Journal of ELECTRICAL ENGINEERING, VOL 71 (2020), NO2, 96–102

sciendo

PAPER

Properties of a modelled call centre

Ivan Baronak, Matej Hartmann, Robert Polacek[1]

To detect the number of agents needed to serve customers, it is necessary to consider the call centre as a mass service system. Then it is possible to asses the convenient number of agents according to the probability of the system receiving a request and the time in which the request is serviced by employing a Markov chain and the Erlang model. In an archetypal call centre, the incoming calls are added to a waiting queue and subsequently they are assisted by an agent. In case all agents are occupied, the customer has to wait in the queue until one of the agents becomes available. It is, therefore, important to compromise on the number of agents and the time the customers spend waiting in the queue. The result should be that there are enough agents in the call centre to serve the customers in the time required. This article focuses on solving this problem.

K e y w o r d s: call centre, IVR, mass service system, call agent, Erlang equations

1 Introduction

The task of the call centre is to provide the calling customer efficient service by one or more agents. Concurrently, each agent must be sufficiently trained and properly equipped [1].

The call centre is a complex of technical equipment formed into an info communication structure of agents (operators) and a supervisor [1].

The call centre typically consists of several key components:

- interactive voice reply (IVR) [2],
- automatic call distribution (ACD) [2],
- work facilities [2],
- session border controller (SBC) [2].

Call centres comprise many other elements (a database server, a recording system, a campaign manager), whose description is not included in this article. The interconnection of the components is shown in Fig. 1.

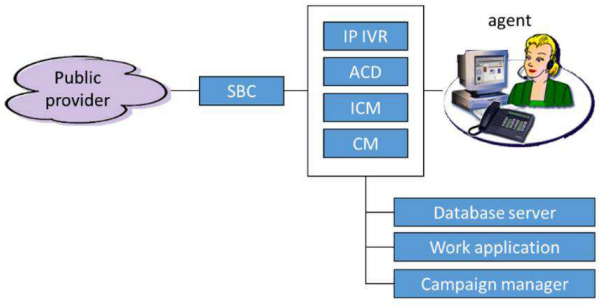

Fig. 1. Components of a call centre.

2 Mass service system

To determine the optimal number of agents, it is necessary to consider the call centre as a mass service system. This system consists of one or more service channels that control the flow of requests entering the system. The already serviced requests then depart from the system and are considered to be successfully accomplished.

However, a situation may appear in which there are not enough servers in the system to allow service to all arriving requests. In that case, the request is either refused or added to a queue, where it waits for the server to become available.

The mass service system is composed of several parts, see Fig. 2. The parts are:

- queue - requests waiting to be served are added there if none of the servers is vacant,
- service channels - servers providing service of the requests from the queue
- departure flow - successfully accomplished requests leave the system.

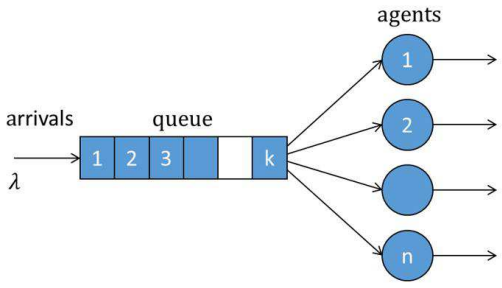

Fig. 2. The mass service system

[1] Slovak Technical University in Bratislava, Faculty of Electrical Engineering, Department of Telecommunication Engineering, Ilkovičova 3, 812 19 Bratislava, Slovakia ivan.baronak@ut.fei.stuba.sk, matej.hartmann@gmail.com, xpolacekr2@is.stuba.sk

DOI:10.2478/jee-2020–0014, Print (till 2015) ISSN 1335-3632, On-line ISSN 1339-309X

A typical example of such a system is a call centre. The first contact of the customer is IVR. According to options selected by the customers in the IVR it is then possible for them to subsequently deal with the request on their own or to redirect them to an agent. If no agents are available now, the customer is added to the queue and waits for an agent to become unoccupied. After the service is successfully provided by the agent, the customer leaves the system.

