MARKOV MODEL M/M/M/K IN CONTACT CENTER

Erik CHROMY[1], *Jan DIEZKA*[1], *Matej KAVACKY*[1]

[1]Institute of Telecommunications, Faculty of Electrical Engineering and Information Technology,
Slovak University of Technology Bratislava, Ilkovicova 3, 812 19 Bratislava, Slovak Republic

chromy@ut.fei.stuba.sk, jan.diezka@gmail.com, kavacky@ut.fei.stuba.sk

Abstract. *Our paper deals with a contact center dimensioning. The main task of a contact center is to offer services to customers. The most simple and genuine way of communication with the contact center agent is by phone. From the customer satisfaction point of view, the shortest serving time is the most important factor. It is known that the highest operating costs of the contact center are costs for agents. Therefore, we need to choose the right number of agents and effectively utilize them. For contact center dimensioning we can use various models. In this paper, we introduce the Markov M/M/m/K model which offers the broad range of parameters suitable for the contact center sizing, e.g. required number of agents based on the probability of call refusal and probability of call enqueue.*

Keywords

Contact center, M/M/m/K, queuing system.

1. Introduction

The contact center belongs to convergent technologies and allows companies to provide services on four basic platforms: voice, data, video and web.

The processing of customer's phone queries by agents of the contact center is the basic contact center service. Utilization of the self-service Interactive Voice Response (IVR) system is tightly coupled with processing of a telephone call. For the correct operation of the contact center it is necessary to ensure:

- easy access,
- one number,
- access through other media (Internet, GSM, WAP),
- specialized numbers (for VIP clients or important products),
- 24/7 availability.

The availability 24 hours and 7 days per week needn't mean permanent occupancy of contact center by agents. During light traffic (e.g. in the night or small hours), it can be satisfactory to operate only self-service parts of contact center.

The paper is divided as follows. The chapter 2 describes basic principles of the contact center operation and summarizes the important traffic parameters with impact on contact center design. The chapter 3 introduces the contact center as a queuing system. Than the modeling of the contact center through the Markov model M/M/m/K is presented. Calculations and relations of important traffic parameters of the contact center are given. The final part of the paper deals with conclusions on Markov M/M/m/K model.

2. Traffic Parameters in Contact Center

The operation of contact center is based on the following principle. Customers make phone calls to the contact center with various queries. The first contact with calling customer is through IVR system. This system identifies the customer and it is capable of automatically responding to some queries from customers. If the help of an agent is needed, the IVR can reroute the call to the Automatic Call Distribution module (ACD) which connects the calls with the most appropriate agents through special routing algorithms. In the case of more calling customers than available agents the waiting queue will occur. The management of waiting queues is also the task of ACD module. At the end, the contact center agent handles the particular query.

It is obvious that customer calls are random, some of them will stay on the line through the whole process (contact with IVR, reroute by ACD, waiting in the queue), some will leave the line after contact with IVR, or during waiting in the queue.

Detailed information about contact center operation can be obtained by monitoring of various traffic parameters. These traffic parameters can be divided into two groups according to their characteristics. They can describe the contact center or grade of Quality of Service (QoS) [1], [2], [3], [4], [5], [6], therefore, the quality of services offered by given contact center. The essential traffic parameters of contact center are:

- traffic load of the contact center – A [erl],
- number of agents – m,
- average rate of incoming calls per hour – λ [calls/h],
- average rate of served calls per hour – μ [calls/h],
- average length of call handling – T_c [s] – and holds that $T_c = 1/\mu$.

The parameters listed above will form hereinafter the basic input values for traffic modeling in the contact center.

3. The Contact Center as a Queuing System

A queuing system is such system in which given number of service stations handle a large number of queries. Also, contact center is such system, the queries are generated by calling customers and service stations are represented by contact center agents.

In general, one agent can handle one query at the same time. There can be also situation, when the agents can’t handle all incoming call immediately. In this case, the customers may leave a system or wait for an available agent in the waiting queue. Of course, instant hang-ups are unwanted, therefore, the proper management of the waiting queue is one of the main task in the contact center.

According to the queuing theory, the properties and behavior of the queuing system is described by five parameters:

- the way of arrival of inputs into the system,
- the way of processing of queries,
- the number of service stations,
- total capacity of the system,
- largeness of population of queries for this system.

Large number of queuing system exists and the most appropriate queuing system should be used for solving of a particular problem. According to Kendall classifying of queuing systems the system can be marked as *A/B/c/N/K*, while [7]:

- A – probability distribution of queries arrivals (inputs) into system,
- B – probability distribution of handling time required for query processing,
- c – the number of service stations (agents),
- N – capacity of the system (number of customers, queries, etc.),
- K – restriction of waiting the queue length.

