as well represented by an i.i.d. exponential random variable with mean rate α and CDF

$$F_A(t) = \begin{cases} 1 - e^{-\alpha t} = 1 - e^{-\frac{t}{t_A}} & \text{if } t \geq 0, \\ 0 & \text{if } t < 0, \end{cases} \tag{6}$$

where parameter t_A is the mean value of administrative times. Once an agent completes the after call work, it is available for next call waiting in queue or coming directly from IVR module if the queue is empty. The next section describes the analytical model of above-described call center model.

2. Analytical Model

From the queuing theory point of view, the Markovian model from the previous chapter can be described by three-dimensional state space (i, j, k) with

$$n = \frac{(N+1)(N+2)(S+1)}{2}, \tag{7}$$

stationary probabilities of states $P(i, j, k)$, where the index i represents the number of calls in IVR module, the index j represents the sum of the number of active conversation with agents and the number of calls waiting in the queue. The last index k represent the number of agents in administrative state – finishing the after call work. The three-dimensional state space is limited by following inequalities

$$\begin{aligned} 0 \leq i \leq N, 0 \leq j \leq N, 0 \leq k \leq S, \\ 0 \leq S \leq N, i + j \leq N. \end{aligned} \tag{8}$$

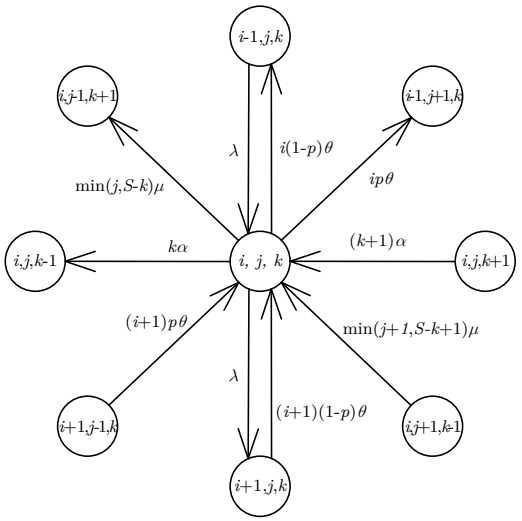

Fig. 2: Possible transitions between inner state and neighbour states.

All possible transitions between an inner (i, j, k) state and all neighbour states are shown in Fig. 2 and generally described by Eq. (9) which take into account five types of following transitions:

- $(i, j, k) \rightarrow (i + 1, j, k)$ or $(i - 1, j, k) \rightarrow (i, j, k)$ represents an incoming call to the call center that found a free line and start its service in IVR module.
- $(i, j, k) \rightarrow (i - 1, j + 1, k)$ or $(i + 1, j - 1, k) \rightarrow (i, j, k)$, represents a call that request an assistance by an agent after completed service in IVR module.

$$\begin{aligned}&[i\,(1-p)\,\theta + ip\theta + \lambda + k\alpha + \min(j, S-k)\,\mu]\,P_{i,j,k} = \lambda\,P_{i-1,j,k} + (k+1)\,\alpha\,P_{i,j,k+1}\\&+\min(j+1, S-k+1)\,\mu\,P_{i,j+1,k-1} + (i+1)\,(1-p)\,\theta\,P_{i+1,j,k} + (i+1)\,p\theta\,P_{i+1,j-1,k}\end{aligned} \tag{9}$$