2.1 Modelling a call centre

In relation to the previous information, we can start modelling the call centre. It consists of fragments shown in Fig. 2. These fragments have specific properties described in this subchapter.

2.1.1 Flow of incoming requests

The intensity of the flow of arriving requests is a deciding factor affecting the functioning of the system. The flow of arriving requests can be characterised as a random process, since requests arrive at the system randomly. It has been discovered by having obtained statistic data from a large number of systems that the flow of arriving requests in call centres follows the Poisson distribution [3–5].

Discrete random variable X follows the Poisson distribution with a parameter $\lambda > 0$, if it possesses value of $k = 0, 1, 2, \ldots$ with probability [6]

$$p_k = P[X = k] = \frac{\lambda^k}{k!} e^{-\lambda} \tag{1}$$

In which:
p_k – probability of k value,
$P[X = k]$ – probability that X value equals k value,
λ – mean value of the Poisson distribution.

The mean value of the Poisson distribution is λ and it simultaneously demonstrates dispersion. This means that requests arrive at the call centre according to the Poisson distribution at an average arrival rate λ. The interval between the incoming requests is then and it can be characterised as exponential distribution [6, 7].

2.1.2 Queue

If the number of service channels in the system is sufficient, the request is accepted and responded. If now the system has no vacant service channel, the request is added to the queue, where it waits for the service channel to become available.

The queue is of FIFO type and the requests, therefore, arrive at the service channel in the same order as in which they were added to the queue - the first request that came into the queue is given the service first.

For the requirements of the call centre we can assume that there is exactly one queue and that it is infinite. Thus, no customer would be refused and there would not occur a busy signal in case all agents are currently occupied. The customers would be added to the queue.

2.1.2 Request service

All incoming requests are serviced by pool of N identical agents at an average rate $N\mu$. It has been discovered that similarly as the request arrival, the request service obeys the exponential distribution [3].

The combined random variable x follows an exponential distribution with a parameter $\lambda > 0$ if the probability density is expressed as [6]

$$f(x) = \begin{cases} \lambda e^{-\lambda} & \text{for } x > 0\,, \\ 0 & \text{else}\,. \end{cases} \tag{2}$$

The request service, therefore, follows the exponential distribution at an average service rate of μ^{-1}. After having been serviced, the request leaves the system.

The exponential and Poisson distributions express the same instance from two different points of view. While the Poisson distribution determines the probability of certain occurrences in a unit of time, the exponential distribution determines the probability of the interval length between two successive occurrences [6].

2.2 Erlang equations

As mentioned in chapter 2.1, requests arrive at the call centre following the Poisson distribution and are given service following the exponential distribution. This means that this is an M/M model according to Kendall's notation.

To model the call centre and assess the optimal number of agents we can employ Erlang models:

- Erlang B (M/M/N/N),
- Erlang C (M/M/N),
- Erlang A (M/M/N+M).

2.2.1 Erlang B

Erlang B model allows to detect the probability of call losses for a group of identical parallel resources at a given number of agents N and operation load R. However, the model is not suitable for assessing the number of agents in a contemporary call centre because it does not include the queue (see Fig. 3).

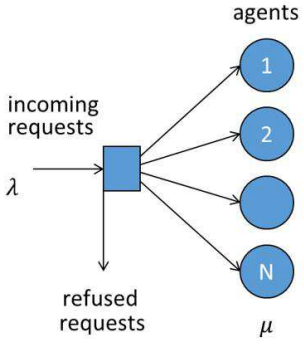

Fig. 3. Erlang B model

Erlang B model assumes that if there is no available agent, the incoming request is refused. For that reason, it is not possible to employ the model. It is unacceptable for a call centre not to serve a customer. Therefore, this model is not going to be explored further [8].

2.2.2 Erlang C

The second Erlang equation defines the probability that the request waits to be provided service in the queue. The model assumes that requests added to the queue stay there until they are serviced, and the queue is infinite. For better illustration of Erlang C model, see Fig. 4 [9].

