Method for Fast Estimation of Contact Centre Parameters Using Erlang C Model

Tibor Mišúth, Erik Chromý, Ivan Baroňák
Department of Telecommunications
Faculty of Electrical Engineering and Information Technology, Slovak University of Technology
Ilkovičova 3, 812 19 Bratislava, Slovak Republic
tibor.misuth@stuba.sk, erik.chromy@stuba.sk, ivan.baronak@stuba.sk

Abstract — The paper deals with Contact centre parameters estimation. The ideas are based on Erlang C mathematical model that is useful during Contact centre design and implementation. We concentrate on modification of the basic Erlangs’ formulae to provide resources non-demanding iterative method of Contact centre parameters’ calculation. Especially we focus on the staffing requirements to meet certain requests handling level with minimum queuing. The accuracy of the computation results based on newly derived formulae is proofed against simulation.

Keywords—Contact Centre; QoS; Erlang C model

I. INTRODUCTION

The public contact today is a must for almost every enterprise or organization. Availability of simpler, faster and more comfortable contact with the clients, customers or trade partners can be a crucial advantage in business competition.

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

The heart of the Contact centre consists of Automatic Call Distribution (ACD) module and several other optional modules ACD is responsible for correct and efficient routing of all types of requests (voice calls, e-mails, etc.) through the Contact centre to an adequate and skilled agent or system capable of handling the request.

For a company or Contact centre operator it is crucial to provide sufficient level of call handing quality (expressed mostly in terms of waiting time) with minimum possible resources. We present an alternative approach to staffing requirements calculation.

The next parts of this paper are structured as follows. First we present some assumption regarding telecommunication traffic and contact centre mathematical models. Next the Erlang C queuing model is presented. In section 3 a possible less resources demanding formulae are derived and presented. Finally, the results are validated by comparing to simulation measurements.

II. MATHEMATICAL MODEL OF CONTACT CENTRE

Each component of ACD (or contact centre) system can be more or less precisely converted to a mathematical model. Since the contact centres usually handle large amount of requests per time unit, the vast majority of these models are based on queuing theory. Accuracy of results depends on a correct model selection. Also the precise description of input parameters and variable dependencies can significantly affect them.

A. Input traffic model

The rate of arriving calls (or other requests) per time unit is considered as random variable. Contact centre’s goal is to provide its services to a wide range of customers therefore we can assume the requests population to be unlimited. Once the population is unlimited the contact centre system can be denoted as queuing system with unlimited request population [1], [2].

In such a wide population the request creators (i.e. the callers) choose the start moment of their call individually and independently of each other. Any events (calls) are considered independent from the mathematical point of view if for any arbitrary set of requests $A_1, A_2, \ldots A_n$ and arrival probabilities $P(A_1), P(A_2), \ldots, P(A_n)$ during randomly chosen time interval the following formula is valid [3]:

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

(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 acquire only countable set of values, so it is a discrete variable. Incoming requests into a queuing system are mostly modelled by Poisson distribution [2], [3]. Following formulae define its probability density function, mean and variance:

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

(2)

$$E(X) = \lambda,$$

(3)

$$D(X) = \lambda,$$

(4)

where $\lambda$ represents the average event arrival rate (i.e. avg. number of requests per time unit).
The inter-arrival time is a very important variable as well. Since the arrival rate is a 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]:

$$f(t) = \lambda e^{-\lambda t} \tag{5}$$

where $t$ represents the inter-arrival time (random variable).

B. Service group model based on Erlang C

The service group consists of a queue and at least one human agent. The agent is (from the queuing theory point of view) a server where the requests are processed. In case of 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.

Even now, a century later since the Erlang models have been published, are these one of the basic approaches used in parameter estimation of telecommunication systems. Since the contact centre is considered to be a telecommunication system the models are valid for it as well.

Both Erlang and Poisson process based Markovian models use the same assumptions and following requirements must be met [5]:
- number of sources (requests population) is much greater than the number of servers,
- requests are generated randomly and independently of each other,
- average number of requests per time unit from all sources is constant,
- request handling time is a random variable with exponential distribution,
- queueing is based on the First in First out (FIFO) algorithm.

As we mentioned earlier the first two requirements are easily met by having large group of potential callers (request creators). Usage of Poisson distribution for arrival process ensures, the average number of request per time interval is stable and equal to $\lambda$. The remaining two requirements were implemented into the simulation application.