For the denotation of probability distributions of queries arrivals into the queuing system and handling time of queries (A and B in Kendall classifying) the following symbols are used [7]:

- M – exponential distribution,
- E – Erlang distribution,
- D – deterministic distribution (time between arrivals of queries or handling time is constant),
- G – general distribution,
- G1 – General distribution with independent interarrival times.

The way of waiting queue management is the substantial problem in quality of service provision in the queuing system. The administrator of queuing system has to ensure no query will stay in waiting queue too long time. From this, we can derive the number of service stations (agents, servers, etc.) and capacity of waiting queue.

In high-speed networks, there is considerable interest in traffic arrival processes where successive arrivals are correlated. Such non-G1 arrival processes include the Markov modulated Poisson process (MMPP) [7].

3.1. Basic Probability Distributions

Call arrival into the contact center is in the most cases modeled by Poisson probability distribution.

Stochastic variable X has Poisson distribution with the mean λ, if set of their possible values is $H(X) = \{0, 1, 2, \ldots, n, \ldots\}$ and the probability density function [8] is given by:

$$f(k) = e^{-\lambda}\frac{\lambda^k}{k!}, \quad \text{for } k \in \{0,1,2,\ldots,n,\ldots\}. \tag{1}$$

Another probability distribution often used in the contact center modeling is an Exponential distribution. Variable X has an exponential distribution if the probability density function [8] is given by:

$$f(x) = \mu e^{-\mu x}, \quad \text{for } x > 0. \tag{2}$$

Phase-type distribution - a method of Markovizing a non-Markovian model and as such has a wide applicability [7].

4. Modeling of Contact Center

Number of agents has a direct impact on QoS of the contact center. On the other hand, the high number of agents has a negative impact on operational costs of contact center.

From the contact center operational costs analysis results that the significant part of costs contains costs for human sources, e.g. wages of agents, managers and supervisors of contact center. These costs are monthly repeated therefore, the proper design of the required number of agents is the one of the key tasks in the contact center design. When determining the required number of agents the following three basic assumptions valid in the majority of contact centers should be taken into consideration [9]:

- agent has to handle call immediately (delayed response decreases QoS),
- call arrivals are not periodical and their number always vary (according to daytime, day of the week, etc.),
- there is no linear relation between number of incoming calls and the required number of agents in order to ensure QoS.

There are various models of contact center applicable to the determination of the required number of agents or calculations of important traffic parameters, from basic Erlang models B and C [10], through Markov queuing system models to complex non-Markov models and simulations of the contact center.

4.1. Markov M/M/m/K Model

The biggest difference of *M/M/m/K* model [11], [12] against *M/M/m/∞* model is a limited length of waiting queue. With such limitation, this model is significantly approaching the real situation in the contact center which is of limited capacity. The principle of Markov *M/M/m/K* model is depicted in Fig. 1.

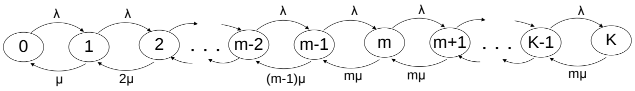

Fig. 1: Markov M/M/m/K model.

Queries arriving into the system with rate $\bar{\lambda}$, servers handles particular queries with rate $m\mu$, and maximum K queries can be in the system at the same time. All other incoming queries are refused.

The customers call the contact center randomly, with the Poisson distribution with mean λ. Therefore, there are λ calls per hour arriving into the contact center. There are m agents in the contact center and each of them serves the calling customers with rate μ customers per hour. Handling time has an exponential distribution with rate μ. There can be maximum of K customers in the contact center, while the maximum of m customers can be served simultaneously. All other customers have to wait in the waiting queue until there is any free agent. The parameter K therefore significantly limits the capacity L of waiting queue and the capacity of waiting queue L is given by equation (3):

$$L = K - m. \tag{3}$$

In the case of full occupancy of the contact center (K customers are in the contact center) the each other incoming call is blocked. Probability of call blocking is P_B. Probability of successful call into the contact center is 1 - P_B. The customers are served by agents with rate $m\mu$, while the rate in which the agents are occupied by customers is given by the value $\bar{\lambda}$:

$$\bar{\lambda} = \lambda(1 - P_B). \tag{4}$$

4.2. Calculation Capacity of Markov M/M/m/K Model

The contact center modeled by Markov *M/M/m/K* model is stable in every traffic load, because in the case of full occupancy each other call is blocked. The parameter ρ (5) is therefore used for determination of traffic load per server, or per agent [11]:

