ERLANG C FORMULA AND ITS USE IN THE CALL CENTERS

Erik CHROMY¹, Tibor MISUTH¹, Matej KAVACKY¹

¹Department of Telecommunications, Faculty of Electrical Engineering and Information Technology, Slovak University of Technology, Ilkovicova 3, 812 19 Bratislava, Slovak Republic
chromy@ktl.elf.stuba.sk, misuth@ktl.elf.stuba.sk, kavacky@ktl.elf.stuba.sk

Abstract. This paper deals with calculation of important parameters of the Call Center using the Erlang C formula and then results are verified through simulations. Erlang C formula is defined as a function of two variables: the number of agents \( N \) and the load \( A \). On the base of their values it is possible to determine the probability \( P_C \), that the incoming call will not be served immediately, but it will have to wait in the waiting queue. Simulations satisfy the assumption of Markov models.

Keywords
Call center, Erlang C formula, Markov models, Quality of Service.

1. Introduction

Call Center is dynamical, technical system (package of technical equipments – hardware, software and human sources) designed for effective connecting people with the requirements for service with operator or with systems able to satisfy their requirements. The core of the Call Center is Automatic Call Distribution (ACD).

Each of the components of ACD can be described with some precision by means of mathematical tools and causalities. Since the ACD systems process a large number of incoming requests, the majority of models are based on the principles of mathematical statistics. The right choice of a statistical model is able to ensure the sufficient accuracy of the results. It is essential to describe the dependency of input variables and parameters that can greatly affect the accuracy of the results. The modeling of Call Center parameters is possible through Markov models, but also through Erlang formulas.

This paper deals with calculation of important parameters of the Call Center (which affect proper operation of such queuing system) using the Erlang C formula and then results are verified through simulations. These simulations satisfy the assumption of Markov models.

1.1 Erlang C formula and M/M/m/∞ model

Immediate rejection of call by reason of occupation of all agents (as expected in Erlang B formula) is in terms of provided services by Call Center inappropriate solution. This shortness is eliminated in second Erlang formula – Erlang C. In the case that the call cannot be served immediately, the call is placed into the waiting queue with unlimited length. If the release of one of the agents happens, it is automatically assigned to the following call from the queue. If the waiting queue is empty, the agent is free and he waits for next call.

Erlang C formula [5] is in the original form defined as a function of two variables: the number of agents \( N \) and the load \( A \). On the base of their values it is possible to determine the probability \( P_C \), that the incoming call will not be served immediately, but it will have to wait in the waiting queue.

\[
P_C(N,A) = \frac{A^N N!}{(N - A)!} \sum_{i=0}^{N-A} \frac{A^i}{i!} - \frac{A^{N+1}}{N(N - A)}
\]

(1)

where

\[
A = \frac{\lambda}{\mu}.
\]

(2)

Now, we use the relationship between load \( A \) (2), the average number of calls per time \( \lambda \) and the average number of requests processed per time \( \mu \). Next, we define the variable \( \eta \), which represents the load of 1 agent as [2], [3], [7]:

\[
\eta = \frac{\lambda}{N\mu}.
\]

(3)

By substituting (2) and (3) into equation (1) we have:
its properties on the distribution function (5), we can get the average number of requests in waiting queue as and by applying of Little theorem [7] and formula (2) we will get from Little theorem a value of the average time T, that the requirement spend in the system:

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

Other important parameter is the average utilization of agents \( \eta \) [2]:

\[ \eta = 1 - \frac{\sum_{k=0}^{N-K} A^k N - k}{\sum_{i=0}^{N-1} i + \frac{A^k N}{N!(N-A)}}. \]

Erlang C formula is the basis for the analysis of parameters and simulation of the call center. Its shortness is the assumption of the unlimited waiting queue. This is not a problem in terms of available storage capacity, and therefore the waiting queue could be potentially unlimited, but no real caller will wait too long. Therefore, the limitation of waiting period represents other parameter that is under consideration in the special Markov models.