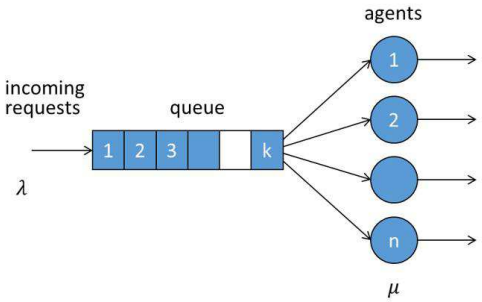

Fig. 4. Erlang C model

To calculate the number of agents, it is necessary first to calculate the operation load

$$R = \frac{\lambda}{\mu} \tag{3}$$

where: R is the operation load, λ is the request arrival rate, and μ is the rate of request service provided by agents.

Then, it is possible to calculate the load per agent

$$\rho = \frac{\lambda}{N\mu} = \frac{R}{N} \tag{4}$$

where ρ is the load per agent, and N is the number of agents.

The load per agent determines how much of the available time the agent uses to serve customers given the assumption that all requests are serviced. The value of the parameter must strictly be less than 1 (the agent must not be loaded for more than 100 %). Otherwise the system becomes unstable the requests would arrive at a higher rate than they are serviced, and the queue would grow to infinity.

Consequently, we can calculate the probability of the request being added to the queue

$$P\{W > 0\} = \frac{\frac{R^N}{N!}\frac{N}{N-R}}{\sum_{m=0}^{N-1}\frac{R^m}{m!} + \frac{R^N}{N!}\frac{N}{N-R}} \tag{5}$$

where, $PW > 0$ is the probability that the request would have to wait, under condition that $N > A$.

In other words, this in way it is possible to calculate the probability that all agents are occupied.

From this value it is possible to figure out other important parameters of the call centre. One of the parameters is ASA (Average Speed of Answer) or how long the request waits for service:

$$ASA = E[W] = P\{W > 0\}E\left[W|W > 0\right] = P\{W > 0\}\frac{1}{N\mu\rho} \tag{6}$$

where, ASA is the average speed of answer, E is the mean value.

Important performance metric of the call centre is TSF (Telephone Service Factor). This value determines the percentage of requests that are serviced in a given time. For example: the percentage of requests that are serviced in 30 seconds. The telephone service factor is

$$TSF = P\{W \leq T\} = 1 - P\{W > 0\}P\{W > T|W > 0\} = 1 - P\{W > 0|\}e^{-N\mu(1-\rho)T}. \tag{7}$$

2.3 Disadvantages of the model

The fourth and last parameter of a call centre is the number of requests that abandon the queue before being provided service. However, employing Erlang C model, it is impossible to asses this parameter, since it assumes that all requests remain in the queue until they are serviced [9].

According to the up-to-date research, Erlang C model is not completely accurate in assessing the number of agents in a call centre, since the model is responsible for certain inaccuracies or mistakes in the process. The highest contribution to the mistake is the fact that Erlang C model neglects the possibility of a request abandoning the queue before it gets to the agent to be serviced. Erlang A model, described in the following chapter, amends this disadvantage [8].

The absence of cancellation of the waiting request in this model results in Erlang C model requires a rather higher number of agents in the call centre than needed. However, agents represent the biggest expense of a call centre (60–75 %). For that reason, the excessive number of agents is not economically favourable [9].

Another potential inaccuracy is the assumption that requests arriving at the system follow the Poisson distribution. This distribution does not always accurately describe how the requests arrive at the call centre because this random process is very hard to predict. The request arrival rate can be affected by instances such as the weather or an ongoing advertising campaign, which could cause greater deviations than the Poisson process is able to predict. This disadvantage can be compensated to some extent by a manager of the call centre making early predictions of operation according to the ongoing advertising campaigns and system statistics from the past [9].

