Prediction of traffic in the Contact Centers

Tibor Mišuth, Erik Chromý, Matej Kavacký

Department of Telecommunications, Faculty of Electrical Engineering and Information Technology, Slovak University of Technology, Ilkovicova 3, 812 19 Bratislava, Slovakia
tibor.misuth@gmail.com, erik.chromy@stuba.sk, matej.kavacky@stuba.sk

Abstract

Our paper deals with description of the Contact center parts by mathematical models and prediction of traffic in the Contact centers. This knowledge is useful issue during design of the Contact center. The model of the Contact center is one of the options how we can estimate the traffic and important parameters of Contact center. In our work we deal with the description of the Contact center components by use of Erlang formulas and Markov models.

1. Introduction

The public contact today is the must for almost every enterprise or organization. Availability of simple, fast and comfortable contact with the clients, customers or trade partners is the crucial need in the competitors fight.

Contact center is one of many other ways how the institution is able to provide all its communication requirements towards clients and partners. Contact center is structured communication system consisting of human and technological resources, which improves the communication between organization and customers. It is software and hardware upgrade of private branch exchange which is primary medium for call access to organization. It combines telephone processes and data processing to achieve the most effective business transactions.

At the heart of Contact center there is Automatic Call Distribution (ACD) module. ACD is responsible for the correct and efficient routing of all types of requests (voice calls, e-mails, etc.) from their arrival to the Contact center until they are served.

2. Contact Center Model

Each component of ACD system can be more or less precisely converted to a mathematical model. Since the ACD systems usually handle large amount of requests per time unit, the most of these models are based on statistical principles. The accuracy of results lies in correct model selection. Also the precise description of input parameters and variable dependencies can significantly affect them as well.

2.1. Model of input traffic

The rate of arriving calls (or other requests) per time unit can be considered as random variable. Contact center's goal is to provide its services to wide range of customers therefore we can assume that the requests population is unlimited. Once the population is unlimited, we can regard the ACD system as queuing system with unlimited requests population [1], [2]. Each request arriving to Contact center originates randomly and independently from each other requests. The events (calls) are independent from the mathematical point of view if for any arbitrary set of requests A_1, A_2, ... A_n and arrival probabilities in selected time interval $P(A_1)$, $P(A_2)$, ..., $P(A_n)$ we have [3]:

$$P(\bigcap_{i=1}^{n} A_i) = \prod_{i=1}^{n} P(A_i). \tag{1}$$

So the arrival probability of n requests during the same time interval equals to conjunction of individual requests arrival probabilities for given interval.

It is obvious that the arrival rate of calls can obtain only limited number of values, so it is a discrete variable. Any random variable can be described by the probability density function. Discrete random variable X with values x can be described by the formula:

$$f(x) = P(X = x). \tag{2}$$

It is obvious that $\sum_{x \in H} f(x) = 1$, where H is the set of all values the random variable X can obtain.

The mean and variance are also important characteristic of random variables. Let random variable X has values $x_1, \ldots, x_n$ with corresponding probabilities $p_1, \ldots, p_n$. Then mean is defined as:

$$E(X) = \sum_{i=1}^{n} x_i p_i \,, \tag{3}$$

and variance as:

$$D(X) = \sum_{i=1}^{n} \left(x_i - E(X)\right)^2 p_i \,. \tag{4}$$

Incoming requests into queuing system are mostly modeled by Poisson distribution [2], [3]. Following formulas define its probability density function, mean and variance

$$p_k = f(k) = P(X = k) = \frac{\lambda}{k!} e^{-\lambda}, \tag{5}$$

$$E(X) = \lambda, \tag{6}$$

$$D(X) = \lambda, \tag{7}$$

where λ represents the average number of requests per time unit.

The inter-arrival time is also very important variable as well. Since the arrival rate is random variable the inter-arrival time must be random variable as well. Based on the independence of each request the inter-arrival time for Poisson traffic flow is exponentially distributed continuous random variable with the mean equals to $1/\lambda$ [4]. The probability density of this random variable is defined by the following formula [2], [3]:

$$f(t) = \lambda e^{-\lambda t}, \tag{8}$$

where t represents the inter-arrival time and λ represents the average number of requests per time unit.

The random variable with Poisson distribution is the most frequently used case how to model the input traffic.

2.2. Model of Interactive Voice Response

Interactive voice response (IVR) module collects and provides information to customers using self-service information services. It can answer questions, connect calls, or proceed the whole transaction. The caller selects from options by telephone keyboard, or by voice if system allows it.