$$\sum_{i=0}^{N}\sum_{j=0}^{N-i}\sum_{k=0}^{S} P_{i,j,k} = 1 \tag{10}$$

$$\begin{aligned}\lambda P_{0,0,0} &= (1-p)\,\theta P_{1,0,0} + \alpha P_{0,0,1}\\(\alpha+\lambda)\,P_{0,0,1} &= (1-p)\,\theta P_{1,0,1} + \mu P_{0,1,0}\\(\lambda+\mu)\,P_{0,1,0} &= p\theta P_{1,0,0} + (1-p)\,\theta P_{1,1,0} + \alpha P_{0,1,1}\\(\alpha+\lambda)\,P_{0,1,1} &= p\theta P_{1,0,1} + (1-p)\,\theta P_{1,1,1} + \mu P_{0,2,0}\\\mu P_{0,2,0} &= p\theta P_{1,1,0} + \alpha P_{0,2,1}\\\alpha P_{0,2,1} &= p\theta P_{1,1,1}\\(\lambda + (1-p)\,\theta + p\theta)\,P_{1,0,0} &= 2\,(1-p)\,\theta P_{2,0,0} + \alpha P_{1,0,1} + \lambda P_{0,0,0}\\(\alpha+\lambda+(1-p)\,\theta + p\theta)\,P_{1,0,1} &= 2\,(1-p)\,\theta P_{2,0,1} + \lambda P_{0,0,1} + \mu P_{1,1,0}\\(\mu + (1-p)\,\theta + p\theta)\,P_{1,1,0} &= 2p\theta P_{2,0,0} + \alpha P_{1,1,1} + \lambda P_{0,1,0}\\(\alpha + (1-p)\,\theta + p\theta)\,P_{1,1,1} &= 2p\theta P_{2,0,1} + \lambda P_{0,1,1}\\(2\,(1-p)\,\theta + 2p\theta)P_{2,0,0} &= \alpha P_{2,0,1} + \lambda P_{1,0,0}\\(\alpha + 2\,(1-p)\,\theta + 2p\theta)\,P_{2,0,1} &= \lambda P_{1,0,1}\end{aligned} \tag{11}$$

- $(i,j,k) \rightarrow (i-1,j,k)$ or $(i+1,j,k) \rightarrow (i,j,k)$, represents a call leaving the call center after completed service in IVR module.
- $(i,j,k) \rightarrow (i,j-1,k+1)$ or $(i,j+1,k-1) \rightarrow (i,j,k)$, represents the transition of an agent from the conversation phase to the administrative phase.
- $(i,j,k) \rightarrow (i,j,k-1)$ or $(i,j,k+1) \rightarrow (i,j,k)$, represents the completion of administrative phase and leaving the system.

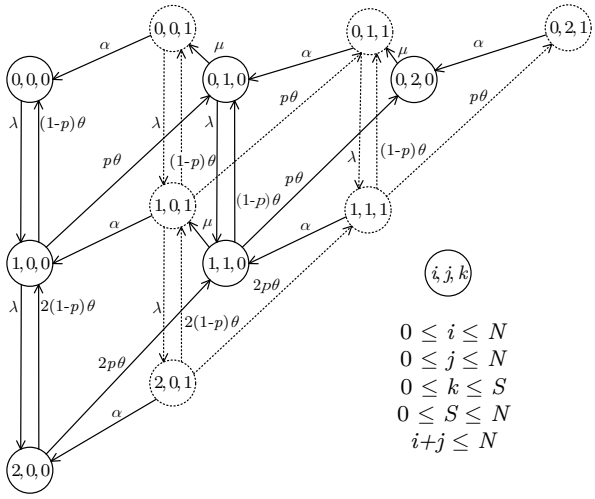

Fig. 3: Example of state space for the system $N = 2$ and $S = 1$.

The Eq. (9) and state space limits in Eq. (8) form a set of n linear equations where each one represents the equality of rate of transitions out of a given state (i,j,k) and the rate of transitions into that state, in steady state [3], [4] or [5]. In case of boundary states, it is important to omit terms that correspond to transitions that do not exist. The set of above mentioned linear equations should be normalized with condition described in Eq. (10).

An example of three-dimensional state space with all possible transitions for small model with two trunk lines and one agent ($N = 2$, $S = 1$) is shown in the Fig. 3.

The concrete set of equations for this small system is derived in Eq. (11). All twelve states belongs in this case to the boundary states.

Index k corresponds to a layer of the three-dimensional model and indexes i, j are the same in each layer. Figure 3 has two layers, zero layer is illustrated by the solid line and layer one use dotted line. If the transition rate α is going to infinity the three-dimensional model converges to the two-dimensional model presented in [1] or [2].

Numerical solution for larger values of N and S can be challenging. For the system with $N = 100$ trunk

lines and $S = 70$ agents the state space has $n = 365\,721$ states. The number of elements in a matrix that represent the set of linear Eq. (9) is $365\,721^2 = 1.3 \cdot 10^{11}$ and required memory to save them in double precision format is approximately 1 TB. Fortunately, the matrix belongs to the category of band-diagonal sparse matrices (each row has only six nonzero elements) [6]. The sparsity is 0.00016 in this case. Therefore, the solution of probability state space leads to the application of methods that uses sparse arrays and is solvable in an acceptable time.