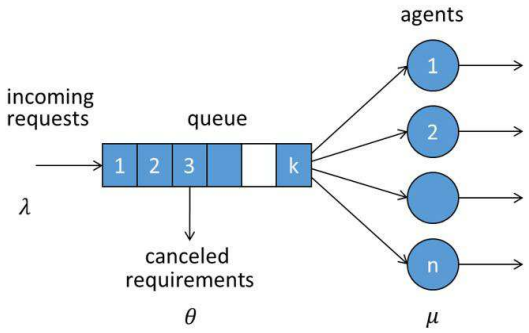

Fig. 5. Erlang A model

Table 1. Requests for a call centre

	Small company	Large company
Number of calls per hour	100	900
Service time (min)	3	3
Number of customers served in 20 seconds (service level) (%)	80	80
Patience (min)	1	1

To assume that the request service follows the exponential distribution is also not utterly correct. It is assumed that all agents are homogeneous, they have the same skills and service is performed equally regardless of a specific agent. However, this might not be the truth. One agent might be providing service more efficiently than the other. This problem could be removed to some extent by creation of groups of agents focused on specific problems according to the skills of agents so that the highest possible efficiency of service is achieved [9].

2.3.1 Erlang A

Erlang A model is an extension of Erlang C model. It gives the requests the possibility to abandon the queue before they are serviced. Erlang A gives the request a new property patience θ^{-1} . Patience expresses how much time the customers are willing to spend in the queue until they are served. If a customer is waiting for longer than the given time, he or she will abandon the queue. In this way, the major disadvantage of Erlang C model, the mistake caused by the inability to abandon the queue, is eliminated.

The patience of customers obeys the exponential distribution.

For better demonstration of Erlang A model, see Fig. 5. The scheme shows that in comparison with Erlang C, the abandoned request has been added to the model. The abandoned requests comprise those requests that exceeded the time limit for waiting and left the queue. In this way, another disadvantage of Erlang C model has been removed - the system will never become unstable. Since the waiting requests eventually leave the queue, the queue cannot grow to infinity [8].

Erlang A model is, therefore, characterised by four parameters:

- λ - request arrival rate (calls per unit of time),
- μ – request service rate (the average service time $1/\mu$),
- N - number of agents,
- θ - queue abandonment rate (the average time the customer is willing to wait is $1/\theta$).

All parameters are unrelated.

Erlang A model consists of two blocks: J and ε, as defined in [8, 10, 11].

$$J = \frac{e^{\lambda/\theta}}{\theta}\left(\frac{\theta}{\lambda}\right)^{N\mu/\theta}\gamma\left(\frac{N\mu}{\theta}, \frac{\lambda}{\theta}\right), \tag{8}$$

$$\varepsilon = \frac{\sum_{j=0}^{N-1}\frac{1}{j!}\left(\frac{\lambda}{\mu}\right)^{j}}{\frac{1}{(N-1)!}\left(\frac{\lambda}{\mu}\right)^{N-1}}. \tag{9}$$

The probability that the request will wait in the queue can be calculated as

$$P\{W > 0\} = \frac{\lambda J}{\varepsilon + \lambda J}(1-\theta). \tag{10}$$

As demonstrated, the addition of the possibility to abandon the queue to the model of the call centre caused notable complications in calculation. Due to the calculation being so demanding, to model a call centre according to Erlang A model in practice, Java CCOptim library is commonly used [12]. Subsequently, the differences in results of predictions of the number of agents for Erlang C and Erlang A models are compared.

3 Call centre optimisation

The most expensive part of a call centre are the human resources. It is, therefore, important that the call centre have an optimal number of agents. An excessive number of agents would, however, cause the business operation to become unnecessarily expensive. A small number of agents would cause the requests to wait in the queue for too long and, therefore, customer satisfaction would be affected.

We are, therefore, going to focus on the comparison of Erlang C and Erlang A models and an assessment of their influence on the number of agents in a call centre. As mentioned in the previous chapter, Erlang B model is not convenient for call centres, as it does not include the queue.

When designing a call centre, it must be decided what parameters should the call centre meet. Both models are going to be used to calculate the number of agents for two different hypothetic companies - one small (with a small number of agents) and one large (more agents) so that the differences of the two models in different situations are clearly demonstrated.

The requests for call centres of the two companies are in Tab. 1. As shown, only the numbers of incoming calls differ, other parameters remain unchanged so that it is possible to compare the results.