### 1.2 M/M/m/M Model with Limited Length of Waiting Queue

A limited number of requirements placed in the queue of queuing system can be described by Markov model M/M/m/M, where the maximum number of requests in the system \( M \) is greater or equal to the number of servers \( m = N \) [2].

The probability \( p_0 \) [2], that in the system occurs exactly 0 requirements (i.e. empty) is defined as:

\[ p_0 = \begin{cases} \sum_{k=0}^{N-M} A^k N - k, & \text{if } M < N \cr \sum_{k=0}^{N-M-1} A^k N - k, & \text{if } M = N \end{cases} \]

and then we can define the probability of the call rejection \( P_R \) [2] as:

\[ P_R = \frac{A^M}{N!} \cdot p_0 \]

and the probability, that the call will be assigned into the waiting queue \( P_C \) [2] as:

\[ P_C = \frac{A^N}{N!} \cdot p_0. \]
Furthermore, for the values $Q$ [2] and $K$ [2] can be calculated as

$$Q = \frac{p_0 A^{N+1}}{(N-1)! A^N} \left(1 - \frac{1}{N\lambda}ight)^{M-N+1},$$

$$K = Q + A \left(1 - \frac{A^M}{N^{M-N+1}}\right).$$

(15)

(16)

The time characteristics can be calculated on the basis of Little theorem [7].

These relations are relatively complicated. It is therefore possible a consideration, if the approximation using standard Erlang C formula is not sufficient, or what are the conditions, that such approximation is sufficiently exact. If the average number of requests in the system $K$ will be far less than the limit $M$, it is possible to apply Erlang C formula without thinking of the maximum capacity. The more closely will be the value $K$ to the limit, the less accurate results Erlang C formula will provide.

2. The Principle of Realized Simulations

The basic of the simulated algorithm consists of three blocks:

- set up of inputs,
- traffic simulations,
- processing of measured values and their presentation.

2.1 Modeling of Inputs

The basic parameters are:

- average number of incoming calls into the call center per time $\lambda$,
- average time of call processing by agent $1/\mu$,
- number of agents $N$.

The group of time parameters consists of:

- the total length of simulated traffic in seconds $T_{SIM}$,
- time step of simulation $T_{STEP}$ (by default 1 second),
- time to steady state $T_{NAB}$.

Vector with arrival times of each call based on the average number of requests per time unit and length of simulation is created by random generator. There is used a property about exponential distribution of length of the intervals between arrivals. Generated random variables are then tested by the statistical chi-square [4]. There is test of vector consistency with exponential distribution on significance level $\alpha = 0.05$. In the case of unsuccessful test, the entire vector is randomly generated once again. Number of generated calls is by 20% higher than the average number of calls that should enter into the call center through the simulated period.

Furthermore, vector with service times of individual calls that will be allocated during the simulation, is generated for each agent based on his average service time.

All these operations are performed in advance by reason of high rate and efficiency of MATLAB by working with vectors and matrices. The generation of set of values is then faster than a gradual generation of individual values during the simulation.

2.2 Traffic Simulation

The core of simulation is realized as a cycle, in which the each iteration represents one time step in simulation. In the each iteration is the vector of incoming calls compared with actual time and if it is necessary, the call enters into the simulated call center in proper timer. Concurrently the time of incoming call is recorded for later calculations. Call is placed into the waiting queue or directly assigned to a free agent. The time index is stored for later calculations (average waiting time) at the moment of call assignment.

Furthermore, in the each iteration the status and the occupancy of all agents are checked. If any of the agents is free, it will be assigned the first call from the waiting queue. If more agents are free, the call is assigned to agent that didn’t work for the longest time. If there is not a call in the waiting queue, the agent remains as free and expects the arrival of another request into the system.

During the one iteration there is possibility of entry into the system and also assignment to the agent of more than one call simultaneously. The number of calls in the system is stored in each step so then it is possible to determine the number of calls in the waiting queue.

The mathematical model of queuing system is defined for the steady state. It means, that in the moment of parameters monitoring, the simulation runs infinitely long time.