3. System Parameters

The probability of blocking or loss B of an incoming call is given by sum of boundary probabilities of states $P(i,j,k)$ that doesn't have the transition to neighbour state (i,j,k)

$$B = \sum_{i=0}^{N} \sum_{k=0}^{S} P_{i,N-i,k}. \tag{12}$$

The probability of zero waiting time of an incoming call after its service in IVR module is

$$P(W_{IA} = 0) = \frac{\sum_{i=1}^{N} \sum_{j=0}^{N-i} \sum_{k=0}^{S-1-j} iP_{i,j,k}}{\sum_{i=1}^{N} \sum_{j=0}^{N-i} \sum_{k=0}^{S} iP_{i,j,k}}. \tag{13}$$

The mean number of calls in IVR module $\mathrm{E}[X_I]$ is

$$\mathrm{E}[X_I] = \sum_{i=1}^{N} \sum_{j=0}^{N-i} \sum_{k=0}^{S} iP_{i,j,k}. \tag{14}$$

For the system in state (i,j,k) is the number of waiting calls $\max(0, j+k-S)$ and the mean number of calls in the queue is then

$$\mathrm{E}[X_Q] = \sum_{i=1}^{N} \sum_{j=0}^{N-i} \sum_{k=0}^{S} \max(0, j+k-S)\, P_{i,j,k}. \tag{15}$$

Similarly the mean number of active conversations with agents is

$$\mathrm{E}[X_C] = \sum_{i=1}^{N} \sum_{j=0}^{N-i} \sum_{k=0}^{S} [j - \max(0, j+k-S)]\, P_{i,j,k}, \tag{16}$$

and the mean number of agents in administrative phase is given by equation

$$\mathrm{E}[X_A] = \sum_{i=0}^{N} \sum_{j=0}^{N-i} \sum_{k=1}^{S} kP_{i,j,k}. \tag{17}$$

The mean number of calls $\mathrm{E}[X_T]$ in call center is equal to the sum of mean values of calls in IVR module $\mathrm{E}[X_I]$, calls in queue $\mathrm{E}[X_Q]$ and active conversations with agents $\mathrm{E}[X_C]$

$$\mathrm{E}[X_T] = \mathrm{E}[X_I] + \mathrm{E}[X_Q] + \mathrm{E}[X_C]. \tag{18}$$

The definition of mean waiting time is significantly influenced by the location of measurement. It is important to know, to what portion of calls the mean waiting time is related. The mean waiting time $\mathrm{E}[W_0]$ related to all offered calls is according to Little's law equal to

$$\mathrm{E}[W_0] = \frac{\mathrm{E}[X_Q]}{\lambda}, \tag{19}$$

similarly the mean waiting time $\mathrm{E}[W]$ related to all calls served by the call center is equal to

$$\mathrm{E}[W] = \frac{\mathrm{E}[X_Q]}{\lambda(1-B)}, \tag{20}$$

and mean waiting time $\mathrm{E}[W_{IA}]$ related to all calls requested service by an agent (calls that decide to continue to the second phase of service) is equal to

$$\mathrm{E}[W_{IA}] = \frac{\mathrm{E}[X_Q]}{\lambda_{IA}} = \frac{\mathrm{E}[X_Q]}{\lambda(1-B)\,p}. \tag{21}$$

Finally the men waiting time $\mathrm{E}[W_W]$ related to all call that really wait in queue

$$\mathrm{E}[W_W] = \frac{\mathrm{E}[X_Q]}{\lambda(1-B)\ p\ (1 - \mathrm{P}(W_{IA} = 0))}. \tag{22}$$

The next chapter presents some numerical and simulation results.

4. Numerical Results

To verify the correctness of presented model a simulation program in C++ has been written. In Tab. 1 is a short summary of results for call center with $N = 100$ trunk lines and $S = 70$ agents. Simulation time of one hundred intervals with length 500 hours takes 46 seconds on Intel CPU i7-3520M 2.9 GHz.