Erlang presented two basic models that differ in presence of a queue. The simpler Erlang B model does not contain a queue and requests over the current capacity of the system (no idle agent available) are permanently lost. This model is more suitable for dimensioning of telephone trunks or data throughput capacity (in case of VoIP traffic). In this paper we focus on the second Erlang’s model, so called Erlang C model. Presence of simple FIFO queue eliminates the main disadvantage of the previous model. If a temporary overload situation occurs, the request is put into the queue until any of the $N$ agents become available, i.e. the handling of any previous requests is finished. The FIFO strategy ensures the caller waiting the longest time to be served the first.

Erlang C model [5] is originally defined as function of two variables: the number of agents $N$ and the traffic load $A$ (expressed in Erl). Based on these parameters it calculates the probability $P_c$ (6) that the arriving request is not assigned to the agent immediately and it has to wait in the queue.

$$P_c(N,A) = \frac{A^N N}{\sum_{i=0}^{\infty} \frac{A^i}{i!} + \frac{A^N N}{N!(N-A)}} \tag{6}$$

If we know the rate of calls per unit time $\lambda$ and the average number of served requests per the same unit $\mu$ (so the average handling time is $1/\mu$) then the traffic load can be easily evaluated as [6]:

$$A = \frac{\lambda}{\mu} \tag{7}$$

Furthermore we define the variable $\rho$, that represents the average utilization of each agent as [1], [2], [7]:

$$\eta = \frac{\lambda}{N\mu} \tag{8}$$

For system stability reason (i.e. the number of requests present in the queue does not extend to infinity) the value of $\eta$ is required to be less than 1 (stability criterion). From the conjunction of (6), (7) and (8) we can obtain following formula:

$$P_c(N,\eta) = \frac{(N\eta)^N}{\sum_{i=0}^{\infty} \frac{(N\eta)^i}{i!} + \frac{(N\eta)^N}{N!(1-\eta)}} \tag{9}$$

If we start from Markov 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], [7]. 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 a queue, there are some more parameters and variables, 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 the time the request spends in the queue before it is served by an agent. This value is a random variable described by the probability density function [7]:

$$F_W(\tau) = \begin{cases} 1 - P_c & \tau = 0 \\ 1 - P_c \cdot e^{-\mu(N-A)\tau} & \tau > 0 \end{cases} \tag{10}$$

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

$$W = \frac{P_c}{\mu(N-A)} \tag{11}$$

and using Little’s theorem [8] and (10) we can obtain the average number of requests in the queue $Q$:
\[ Q = \lambda W = \frac{\lambda}{\mu(N-A)} P_C = \frac{A}{(N-A)} P_C \cdot \]

The general definition of probability density function of any statistical distribution and its properties [3] gives us an opportunity to derive Grade of Service (GoS) parameter value from equation (10) once the Acceptable Waiting Time (AWT) value is known [1], [7]:

\[ GoS = 1 - P_C \cdot e^{-\mu(N-A)AWT} \]  

(13)

GoS defines the percentage of incoming calls that are answered no later than defined AWT (usually 20 sec.). Following equation is valid for the average number of requests \( K \) in queuing system [7]:

\[ K = N\eta + \frac{\eta}{1-\eta} P_C = A + \frac{A}{(N-A)} P_C = A + Q \]  

(14)

Again, the Little’s theorem allows us to obtain the average time \( T \) the request will spend in the system:

\[ T = \frac{K}{\lambda} = A + \frac{Q}{\lambda} = \frac{1}{\mu} + W = \frac{1}{\mu} + \frac{P_C}{N-A} \]  

(15)

### III. EVALUATION OF ERLANG C MODEL

All calculations using Erlang C model are based on three variables \( A, N \) and \( P_C \), of which any two can be the input variables and the remaining one is computed by the formulæ. Once all three values are known, more queuing system parameters can be evaluated using set of equations in previous section.

The computation of a queuing probability \( P_C \) once the traffic load \( A \) and the number of agents \( N \) are defined is the most straightforward problem. To avoid numerical errors (division of two relatively large values), the basic formula (6) can be easily modified into

\[ P_C(N,A) = \frac{1}{1 + \frac{N-A}{N} \left( \sum_{k=0}^{N} \frac{1}{(N-k)A} \right)} = \frac{1}{1 + \frac{N-A}{N}} \]  

(16)

Using the same principle as for Erlang B model in [8]. Furthermore the Horner scheme can be easily applied to reduce number of required operations to compute the sum part of the denominator \( s \) as

\[ s_i = (1 + s_{i-1}) \cdot (i/A) \]  

(17)

for \( i = 1 \ldots N \). Then

\[ s_N = s. \]  

(18)