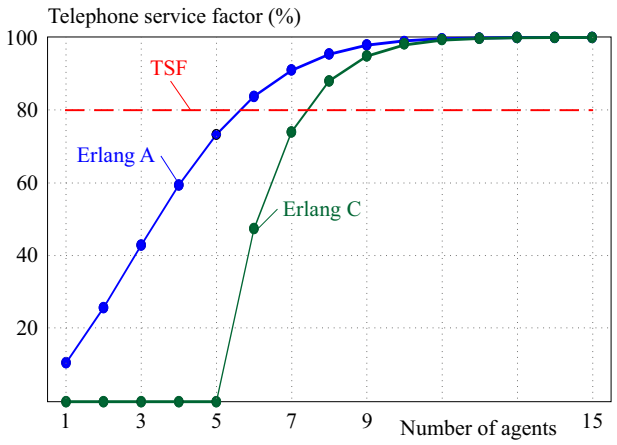

Fig. 6. Dependence of the telephone service factor on the number of agents

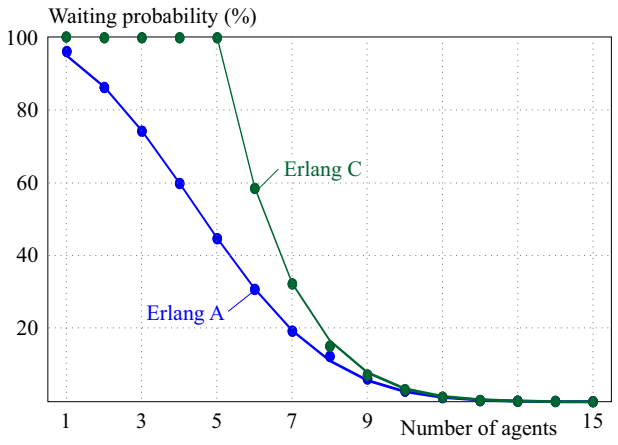

Fig. 7. Dependence of the probability of waiting on the number of agents

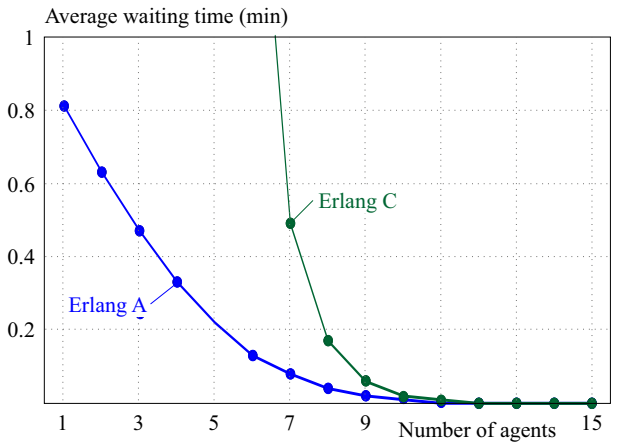

Fig. 8. Dependence of the waiting time on the number of agents

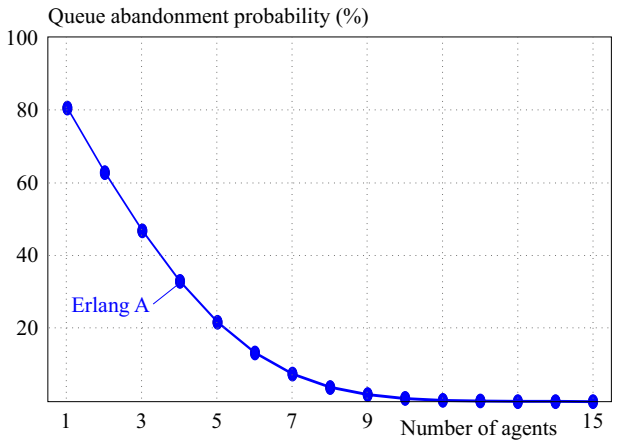

Fig. 9. Dependence of the queue abandonment on the number of agents

3.1 Assessment of several agents

The telephone service factor is one of the most important parameters of the call centre. It determines the percentage of requests serviced in each time. In this case, 80 % of customers must be assisted within 20 seconds. The number of agents must be adjusted to that so that the call centre meets this parameter.