The graph theory can significantly help us to create a model for IVR system. Oriented graph $G = (V, E)$ is defined by the set of peaks (nodes) V and set of edges E that interconnect pairs of peaks [5]. The cases when caller inputs some information or his request is processed and he obtains the information we can model as peaks of graph in our model. User's choices in particular peaks represent the edges of the graph. Each edge has assigned a value representing the probability of choosing the particular edge to get to the next peak. It is obvious that the sum of these values for all edges from one peak equals to 1.

2.3. Model of Service group

The service group consists of a queue and at least one agent. The agent is a server where the requests are processed from the queuing theory point of view. In case if there are more requests at the same time, the later coming requests has to wait in the queue.

Danish mathematician A. K. Erlang is very closely tied to the origin of queuing theory [1]. He published the paper concerning application of statistics in telephone service in 1909. This document begins further research of queuing theory. Together with Markov chains theory his ideas helped to define and describe more complicated queuing models.

In 1953 D. G. Kendall introduced dedicated queuing system categorizing scheme. It is based on combination of letters and numbers in pattern [1], [2] (A/B/X/Y), where A represents the inter-arrival time distribution, B defines the handling time distribution, X denotes the number of servers and Y defines the maximum system capacity. Comparison of theory with real systems has shown that the best distribution to describe the inter-arrival time is exponential distribution denoted as M. Therefore our research is based on M/M/X/Y models.

Both Erlang and M/M/X/Y models are based on the same assumption and following requirements must be met [6]:

- number of source (requests population) is much more greater than the number of servers,
- the requests are generated randomly and independently from each other,
- the average number of requests per time unit from all sources is constant,
- the handling time is random variable with exponential distribution,
- queuing is based on the First in First out (FIFO) algorithm.

Several metrics can be used to evaluate the Contact center service. The following parameters belong to the most important and observed [1], [2], [6], [7], [8]:

- *GoS (Grade of Service)* – denotes the percentage of requests served (assigned to agent) before defined threshold AWT (Acceptable Waiting Time),
- *ASA (Average Speed of Answer)* – defines the average waiting time the request spends in the queue before it is assigned to the agent (in M/M/X/Y models named as W),
- *average queue length* – denotes the average number of requests present in the queue at any moment (in M/M/X/Y models named as Q),
- *average number of requests in the system* – this parameter informs about Access lines utilization (in M/M/X/Y systems named as K),
- *average time spent in the system* – shows the average time the request spends in system from entrance to the queue until it is served and exits the system (in M/M/X/Y named as T),
- *agent utilization* – informs about percentage of time used by agents to handle the requests,
- *queuing probability* – probability the request entering the system is not assigned to a free agent immediately but has to wait in the queue,
- *blocking probability* – probability the request is not served at all due to some reason (insufficient capacity of lines, maximum waiting time exceeded, etc.),
- *AHT (Average Handling Time)* – denotes the average time the agent is occupied by handling one request including the time spent for „paper-work" before and after the request handling.

The first two metrics are the mostly used to monitor the Contact center service level. They can describe almost all situations that can occur in Contact center.

2.4. Erlang B model - model M/M/m/m

Erlang B model is the basic model which does not contain the queue. Incoming calls are assigned to the idle agent directly if there is any, otherwise they are blocked and will receive busy signal [6]. An analogy to this behavior is the usage of all trunks that interconnects Contact center to PSTN (Public Switched Telephone Network) or any other communication network. Once there is no more available lines, the requests can not be put through to the Contact center and they are blocked (rejected). This implies the Erlang B model is widely used to dimension the trunk capacity between Contact center and communication networks. Today, the Voice over IP (VoIP) technology is more and more important, but the basic capacity problem is only slightly modified to available data throughput of the connection. Thus Erlang B model can be used in this case as well.

The Erlang equation uses three basic parameters:

A – the traffic load in Erlang,

N – number of lines / trunks (requested simultaneous connections),

P_B – probability of call blocking.