The remaining two situations (\( A \) or \( N \) values to be calculated) are complicated to solve analytically, but the numerical solution provides us the expected results with adequate accuracy. Probably the most important calculation for a Contact centre provider is determination of number of required agents \( N \) to handle given traffic load \( A \) with queuing probability not more than \( P_{C_{max}} \). The computation is based on cyclic calculations of \( P_C \) for incrementing values of \( N \) until the result is below expected threshold. However using (16) the stop condition can be easily transformed to

\[ s_i \leq \frac{1-P_{C_{max}}}{N_i-A} \frac{N_i}{N_{i+1}} \]  

(19)

and the for each step we can use

\[ N_{i+1} = N_i + 1 \]  

(20)

\[ s_{i+1} = (1+s_i) \frac{N_{i+1}}{A} \]  

(21)

This extremely shortens the solving time since each iteration requires only a few simple operations. Obviously to fulfill the stability criterion the number of agents \( N \) must be higher than expected traffic load \( A \) so the first check of stop condition could be postponed until \( i > A \).

The last combination of variables (i.e. unknown traffic load \( A \)) the \( N \) agents are capable handling at queuing probability less or equal to \( P_C \) is the most complicated case. It requires finding the solution of

\[ f(x) = P_C(x,N) - P_C_{INPUT} = 0 \]  

(22)

where \( N \) and \( P_C_{INPUT} \) are defined and \( P_C(x,N) \) is the queuing probability for current value of \( x \). Since the \( P_C(x,N) \) is monotonic function, only one real solution exists and its interval can be limited to \((0;N)> \) to fulfill the stability criterion. Any available numerical solution method for such an equation can be applied.

### IV. MODEL VERIFICATION

We decided to verify the model’s reliability and accuracy (including the modified calculations) by means of simulation [9]. It was implemented in MATLAB taking all Erlang C model’s requirements into account (see above). The simulation algorithm consists of three basic sections:

- input parameters verification, request arrival times determination and handling times determination (based on exponential distribution),
- simulation with discrete time steps (1 sec),
- results calculation and presentation.

Both the arrival traffic flow and requests handling times values are tested against its statistical distribution using \( \chi^2 \)-square test with \( \alpha = 0.05 \). Since the Erlang models describe the steady state of queuing system, the very beginning part of simulation results is discarded until the model reaches the steady state. The simulation length was set to 30 hours of simulated traffic with avg. arrival rate set to \( \lambda = 667 \) requests per hour and avg. handling time of 150 seconds. The number of agents was variable from 28 (minimum required to fulfill the stability criterion since \( A = 27.8 \) Erl) up to 37.

<table>
<thead>
<tr>
<th>( N )</th>
<th>( P_c ) [%]</th>
<th>( K ) [h]</th>
<th>( T ) [h]</th>
<th>( Q )</th>
<th>( W ) [s]</th>
<th>( GoS ) [%]</th>
<th>( p ) [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>28</td>
<td>95.4</td>
<td>155.0</td>
<td>836.6</td>
<td>127.2</td>
<td>686.6</td>
<td>7.2</td>
<td>99.3</td>
</tr>
<tr>
<td>29</td>
<td>75.3</td>
<td>45.1</td>
<td>243.5</td>
<td>17.3</td>
<td>93.5</td>
<td>35.9</td>
<td>95.8</td>
</tr>
<tr>
<td>30</td>
<td>58.7</td>
<td>35.2</td>
<td>189.9</td>
<td>7.4</td>
<td>39.9</td>
<td>56.3</td>
<td>92.6</td>
</tr>
<tr>
<td>31</td>
<td>45.1</td>
<td>31.7</td>
<td>171.1</td>
<td>3.9</td>
<td>21.1</td>
<td>70.6</td>
<td>89.7</td>
</tr>
<tr>
<td>32</td>
<td>34.1</td>
<td>30.9</td>
<td>162.1</td>
<td>2.2</td>
<td>12.1</td>
<td>80.6</td>
<td>86.8</td>
</tr>
</tbody>
</table>
Direct comparison of the tables above shows that the simulation results are very similar to Erlang C model calculations. The biggest difference can be seen for low number of agents or the situation when the stability criterion is matched by the minimum number of agents. However, the more overstaffed the contact centre is, the better precision of model results is gained.

From the callers point of view the most important characteristics are probability of call queuing ($P_C$), average waiting time (how much time does the caller spend waiting for an agent) $W$ and Grade of Service (GoS) level (the probability of call reception in 20 seconds). Fig. 1, Fig. 2 and Fig. 3 display the relation of these variables to the number of agents present in the contact centre ($N$).