Tab. 1: Results comparison of analytical and simulation model.

$N = 100, S = 70, \lambda = 0.1818\ \mathrm{s}^{-1}$ $t_I = 100\ \mathrm{s}, t_C = 360\ \mathrm{s}, t_A = 180\ \mathrm{s}, p = 0.7$		
Parameter	**Model**	**Simulation**
B	0.01074	0.01062 ± 1.99 %
$\mathrm{E}[W_0]$	58.9930 s	58.9031 s ± 0.94 %
$P(W_0 > 120\,\mathrm{s})$		0.22567 ± 1.19 %
$P(W_0 = 0)$	0.50451	0.50468 ± 0.44 %
$P(W_0 > 0)$	0.49549	0.49532 ± 0.45 %
$\mathrm{E}[W]$	59.6336 s	59.5378 s ± 0.96 %
$P(W > 120\,\mathrm{s})$		0.22810 ± 1.20 %
$P(W = 0)$	0.49913	0.49936 ± 0.47 %
$P(W > 0)$	0.50087	0.50064 ± 0.47 %
$\mathrm{E}[W_{IA}]$	85.1908 s	85.0584 s ± 0.94 %
$P(W_{IA} > 120\,\mathrm{s})$		0.32588 ± 1.19 %
$P(W_{IA} = 0)$	0.284471	0.28475 ± 1.13 %
$P(W_{IA} > 0)$	0.71553	0.71525 ± 0.45 %
$\mathrm{E}[W_W]$	119.060 s	118.999 s ± 0.57 %

All simulation results correspond to the results of presented analytical solution. The confidence level 95 % was used for all simulation outputs.

5. Influence of Wrap-Up Time

There are many models that do not respect the last phase of service – the after call work. To this category also belongs the two-dimensional model published in [2] or simpler and fundamental Erlang queuing model $M/M/S/N$. All these models often use a simple correction of wrap-up time absence, they only add the mean wrap-up time to the mean conversation time. The following analysis tries to quantify the impact of such simple correction on key parameters of the proposed model that exactly respect important phases of service.

Following parameters are constant: number of trunk lines $N = 40$, number of agents $S = 23$, probability of assistance of an agent $p = 0.7$, input intensity $\lambda = 0.1\ \mathrm{s}^{-1}$, mean service time in IVR module $t_I = 120$ s and mean occupation time of an agent $t_C + t_A = 300$ s.

All graphs in this chapter use on x-axis the t_A/t_C ratio. A simple addition of the wrap-up time to conversation time would increase the average load of agents. To minimize this negative effect the following analysis has been done with constant mean agent occupation time $t_A + t_C$. This means that for $t_A/t_C = 1$ is $t_A = t_C = 150$ s and for $t_A/t_C = 2$ is mean conversation time $t_C = 100$ s and mean wrap-up time $t_A = 200$ s.

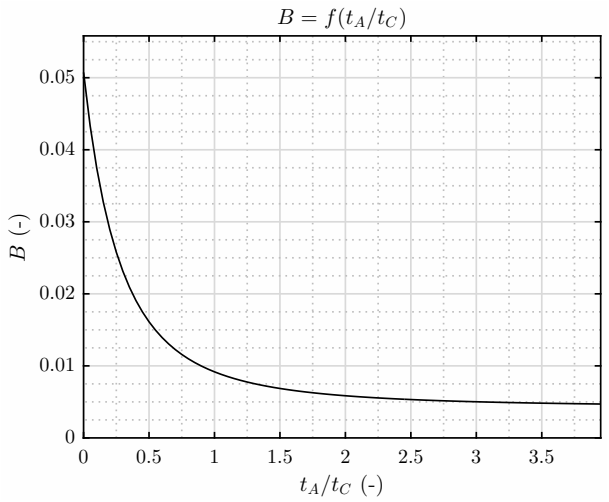

Fig. 4: Influence $B = f(t_A/t_C)$, for system with parameters $N = 40$, $S = 23$, $p = 0.7$, $\lambda = 0.1\ \mathrm{s}^{-1}$, $t_I = 120$ s, $t_C + t_A = 300$ s.