$$\rho = \frac{\lambda}{m\mu}. \tag{5}$$

Probability that the system is empty [11] (there is not any query) is given by:

$$P_0 = \left[\sum_{i=0}^{m-1}\frac{m^i}{i!}\rho^i + \frac{m^m}{m!}\rho^m\frac{1-\rho^{K-m+1}}{1-\rho}\right]^{-1},$$

$$\text{for } \rho \neq 1, \tag{6}$$

and

$$P_0 = \left[\sum_{i=0}^{m-1}\frac{m^i}{i!} + \frac{m^m}{m!}(K-m+1)\right]^{-1},$$

$$\text{for } \rho = 1. \tag{7}$$

Probability that in the system there are just n queries [11] is given by:

$$P(n) = \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n P_0 = \frac{m^n}{n!}\rho^n P_0,$$

$$\text{for } n = 0,1,\ldots,m, \tag{8}$$

and

$$P(n) = \left(\frac{\lambda}{m\mu}\right)^{n-m} P_m = \frac{m^n}{m!}\rho^n P_0,$$

$$\text{for } n = m, m+1,\ldots,K, \tag{9}$$

where P_m is the probability that just m queries are in the system.

In the case that just K queries are in the system each other incoming call is blocked. Probability *P(K)* specifies the probability of call blocking P_B, and from equation (7) we have:

$$P_B = \frac{m^n}{m!} \rho^K P_0 . \tag{10}$$

From the definition of the Markov *M/M/m/K* system results that in the case of *m* calls in the contact center, each other incoming call is rerouted to the waiting queue until there are *K* queries in the contact center, then each other incoming call is blocked. Probability of call enqueue P_Q can be determined by use of theorem of full probability (11):

$$P = \sum_{i}^{\infty} p_i \leq 1 . \tag{11}$$

By use of theorem (11) it is possible to construct an equation for P_Q (12):

$$P_Q = 1 - P_0 - \sum_{n=1}^{m-1} P(n) . \tag{12}$$

5. Calculation of Important Traffic Parameters of Contact Center

Calculation of particular parameters in the analysis of the contact center by Markov *M/M/m/K* model is based on the following reference values:

- λ = 60 incoming calls per hour,
- service time is T_{serv} = 5 minutes,
- the maximal number of queries in the system K = 15.

Thus, the average traffic load of such contact center is 5 *erl*.

5.1. Calculation of the Required Number of Agents

The probability of call blocking P_B can be used as QoS parameter for determination of the required number of agents in the contact center modeled by the Markov *M/M/m/K* model. In Fig. 2 we can see that this probability is sheer decreasing with each new agent. This decreasing stops at the number of agents m = 6, when the P_B is about 2 %, what is an acceptable value. Exact values are stated in Tab. 1.

It must be noted that with each additional agent also the capacity L of the waiting queue vary according to equation (3). This property comes from the definition of *M/M/m/K* model and in all calculations of traffic parameters it is necessary to take into account the variation of the waiting queue capacity *K*.

In Tab. 1 there is also parameter $\bar{\lambda}$. From Tab. 1 we can see that for example in the case of 6 agents and reference value of traffic load of contact center 5 erl the call arrivals rate is λ = 60 calls/h. However, agents of the contact center are occupied with rate 58,74 calls/h. Other calls are waiting in the queue, or are blocked. Other calls are waiting in the queue or are blocked.

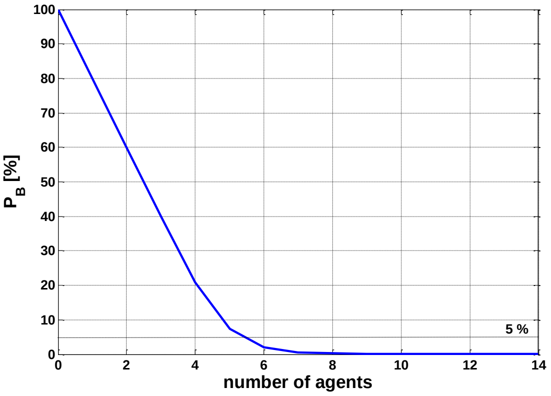

Fig. 2: Relation between P_B [%] and number of agents m when A = *5 erl* and K = *15*.