![Figure 1](image1.png)  
**Figure 1.** Call queuing probability dependency on number of agents.

![Figure 2](image2.png)  
**Figure 2.** Grade of Service level dependency on number of agents.

![Figure 3](image3.png)  
**Figure 3.** Average waiting time dependency on number of agents.

Table II.

<table>
<thead>
<tr>
<th>$N$</th>
<th>$P_C$ [%]</th>
<th>$K$</th>
<th>$T$ [s]</th>
<th>$Q$</th>
<th>$W$ [s]</th>
<th>GoS [%]</th>
<th>$p$ [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>28</td>
<td>91.9</td>
<td>88.5</td>
<td>477.7</td>
<td>60.9</td>
<td>328.3</td>
<td>12.3</td>
<td>98.8</td>
</tr>
<tr>
<td>29</td>
<td>74.0</td>
<td>43.2</td>
<td>233.3</td>
<td>15.4</td>
<td>83.4</td>
<td>38.0</td>
<td>95.6</td>
</tr>
<tr>
<td>30</td>
<td>59.3</td>
<td>35.1</td>
<td>188.9</td>
<td>7.2</td>
<td>38.8</td>
<td>55.9</td>
<td>93.0</td>
</tr>
<tr>
<td>31</td>
<td>46.2</td>
<td>32.4</td>
<td>173.9</td>
<td>4.5</td>
<td>24.1</td>
<td>69.3</td>
<td>90.0</td>
</tr>
<tr>
<td>32</td>
<td>33.7</td>
<td>30.0</td>
<td>162.2</td>
<td>2.2</td>
<td>11.9</td>
<td>81.3</td>
<td>86.9</td>
</tr>
<tr>
<td>33</td>
<td>25.1</td>
<td>29.2</td>
<td>157.7</td>
<td>1.4</td>
<td>7.3</td>
<td>87.5</td>
<td>84.4</td>
</tr>
<tr>
<td>34</td>
<td>19.2</td>
<td>28.8</td>
<td>155.4</td>
<td>0.9</td>
<td>4.7</td>
<td>91.5</td>
<td>82.1</td>
</tr>
<tr>
<td>35</td>
<td>13.0</td>
<td>28.3</td>
<td>152.5</td>
<td>0.5</td>
<td>2.5</td>
<td>95.5</td>
<td>79.4</td>
</tr>
<tr>
<td>36</td>
<td>8.7</td>
<td>28.0</td>
<td>151.2</td>
<td>0.3</td>
<td>1.5</td>
<td>97.4</td>
<td>77.1</td>
</tr>
<tr>
<td>37</td>
<td>6.4</td>
<td>28.1</td>
<td>151.4</td>
<td>0.2</td>
<td>1.0</td>
<td>98.2</td>
<td>75.3</td>
</tr>
</tbody>
</table>

The agent utilization decreases (Fig. 4) almost linearly with increasing number of agents. This results into fact, that small increment in number of agents does not cause steep agent utilization decrease and the increased costs of added staff is compensated by significant improvements of relevant call handling metrics that leads to higher satisfaction of all potential callers. This phenomenon is visible the more the contact centre is operated near its stability bound in relation to number of agents and incoming traffic load.

The most important decision of the contact centre manager is to define equilibrium point in number of agents. In our model case, according to simulation and computation...
results 33 – 35 agents should be sufficient to provide smooth operation of the centre with up to approx. 95% of all incoming requests assigned to an agent in 20 seconds and average waiting time not more than 7 seconds. From the other point of view agents spend at least 80% of their working time by call handling and the remaining time forms the idle period that can be utilized to e.g., related paper work, agent education etc.

V. CONCLUSION

The aim of this paper is to show some of the approaches towards Contact center modelling and simulation that leads to parameter estimation and verification. The Erlang C model is probably one of 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 with acceptable level of accuracy. Simulation results show almost no difference to calculation results is steady state of operation.

The basic form of Erlang C model and its formulae can be adjusted to provide higher level of computation accuracy and more importantly less resource demanding algorithms. These modified formulae enable calculation of any model parameter using fast iterative process.

ACKNOWLEDGEMENT

This work is a part of research activities conducted at Slovak University of Technology Bratislava, Faculty of Electrical Engineering and Information Technology, Department of Telecommunications, within the scope of the projects VEGA No. 1/0565/09 „Modelling of traffic parameters in NGN telecommunication networks and services“ and ITMS 26240120005 “Support for Building of Centre of Excellence for SMART technologies, systems and services”.

REFERENCES