The original form of the equation allows us to find the blocking probability if A and N values are known:

$$P_B(N,A) = \frac{\frac{A^N}{N!}}{\sum_{i=0}^{N}\frac{A^i}{i!}}. \quad (9)$$

If we know the rate of calls per time unit λ and the average number of served requests per the same time unit μ (so the average handling time is $1/\mu$) then the traffic load can be easily evaluated as [7]:

$$A = \frac{\lambda}{\mu}. \quad (10)$$

If we substitute A in equation (9) we receive following form:

$$P_B(N,\lambda,\mu) = \frac{\left(\frac{\lambda}{\mu}\right)^N}{N!}\frac{1}{\sum_{i=0}^{N}\left[\left(\frac{\lambda}{\mu}\right)^i\frac{1}{i!}\right]} \quad (11)$$

We can see that it is the same formula as obtain from M/M/m/m queuing system [1], [2]. We have just analytically shown that the two mentioned models are identical in fact.

The original Erlang B model can be modified in the way that it will assume also the traffic generated by repeated attempts in the case if the previous attempt was unsuccessful (it was rejected). Then we have extended Erlang B model [9]:

The new variable is recall factor r. It denotes the probability that a blocked caller will try to call again immediately. The traffic load towards the Contact center consists of two parts: first-time attempts and repeated calls [10]:

$$A = A_0 + R = A_0 + A.P_B(A,N).r, \quad (12)$$

where A_0 is the traffic load generated by first-time attempts (in Erlang), R represents the load of repeated calls and P_B is the blocking probability using the original Erlang B model (9). Equation (12) can be easily transformed to the form:

$$A\left(1 - P_B(A,N).r\right) = A_0. \quad (13)$$

By using iterative numerical methods we can then find out the original traffic load A_0 that the Contact center can serve.

2.5. Erlang C model - model M/M/m/∞

The immediate blocking of incoming call, if there is no idle agent available (Erlang B model principle), is not the best approach how to handle requests in Contact centers. However the second Erlang model, the Erlang C model, eliminates this disadvantage. This model contains the queue, where the incoming requests can wait until an agent is available. The maximum queue length in Erlang C model is unlimited. Therefore this model is better to use, but it still has some limitations. FIFO queuing strategy is used, so when any agent becomes available, the first request from the queue is immediately assigned to him. If the queue is empty, the agent remains in available state and waits for the next incoming call, which will be assigned to him without queuing.

Erlang C model [6] is originally defined as function of two variables: the number of agents N and the traffic load A. Based on these parameters it calculates the probability P_C (14) that the arriving request is not assigned to the agent immediately and it has to wait in the queue.

$$P_C(N,A) = \frac{\frac{A^N N}{N!(N-A)}}{\sum_{i=0}^{N-1}\frac{A^i}{i!} + \frac{A^N N}{N!(N-A)}}. \quad (14)$$

The comparison of both Erlang equation (9) and (14) shows that they define the probability the system is not capable of immediate request processing. The only difference is the way, how the blocked calls are handled.

At this point we can use the equation (10) once again. Moreover we define the variable ρ, that represents the utilization of 1 agent as [1], [2], [8]:

$$\rho = \frac{\lambda}{N\mu}. \quad (15)$$

From the conjunction of equations (10), (15) and (14) we can obtain following:

$$P_C(N,\rho) = \frac{\frac{(N\rho)^N}{N!(1-\rho)}}{\sum_{i=0}^{N-1}\frac{(N\rho)^i}{i!} + \frac{(N\rho)^N}{N!(1-\rho)}} \quad (16)$$

If we start from M/M/m/∞ queuing model we can derive identical equation that will define probability that m or more requests are present in the queuing system so the new incoming request will be inserted into the queue [1], [2], [8]. This means that Markov M/M/m/∞ and Erlang C models are identical and this relationship can be easily proved analytically.

Since the Erlang C model contains the queue, there are some more parameters and variables which can be measured and more or less influenced by model inputs. From the callers point of view the most important value is the waiting time or time the request spends in the queue before it is served by an agent. This value is random variable described by probability density function [8]:

$$F_W(\tau)=\begin{cases}1-P_C & \tau=0\\ 1-P_C\cdot e^{-\mu(N-A)\tau} & \tau>0\end{cases}. \quad (17)$$