The run of the simulation is terminated after predefined simulation time. Only the calls that are terminated (and are thus served by agent) before expiration of the simulated time are included into the statistics.
2.3 Processing of the Simulation Results

This phase of the algorithm ensures the processing of all measured data during the simulation (time of events generation, number of calls in the system, agent occupancy, …). Based on these the particular parameters monitored in the call center are calculated. Thus obtained simulation results can be then compared with the expected values obtained by calculations through the mathematical model.

The general variables available from the results of the simulated model are:

- real length of the simulation (calculation time),
- number of simulated calls,
- average service time per one call (1/μ),
- average number of calls in the system (K),
- average time spent by user in the system (T),
- number calls placed in the waiting queue,
- probability of blocking P_B, respectively probability of waiting P_C,
- average utilization of agents η.

3. Calculations Using the Erlang C Model

Erlang C model works in the basic form with 3 input parameters (A, N, P_C). By use of relation (2) it is possible to divide the value of load A (generated by incoming calls) into 2 components: λ and 1/μ. Calculator thus always works with 4 values, while 3 of them act as input parameters and the last parameter is the output parameter. By adding of the waiting queue we can obtain a set of new parameters that can be calculated and then compared with the simulation results:

- average number of requests in the waiting queue Q (using (7)),
- average number of requests in the call center K (using (9)),
- average waiting time in the waiting queue W (using (6)),
- average time spent in the call center T (using (10)),
- value of GoS for AWT=20 s (using (8)),
- average utilization of agents η (using (11)).

Probability of insertion into waiting queue P_C can be entered in three different variants:

- direct entry of value P_C,
- entry of average waiting time W,
- entry of GoS.

When using a P_C as input variable, this variable is considered to the upper limit.

3.1 Calculation of the Parameter Pc

The calculation of the unknown value of probability of waiting can be easily realized by basic relation for Erlang C model (1). The dividing of two large number could lead to numerical errors and the obtained result might not be accurate. Therefore we derived an alternative relation, that is identical to the original (1) (in terms of result):

$$P_C(N, A) = \frac{1}{1 + \frac{N - A}{N} \sum_{k=0}^{N-1} \frac{N - k}{k!} \frac{1}{A}} \quad (17)$$

Similarly, we can use Horner scheme [8] also for more efficient calculations.

3.2 Calculation of the Parameter N

Analytical calculation according to (1) respectively (17) would be very difficult. The solution is therefore through a numerical method. The easiest technique is to gradually increase of the number of agents and continuous checking the stop conditions. This depends on the form, in which the input value P_C is inserted (thus direct, or as W, or as GoS). At the moment, where the current value in the calculation is less or equal to the required value, the necessary number of agents is found. The implementation uses this idea, but applying of the relation (17) in every step of the calculation it is possible to use the current result obtained from the previous iteration cycle for the lower value of N. So, this is very quickly method to find the unknown value N.

3.3 Calculation of the Parameter λ or 1/μ

The calculation of one of these unknown quantities in terms of Erlang C model means the find out of the load A, that can the system process at the specified parameters N and P_C (respectively W and GoS parameters). Consequently, we can calculate the second one by applying the formula (2) and one of the value λ and 1/μ.

In terms of implementation, the finding of the unknown value A is the most difficult of all three combinations. As the solution we can use the feature, that allows to search the value of unknown x, for which it holds f(x) = 0. In this case, the function f(x) for input P_C is:

$$f(x) = P_C(x, N) - P_{C,\ \text{INPUT}} = 0 \quad (18)$$

If the input value is defined as the average waiting time W, respectively GoS, we can use the following substitutions by (6) and (10):

$$P_C = W/M \cdot (N - A) \quad (19)$$
\[
P_C = \frac{1 - \text{GoS}}{e^{\mu(N-A)^{AWT}}}.
\]

In both cases we have the value \( \mu \). Therefore, if we need the solution for the unknown value \( 1/\mu \), we must find the solution for the value \( \mu \) and only then to calculate the load \( A \) by (2). The functions \( f(x) \) for input \( W \) (by use of substitution (19)) are:

\[
f(x) = P_C(x, N) - W \cdot \mu \cdot (N-x) = 0,
\]

respectively, if \( \mu \) is unknown, then:

\[
f(x) = P_C \left( \frac{\lambda}{3600 \cdot x}, N \right) - \frac{W}{x} \left( N - \frac{\lambda}{3600 \cdot x} \right) = 0.
\]