3.1.1 Small company

Results of the calculations for a small company after filling the equations from the previous chapter are shown in Fig. 6.

The dashed line in the graphic marks the minimal telephone service factor that must be maintained in the call centre. As shown in the picture, when Erlang C model is used to design a call centre, a minimum of 8 agents are needed to maintain the telephone service factor. In the case of Erlang A, 6 agents are needed. For that reason, Erlang A model is more tolerant and economical, so it is done with fewer agents.

The telephone service factor, however, is not the only parameter to which a call centre needs to be adjusted. From the point of view of the customer, it is important if or how long waiting time in the queue is needed until the service is provided.

The probability that a customer would have to wait depending on the number of agents is demonstrated in Fig. 7.

We already know how many agents are needed for the call centre to work efficiently. We also know what the probability is of a customer being added to the queue. The average waiting time for customers still needs to be figured out. The dependence of the average waiting time on the number of agents is shown in Fig. 8.

As demonstrated in the graphic, when employing Erlang A model, the average time spent in a queue is shorter than in case the Erlang C model is utilised. Consequently, if we use Erlang A lowers the criteria. If the centre is constructed according to this model, it will work worse, slower, with longer waiting. Instability of the queue in the Erlang C model, described in chapter 2.2.2, is also shown in the graphic. At this state the requests arrive at the queue faster than the agents are responding them and,

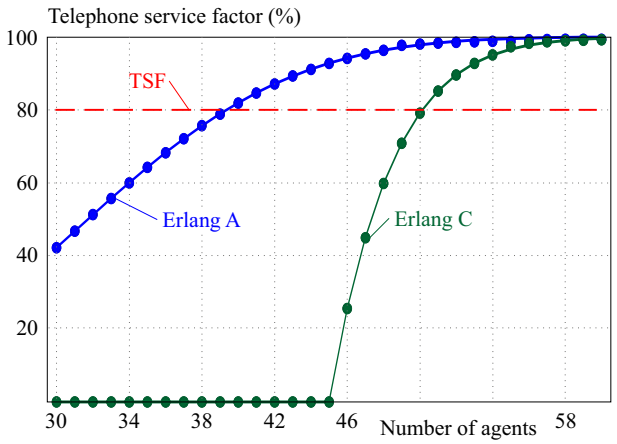

Fig. 10. Dependence of TSF on the number of agents in a large company

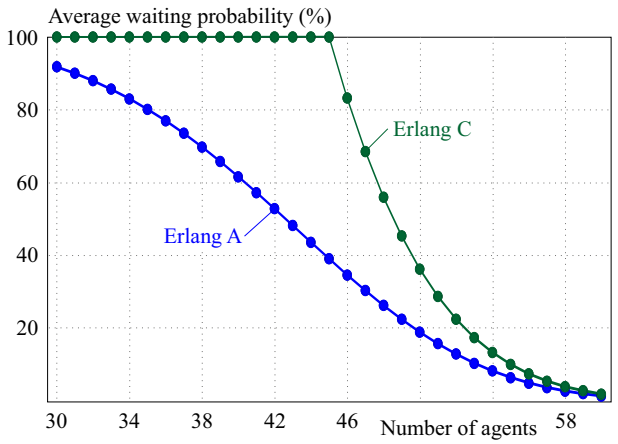

Fig. 11. Dependence of waiting probability on the number of agents

Table 2. Comparison of the results for a small and a large company

	small company		large company	
	Erlang C	Erlang A	Erlang C	Erlang A
Number of agents	8	6	51	40
Telephone service factor	88 %	83 %	85 %	82 %
Probability of waiting	17 %	31 %	29 %	62 %
Average waiting time	10.2 s	7.8 s	8.4 s	9 s
Probability of queue abandonment	–	14 %	–	15 %