Tab.1: Traffic parameters if A = *5 erl*.

m	P_B [%]	P_Q [%]	L	ρ	$\bar{\lambda}$
5	7,40	81,42	10	1,00	55,56
6	2,10	54,43	9	0,83	58,74
7	0,64	31,34	8	0,71	59,62
8	0,23	16,40	7	0,63	59,86
9	0,11	7,93	6	0,56	59,94

5.2. Traffic Parameters and Traffic Load Variation

In Fig. 3 we can see the relation between probability of call blocking P_B [%] and traffic load A [*erl*] in the case of 6 agents. When the value of traffic load of 7 erl is exceeded, the probability of call blocking is above 5 %. Traffic load of 7 erl at the average call handling time T_{obs} = 5 min represents average of 84 incoming calls per hour.

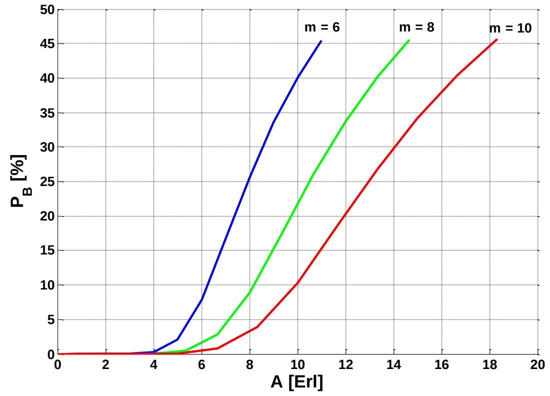

Fig. 3: Relation between P_B [%] and A [erl].

Figure 3 shows the cases when the traffic load is increasing and the number of agents is 8 or 10.

The impact of increasing contact center traffic load *A* on probability of call enqueue P_Q can be seen in Fig. 4. Also, cases with increasing traffic load and 8 or 10 agents are depicted in Fig. 4.

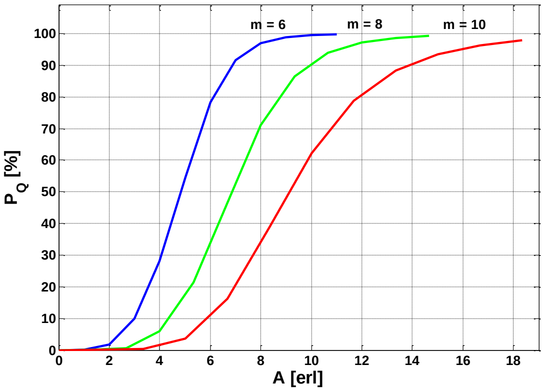

Fig. 4: Relation between P_Q [%] and *A* [erl].

Tab.2: Traffic parameters if *m* = 6.

A [erl]	P_B [%]	P_Q [%]	ρ	$\bar{\lambda}$
3	0,01	9,91	0,50	36,00
4	0,25	28,12	0,67	47,88
5	2,10	54,43	0,83	58,74
6	7,83	78,28	1,00	66,36
7	16,65	91,61	1,17	70,01
8	25,69	96,98	1,33	71,34
9	33,55	98,89	1,50	71,77

Table 2 includes all interesting traffic parameters for 6 agents.

6. Conclusion

The contribution of this paper is the application of M/M/m/K model on designed contact center model. This model is possible to use for dimensioning of contact centers and prediction of their traffic parameters. By restriction of waiting queue capacity this model is approaching the situation in the contact center. The calculation capacity of the model allows the calculations of the most of the important traffic parameters of the contact center.

By use of Markov *M/M/m/K* model the following QoS parameters of contact center have been determined:

- *m* – the number of agents in the contact center,
- P_B – the probability that customer will be rejected,
- P_Q – the probability that customer will have to wait in the waiting queue.

The model *M/M/m/K* does not take into account the situation when the customer will hang up after he is enqueued, so the call is lost.

The model of the contact center can be extended with parameters describing the waiting queue such as:

- capacity (length) of the waiting queue *L*,
- number of all calls in the contact center (served and waited) *N*,
- average number of customers in the waiting queue *Q*,
- average time that the query will spent in the system *T* [s] (waiting time in the queue and handling time by agent),
- average time that the query will spent in the waiting queue *W* [s].

So the interesting results can be obtained by adding of calculations of above parameters which can better monitor the QoS level in the contact center. Based on this it is possible to respond to other situations that will occur in the contact center.

Acknowledgements

This work is a part of research activities conducted at Slovak University of Technology Bratislava, Faculty of Electrical Engineering and Information Technology, Institute of Telecommunications, within the scope of the projects Grant programme to support young researchers of STU - Modeling of Traffic in NGN Networks and Support of Center of Excellence for SMART Technologies, Systems and Services II., ITMS 26240120029, co - funded by the ERDF.