If the input parameter is GoS, the substitution (20) is used and the function for parameter \( \lambda \) is:

\[
f(x) = P_C \left( x, N \right) - \frac{1 - \text{GoS}}{e^{\mu(N-A)^{AWT}}} = 0,
\]

respectively, if \( \mu \) is unknown, then:

\[
f(x) = P_C \left( \frac{\lambda}{3600 \cdot x}, N \right) - \frac{1 - \text{GoS}}{e^{\mu(N-A)^{AWT}}} = 0.
\]

4. Simulations

The whole process of simulation for Erlang C model is shown in the following Fig. 1:

![Fig. 1: Process of simulation for Erlang C model.](image)

4.1 Simulation Results

The parameters of simulated model are:

- 667 incoming calls per 1 hour,
- average call service time is 150 seconds,
- simulation time of the call center work is 30 hours with 1 second step (run to steady state is 1 hour).

The calls that cannot be processed immediately are placed into the waiting queue (according to Erlang C formula). For the system stability it is necessary to satisfy the condition \( A < N \). Therefore the calculations and simulations are realized for the number of agents in the range of 28 to 37. In this range are the most notable changes observed in output parameters. The Tab. 1 and Tab. 2 shows the results obtained by calculation using the mathematical model and results also obtained by simulations.

<table>
<thead>
<tr>
<th>N</th>
<th>Pc [%]</th>
<th>K</th>
<th>T [s]</th>
<th>Q</th>
<th>W [s]</th>
<th>GoS [%]</th>
<th>η [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>28</td>
<td>95,4</td>
<td>155</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</td>
<td>162,1</td>
<td>2,2</td>
<td>12,1</td>
<td>80,6</td>
<td>86,8</td>
</tr>
<tr>
<td>33</td>
<td>25,3</td>
<td>29,1</td>
<td>157,3</td>
<td>1,4</td>
<td>7,3</td>
<td>83,7</td>
<td>84,2</td>
</tr>
<tr>
<td>34</td>
<td>18,5</td>
<td>28,6</td>
<td>154,5</td>
<td>0,8</td>
<td>4,5</td>
<td>91,9</td>
<td>81,7</td>
</tr>
<tr>
<td>35</td>
<td>13,3</td>
<td>28,3</td>
<td>152,8</td>
<td>0,5</td>
<td>2,8</td>
<td>94,9</td>
<td>79,4</td>
</tr>
<tr>
<td>36</td>
<td>9,4</td>
<td>28,1</td>
<td>151,7</td>
<td>0,3</td>
<td>1,7</td>
<td>96,8</td>
<td>77,2</td>
</tr>
<tr>
<td>37</td>
<td>6,5</td>
<td>28</td>
<td>151,1</td>
<td>0,2</td>
<td>1,1</td>
<td>98,1</td>
<td>75,1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>N</th>
<th>Pc [%]</th>
<th>K</th>
<th>T [s]</th>
<th>Q</th>
<th>W [s]</th>
<th>GoS [%]</th>
<th>η [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>28</td>
<td>91,9</td>
<td>88,5</td>
<td>478</td>
<td>60,9</td>
<td>328</td>
<td>12,3</td>
<td>98,8</td>
</tr>
<tr>
<td>29</td>
<td>74</td>
<td>43,2</td>
<td>233</td>
<td>15,4</td>
<td>83,4</td>
<td>38</td>
<td>95,6</td>
</tr>
<tr>
<td>30</td>
<td>59,3</td>
<td>35,1</td>
<td>189</td>
<td>7,2</td>
<td>38,8</td>
<td>55,9</td>
<td>93</td>
</tr>
<tr>
<td>31</td>
<td>46,2</td>
<td>32,4</td>
<td>174</td>
<td>4,5</td>
<td>24,1</td>
<td>69,3</td>
<td>90</td>
</tr>
<tr>
<td>32</td>
<td>33,7</td>
<td>30</td>
<td>162</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>158</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</td>
<td>0,9</td>
<td>4,7</td>
<td>91,5</td>
<td>82,1</td>
</tr>
<tr>
<td>35</td>
<td>13</td>
<td>28,3</td>
<td>153</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</td>
<td>151</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</td>
<td>0,2</td>
<td>1</td>
<td>98,2</td>
<td>75,3</td>
</tr>
</tbody>
</table>