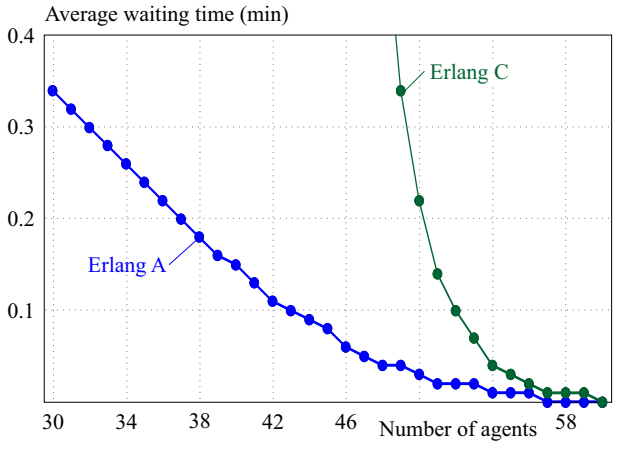

Fig. 12. Dependence of the waiting time on the number of agents

consequently, the queue grows to infinity (orange line in the graphic).

In the case of Erlang A model, it is useful to indicate the percentage of requests that leave the queue before they are serviced. This parameter helps to conclude how many customers leave the queue without being served (potentially unsatisfied customers).

The probability of queue abandonment is shown in Fig. 9. Given the minimum number of agents 6, the probability of queue abandonment is 13 %. It then depends on the consideration of the manager whether the value is acceptable. Adding one agent, the probability decreases to less than 8 % [13, 14].

The graphic also shows the instability of the queue of Erlang C model, which is more closely described in chapter 2.2.2. In this state the requests enter the queue faster than they are being serviced. Therefore, the queue then grows to infinity (orange line in the graphic).

3.1.2 Large company

The results of calculations after utilising the equations from the previous chapter for a large company are shown in Fig. 10.

The red line marks the minimum required value of the telephone service factor. As shown in the graphic, to achieve TSF 80%, a minimum of 51 agents are required in the case of Erlang C model, and a minimum of 40 agents in the case of Erlang A model.

The following graphic indicates the probability that the customer would have to wait (Fig. 11).

An interesting number of agents is for example 45. With that number of agents, a call centre designed according to Erlang C model is no longer able to serve customers and the probability of waiting is 100 %. Thus, every customer must wait in the queue. A call centre designed by Erlang A model, on the other hand, still operates with the same number of agents and the probability of waiting is only 39 %.

Figure 12 shows the average waiting time in the queue. The graphic indicates that given the same number of agents in a call centre designed according to Erlang A

model, the customers wait notably shorter. Apart from that, it also shows the instability of the queue, which grows without border in Erlang C model.

3.1.3 Result comparison

Comparison of the results for Erlang C and Erlang A models is in Tab. 2. In accordance with the results from the table and the graphics, it is possible to claim that Erlang A model is mostly suitable for large call centres in which the agents are more occupied. Whereas in a small company the number of agents varies only by 2, in a large company it is as much as 11, which might considerably reduce the operation costs of the call centre.

It can also be observed that the probability that customers would have to wait is approximately twice as high in a call centre designed by Erlang A model. The waiting time is equal in both cases.

The last row of the table shows the probability that a customer would abandon the queue early, before the service is provided. For Erlang C model this information is unavailable because the model does not consider that a request could abandon the queue before being serviced. In the case of Erlang A model, the value fluctuates below 15 % and represents customers that had to spend more than 1 minute (parameter given in the beginning of the calculation) in the queue.

We managed to detect the number of agents in a call centre and the following step is ensuring that the hardware of the call centre can support the number of calls.

4 Conclusion

The first part describes the parameters that are important for designing and operation of a call centre. It then compares two Erlang models and their effect on the number of agents in the call centre.