Figure 4 shows the influence of increasing t_A/t_C ratio on blocking probability B of the call center. The blocking probability falls down because the line utilization in the conversation phase falls down. A part of trunk lines unused by agents in conversation phase is used for holding calls in queue or accepting new incoming calls. The starting value $B = 0.05$ for $t_A/t_C = 0$ corresponds with results of the two-dimensional model mentioned in [1] or [2]. The blocking probability B is below one percent if the ratio $t_A/t_C = 1$.

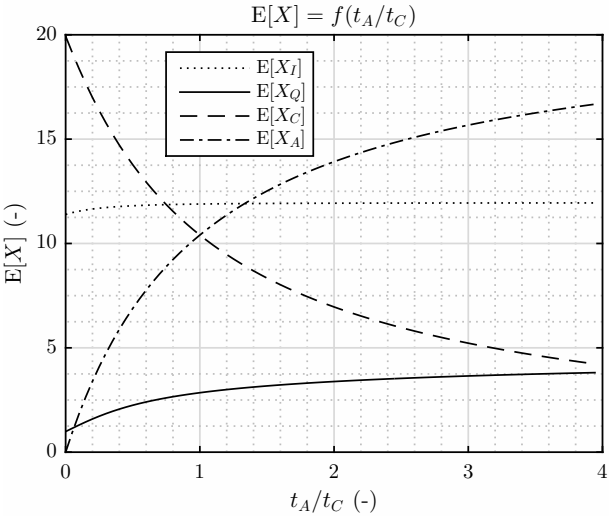

Fig. 5: Influence $\mathrm{E}[X] = f(t_A/t_C)$, for system with parameters $N = 40$, $S = 23$, $p = 0.7$, $\lambda = 0.1\ \mathrm{s}^{-1}$, $t_I = 120$ s, $t_C + t_A = 300$ s.

The same effect is observable in the Fig. 5 where the mean number of calls in queue and also in IVR module has a little increasing trend. There is also shown significantly increasing trend of mean number of calls $\mathrm{E}[X_A]$ in administrative phase and adequately decreasing the mean number of calls $\mathrm{E}[X_C]$ in conversation phase.

Figure 6 displays the values of probability of waiting

$$P(W > 0) = p\,P(W_{IA} > 0) = \frac{P(W_0 > 0)}{1 - B}. \quad (23)$$

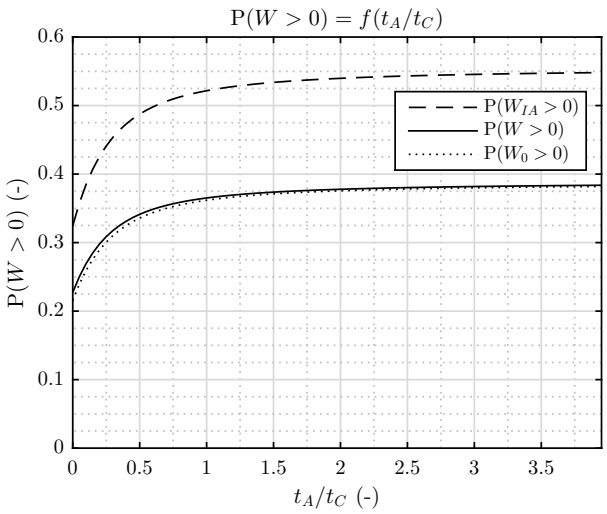

Fig. 6: Influence $P(W > 0) = f(t_A/t_C)$, for system with parameters $N = 40$, $S = 23$, $p = 0.7$, $\lambda = 0.1$ s^{-1}, $t_I = 120$ s, $t_C + t_A = 300$ s.