Obtained simulation results and calculations are very similar. The differences exist only in the case of the minimum number of agents. However, it is probably that any company will not carry on the call center with extremely poor quality of service [9], [10], [11], [12] delivering (mainly the very long waiting time).

From the caller point of view, the most important parameters are the average waiting time in the waiting queue \( W \) and parameter \( \text{GoS} \) (this case is evaluated for the parameter \( \text{GoS} \).
The caller expects the lowest value of $W$ and also the value GoS, which is close to 100%. From the call center point of view, the most important parameters are the number of agents $N$ and their utilization $\eta$, because these two variables significantly affect the financial demands of service. The aim of the operator is to minimize the number of agents and to maximize their utilization. The aim of the analysis is therefore to find such a minimum number of agents $N$, when the operation parameters are yet on the sufficient level. As suitable we can consider the GoS parameter on level 80% and the average waiting time about 10 seconds.

Fig. 2. Simulation results.

The Fig. 2 shows the characteristic curve of the important call center simulation results according to Erlang C model assumptions in relation with the number of agents $N$. The characteristic curve of the average waiting time has a very strong exponential character and thus a slight increase of the number of agent (about 1 to 2 agents) can bring a significant improvement of this parameter. A similar, even though less aggressive, is the characteristic curve of the probability $P_C$. GoS parameter also exponentially converges to the level 100% and we can see that a small change in the number of agents can bring significant improvement. The characteristic curve of agents load $\eta$ is in the displayed range almost linear.

According to the above mentioned requirements on the provided quality of service by the call center is in this case possible to consider 32 agents as sufficient. By adding two agents it is possible to shorten the average waiting time by half and to increase the value of GoS parameter at 10% on very decent level (90%). The utilization of agents does not decrease below 80% and therefore it does not create unnecessarily long pauses, when the agents were redundant.

5. Conclusion

Based on calculations and simulations it can be stated, that in term of simplicity and accuracy of obtained results Erlang C formula is applicable for call center simulations. However, its shortness is the possibility of calculations only for one service group, and also the need to define for all agents the same service time.

Despite these limitations, it is possible to use the basic Erlang C formula also for calculations for call center with several service groups. In this case when it is possible to determine the probability that calls are routed to the individual service groups, then it is possible to use the basic Erlang C formula for calculations for each service group individually. The number of incoming calls per time unit is the aliquot portion of the total number of incoming calls into the call center. The Erlang C formula calculation can be used in the case of different performance of agents. It is possible to determine the value of the average call processing time or average number of the processed calls per time unit by dividing the total number of processed calls per time unit of all agents and the total number of agents in service group.

For the purpose of further study it would be interesting to expand the simulations by more independent service groups at the same time and the random distribution of call between them according to defined probabilities. Another interesting possibility could be different average service time for individual agents. Finally there is possibility to simulate the impact of unequal performance of agents on the results of the whole call center.

Acknowledgments

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 „Modeling of traffic parameters in NGN telecommunication networks and services“ and ITMS 26240120029 „Support for Building of Centre of Excellence for SMART technologies, systems and services II“.

References


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 (FEI STU) 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 FEI STU Bratislava.

Tibor MISUTH is a student of Ph.D. study at Department of Telecommunications, Faculty of Electrical Engineering and Information Technology of Slovak University of Technology Bratislava. He focuses on application of Erlangs’ equations both in classic telecommunication networks and modern IP networks.

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 (FEI STU) Bratislava. In 2006 he submitted Ph.D work “Quality of Service in Broadband Networks”.

Nowadays he works as assistant professor at the Department of Telecommunications of FEI STU Bratislava and his scientific research is focused on the field of quality of service and private telecommunication networks.