References

[1] VOZNAK, M., A. KOVAC and M. HALAS. Effective Packet Loss Estimation on VoIP Jitter Buffer. In: *Networking 2012 Workshops*. Prague: Springer, 2012, pp. 157-162. ISBN 978-3-642-30038-7. DOI: 10.1007/978-3-642-30039-4_21.

[2] KYRBASHOV, B., I. BARONAK, M. KOVACIK and V. JANATA. Evaluation and Investigation of the Delay in VoIP Networks. *Radioengineering*. 2011, vol. 20, no. 2, pp. 540-547. ISSN 1210-2512.

[3] MICUCH, J. and I. BARONAK. Implementation Admission Control Methods for VoIP Applications. In: *Telecommunications and Signal Processing TSP-2010: 33rd International Conference on Telecommunications and Signal Processing*. Vienna: Budapest: Asszisztencia Szervezo Kft., 2010, pp. 391–395. ISBN 978-963-88981-0-4.

[4] POLACEK, P. and I. BARONAK. Enhanced Equivalent Capacity Method. In: *Proceedings of Informatics 2009: IADIS Multi Conference on Computer Science and Information Systems*. Algarve: IADIS Press, 2009, pp. 192-196. ISBN 978-972-8924-86-7.

[5] VOZNAK, M., M. HALAS, B. BOROWIK and Z. KOCUR. Delay Model of RTP Flows in Accordance with M/D/1 and M/D/2 Kendall's Notation. *International Journal of Mathematics and Computers in Simulation*. 2011, vol. 3, iss. 3, pp. 242-249. ISSN 1998-0159.

[6] VOZNAK, M., F. REZAC and M. HALAS. Speech Quality Evaluation in IPsec Environment. In: *Proceedings of the 12th Inter. Conference on Networking, VLSI and Signal Processing: ICNVS 2010*. Cambridge: University of Cambridge, 2010, pp. 49-53. ISBN 978-960-474-162-5. ISSN 1790-5117.

[7] BOLCH, G., S. GREINER, H. MEER and K. S. TRIVEDI. *Queuing Networks and Markov Chains*. 2nd ed. New York: John Wiley & Sons, 2006. ISBN 978-0-471-56525-3.

[8] VOLAUF, Peter. *Numerical and statistical calculations in MATLAB*. Bratislava, 2005. Diploma thesis. Slovak Technical University-STU.

[9] HISHINUMA, Ch., M. KANAKUBO and T. GOTO. An Agent Scheduling Optimization for Call Center. In: *APSCC '07 Proceedings of the 2nd IEEE Asia-Pacific Services Computing Conference*. Tsukuba Science City: IEEE, 2007, pp. 423 – 430. ISBN 0-7695-3051-6. DOI: 10.1109/APSCC.2007.27.

[10] DIAGNOSTIC STRATEGIES. *Traffic Modeling and Resource Allocation in Call Centers*. Needham, MA, 2003. Avaible at: www.fer.hr/_download/repository/A4_1Traffic_Modeling.pdf.

[11] BLANC, J. P. C. TILBURG UNIVERSITY. *Queueing Models: Analytical and Numerical Methods*. Januar 2011.

[12] STOLLETZ, Raik. *Performance Analysis and Optimization of Inbound Call Centers*. Berlin: Springer-Verlag, 2003. ISSN 0075-8450. ISBN 3-540-00812-8.

About Authors

Erik CHROMY was born in Velky Krtis, Slovakia, in 1981. He received the Master degree in telecommunications in 2005 from Faculty of Electrical Engineering and Information Technology of Slovak University of Technology Bratislava. In 2007 he submitted Ph.D. work from the field of Observation of statistical properties of input flow of traffic sources on virtual paths dimensioning and his scientific research is focused on optimizing of processes in convergent networks. Nowadays he works as assistant professor at the Department of Telecommunications of Faculty of Electrical Engineering and Information Technology of Slovak University of Technology Bratislava Bratislava.

Jan DIEZKA was born in Dolny Kubin, Slovakia in 1988. He is a student at the Institute of Telecommunications, Faculty of Electrical Engineering and Information Technology of Slovak University of Technology Bratislava. He focuses on application of Erlangs' formulas in Contact Centers.

Matej KAVACKY was born in Nitra, Slovakia, in 1979. He received the Master degree in telecommunications in 2004 from Faculty of Electrical Engineering and Information Technology of Slovak University of Technology Bratislava. In 2006 he submitted Ph.D. work "Quality of Service in Broadband Networks". Nowadays he works as assistant professor at the Institute of Telecommunications of Faculty of Electrical Engineering and Information Technology of Slovak University of Technology Bratislava and his scientific research is focused on the field of quality of service and private telecommunication networks.