The results indicate that Erlang C model is more exacting and more rigorous. The number of agents is higher when utilising this model than when Erlang A model is employed. This variation is caused by the fundamental disadvantage of Erlang C model, namely that requests waiting to be serviced in the queue cannot abandon it. However, not all calling customers remain waiting until they are served but part of them abandon the queue before that. That is the main reason why employing Erlang A model results in a smaller number of agents and this difference gets bigger as the call centre grows larger. Erlang A model is mostly utilised in large companies, where the agents are the most occupied. Nevertheless, the calculated number of agents should not be strictly determined, and it is necessary to adjust it to the real business operation via following the statistics of the operation, and creation of future prognoses according to them.

Acknowledgment

This article was created with the support of the Ministry of Education, Science, Research and Sport of the Slovak Republic within the KEGA agency project 007STU4/2016 Progressive educational methods in the field of telecommunications multiservice networks and VEGA agency project 1/0462/17 Modelling of qualitative parameters in IMS networks.

References

[1] I. Baroňák, "Call Center", Slovak University of Technology Bratislava, 2010, ISBN 978-80-227-3261-1.

[2] M. Hartmann, "Kontaktné centrum IP", Diploma Thesis, Slovak University of Technology Bratislava, 2014.

[3] Mandelbaum A. Sakov, and S. Zeltyn, "Empirical Analysis of a Call Center", [online], Technical report, Technion, 2001, [cit.2017-04-04], http://ie.technion.ac.il/serveng/References/Sakov_Tel Aviv_talk.ppt.

[4] Garnett A. Mandelbaum, and M. Reiman, "Designing a Telephone Call-Center with Impatient Customers",[online], Davidson Faculty of Industrial Engineering and Management, Technion, 2002, [cit.2017-04-04], http://ie.technion.ac.il/serveng/References/Garnett.pdf.

[5] Borst A. Mandelbaum, and M. Reiman, "Dimensioning Large Call Centers, Operations Research", 2004, [cit.2017-04-04], https://hal.archives-ouvertes.fr/file/index/docid/76467/filename/RR-0094.pdf.

[6] M. Sheldon, and P. Ross, "Introduction to Probability Models", Eights edition. University of California, 2003, ISBN 0-12-598055-8.

[7] F. Chamráz, I. Baroňák, "Impact of Admission Control Methods to the Traffic Management", *Advances Electrical and Electronic Engineering*, vol. 13, no. 4, 2015, ISSN: 1336-1376.

[8] Robbins D. Medeiros, and P. Harrison, "Evaluating the Erlang C and Erlang A Models for Call Center Modeling" [online], East Carolina University, 2012, [cit.2017-04-04] http://myweb.ecu.edu/ROBBINST/PDFs/Erlang%20Compare%20Working%20paper.pdf.

[9] Robbins D. Medeiros, P. Harrison, "Does the Erlang C Model Fit Real Call Centers?" [online], Winter Simulation Conference, 2010, [cit.2017-04-04] http://www.informs-sim.org/wsc10papers/264.pdf.

[10] Baccelli G. Hebuterne, "On Queues with Impatient Customers", [online], North-Holland Publishing Company, 2006, [cit.2017-04-04] https://hal.archives-ouvertes.fr/file/index/docid/76467/filename/RR-0094.pdf.

[11] Mandelbaum S. Zeltyn, "The Palm/Erlang-A Queue, with Applications to Call Centers" [online], Faculty of Industrial Engineering & Management, Technion, 2005.[cit.2017-04-04] http://ie.technion.ac.il/serveng/References/Erlang_A.pdf.

[12] Chan, "CCOptim: Call Center Optimization Java Library", [online], Université de Montréal, 2014. [cit.2017-04-04] http://www-labs.iro. umontreal.ca/ chanwyea/.

[13] Mandelbaum and N. Shimkin, "A Model for Rational Abandonment from Invisible Queues", [online], Queueing Systems: Theory and Applications, 2000. [cit.2017-04-04] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.25.4108&rep=rep1&type=pdf.

[14] Mandelbaum and S. Zeltyn, "The Impact of Customers Patience on Delay and Abandonment: Some Empirically-Driven Experiments with the M/M/N+G Queue", [online], Faculty of Industrial Engineering & Management, Technion, 2004, [cit.2017-04-04] http://ie.technion.ac.il/serveng/References/pat_impact_color.pdf.