Based on this formula we can calculate the average waiting time (or average speed of answer ASA) W:

$$W=\frac{P_C}{\mu(N-A)}, \quad (18)$$

and using Little's theorem [8] and equation (10) we can obtain the average number of requests in the queue Q:

$$Q=\lambda W=\frac{\lambda}{\mu(N-A)}P_C=\frac{A}{(N-A)}P_C. \quad (19)$$

The general definition of probability density function of any statistical distribution and its properties [3] gives us an opportunity to derive *GoS* parameter value from equation (17) once the acceptable waiting time (AWT) value is known [1], [7]:

$$GoS=1-P_C\cdot e^{-\mu(N-A)AWT}. \quad (20)$$

Following equation is valid for the average number of requests K in queuing system [8]:

$$K=N\rho+\frac{\rho}{1-\rho}P_C=A+\frac{A}{(N-A)}P_C=A+Q. \quad (21)$$

Again, the Little's theorem allows us to obtain the average time T the request will spend in the system:

$$T=\frac{K}{\lambda}=\frac{A+Q}{\lambda}=\frac{1}{\mu}+W=\frac{1}{\mu}+\frac{P_C}{\mu(N-A)}. \quad (22)$$

Erlang C represents the basic model for Contact center simulation and parameters estimation. The most important limitation is the unlimited queue length. Nowadays it is not a computer memory related problem however no real caller would be satisfied by extremely long waiting time. Majority of long waiting calls would be interrupted before they are served, but the Erlang C model does not take it into account. So the real results and model results would not correspond in this case.

3. Conclusion

The aim of this paper is to show some of the approaches towards Contact center simulation and parameter estimation and verification. The simple Erlang B, Erlang C models and Markov models are probably the most widely used for this purpose. The greatest advantage of Erlang models is their simplicity and ability to describe the most important operation situations of Contact centers. However due to their limitations some special cases must be described by more complicated Markov models.

In the case of very complicated Contact centers all mathematical models would be extremely complicated. In such cases it is much easier to use simulation software and experimentally test various values of parameters in order to obtain expected results.

4. Acknowledgement

This work is a part of research activities conducted at Slovak University of Technology Bratislava within the scope of the projects VEGA No. 1/0565/09 „Modeling of traffic parameters in NGN telecommunication networks and services“ and AV No. 0015/07 “Optimization of multimedia traffic in NGN networks”.

5. References

[1] Unčovský, L. Stochastic models of operational analysis, 1st edition, Bratislava, ALFA, 1980.

[2] Polec, J., Karlubíková, T. Stochastic models in telecommunications 1, 1st edition, Bratislava, FABER, 1999.

[3] Riečanová, Z. et al. Numerical methods and mathematical statistics, 1st edition, Bratislava, ALFA, 1987.

[4] Vastola, K. Interarrival times of a Poisson process [Online], Troy (New York, USA) : Rensselaer Polytechnical Institue, 15.3.1996.

[5] Kvasnička, V. The graph theory I, Bratislava, [Online]. Available: http://www2.fiit.stuba.sk/~kvasnicka/DiskretnaMatematika/Chapter_10/transparencies10.pdf

[6] Diagnostic Strategies. Traffic Modeling and Resource Allocation in Call Centers, Needham (Mass., USA) : Diagnostic Strategies, 2003. Available: http://www.fer.hr/_download/repository/A4_1Traffic_Modeling.pdf

[7] Koole, G. Call Center Mathematics: A scientific method for understanding and improving contact centers, Amsterdam, Vrije Universiteit, Available: http://www.math.vu.nl/~koole/ccmath

[8] Bolgh, G. et al. ”Queueing Networks and Markov Chains”, 2nd edition, Hoboken (New Jersey, USA), John Wiley, c2006.

[9] The Extended Erlang B Traffic Model [Online], Westbay Engineers Ltd., 2008, Available: http://www.erlang.com/calculatormanual/index.html?the_extended_erlang_b_traffic_model.htm

[10] Yokota, M. Erlang, [Online], Available: http://hp.vector.co.jp/authors/VA002244/erlang.htm

[11] Baroňák, I., Ferenczyová, I. Implementation of Contact Centre to Healthcare Sector. In: 6th International Conference on Emerging eLearning Technologies and Applications ICETA 2008, Stará Lesná, Slovak Republic, September 11 - 13, 2008

[12] Halás, M., Kyrbashov, B. New trend in IP telephony signalization protocols. In 8th conference KTTO 2008, Ostrava, Czech Republic, 24-25.4.2008, pp. 67-69, ISBN 978-80-248-1719-4

[13] Janata, V. Relation between MPLS and Access Network. In 8th Conference KTTO 2008 24th - 25th April, 2008. Ostrava, Czech Republic, 24. - 25.04.2008, VŠB - Technical university of Ostrava, pp. 58-60, ISBN 978-80-248-1719-4