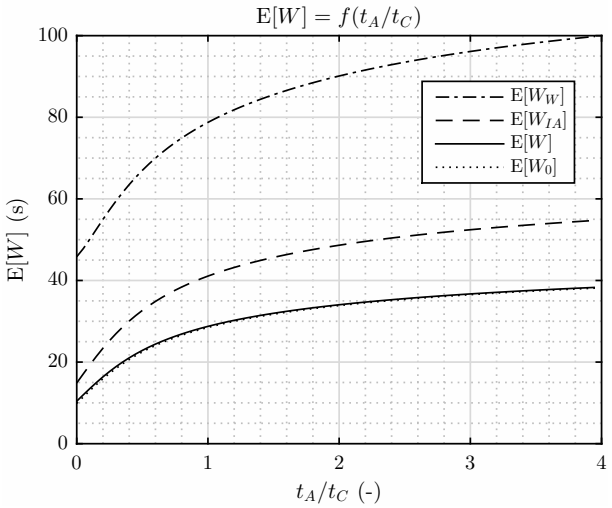

Fig. 7: Influence $\mathrm{E}[W] = f(t_A/t_C)$, for system with parameters $N = 40$, $S = 23$, $p = 0.7$, $\lambda = 0.1$ s^{-1}, $t_I = 120$ s, $t_C + t_A = 300$ s.

Again is shown a significantly increasing trend for very small values of t_A/t_C ratio. The next important parameter from the customer's point of view is mean waiting time $\mathrm{E}[W]$, see Fig. 7.

The ratio t_A/t_C of typical call centers is in the range from 0 to 1, where is shown the greatest increase in mean waiting time. The starting values for $t_A/t_C = 0$ and limiting values for $t_A/t_C \to \infty$ are in Tab. 2.

Tab. 2: Comparison of limiting values of mean waiting times $\mathrm{E}[W_0]$, $\mathrm{E}[W]$, $\mathrm{E}[W_{IA}]$, $\mathrm{E}[W_W]$.

$N = 40, S = 23, \lambda = 0.1$ s$^{-1}, t_I = 120$ s, $p = 0.7$		
Param.	$t_C = 300$ s, $t_A = 0$ s	$t_C = 0$ s, $t_A = 300$ s
$\mathrm{E}[W_0]$	9.9 s	44.6 s
$\mathrm{E}[W]$	10.4 s	44.8 s
$\mathrm{E}[W_{IA}]$	14.9 s	66.8 s
$\mathrm{E}[W_W]$	46 s	119.5 s

6. Conclusion

In this paper is presented analytical solution of call center queuing model under statistical equilibrium that explicitly describe the service in IVR module and after call work of agents. The numerical results of the presented analytical model exactly correspond to values obtained from simulation program that author developed in C++ language for this type of system. The analysis in the section five shows significant negative influence of small values of wrap-up time on the mean waiting time and on the probability of waiting $P(W > 0)$, even if the mean occupation time of an agent is constant $t_C + t_A = 300$ s.

References

[1] KHUDYAKOV, P., P. FEIGIN and A. MANDELBAUM. Designing a call center with an IVR (Interactive Voice Response). *Queueing Systems*. 2010, vol. 66, iss. 3, pp. 215–237. ISSN 0257-0130. DOI: 10.1007/s11134-010-9193-y.

[2] SRINIVASAN, R., J. TALIM and J. WANG. Performance analysis of a call center with interactive voice response units. *Top*. 2004, vol. 12, iss. 1, pp. 91–110. ISSN 1134-5764. DOI: 10.1007/BF02578926.

[3] GROSS, D., J. F. SHORTLE, C. M. HARRIS and J. M. THOMPSON. *Fundamentals of Queueing Theory*. New York: John Wiley & Sons, 2008. ISBN 978-0471791270.

[4] KLEINROCK, L. *Queueing systems - Volume I: Theory*. New York: John Willey & Sons, 1975. ISBN 978-0471491101.

[5] ZITEK, F. *Ztraceny cas - Elementy teorie hromadne obsluhy*. Prague: Academia, 1969.

[6] PRESS, W. H., S. A. TEUKOLSKY, V. T. VETTERLING and B. P. FLANNERY. *Numerical*

Recipes 3rd Edition: The Art of Scientific Computing. New York: Cambridge University Press, 2007. ISBN 978-0521880688.

About Authors

Petr HAMPL was was born in Mestec Kralove, Czech Republic in 1979. He received his Master degree (Ing.) in 2004 and doctor degree (Ph.D.) in 2011 at Faculty of Electrical Engineering, Czech Technical University in Prague, specializing in Telecommunication Engineering. Currently he works as an assistant professor at the Deptartment of Telecommunication Engineering of the Czech Technical University in Prague. His research activities are mainly focused on queuing theory, simulations, models and their applications in telecommunication area.