Service Management 1/2015, Vol. 15, ISSN: 1898-0511 | website: www.wzieu.pl/SM | DOI: 10.18276/smt.2015.15-01 | 5–11

PERFORMANCE ANALYSIS OF AN INBOUND CALL CENTER WITH TIME VARYING ARRIVALS

Maciej Rafał Burak

West Pomeranian University of Technology
e-mail: maciej.burak@zut.edu.pl

Received 15 March 2015
Accepted 1 June 2015

JEL classification C61, C44, C63

Keywords call center, continuous time Markov chains, CTMC, non-stationary model, balking, abandonment

Abstract The paper presents a novel method of performance analysis of a call center with balking and abandonment, i.e. in which the customer may not stay in the queue once realizing he is put on hold, or abandon the waiting queue if the waiting time is too long. In the paper we compare both an inherently transient and a stationary CTMC models of such an inbound call center, using realistic data. The inherently transient method we introduce delivers important characteristics of the model, with the quality close to Monte Carlo simulations, by using modified uniformization method.

Introduction

The provision of service delivery in a remote manner developed historically parallel to the development of the telecommunications technology. For companies, such as financial institutions, airlines, hotels, telecommunication companies and many others, remote contact via telephone or the internet has become already in the last two decades of the 20th century the primary communication channel for their customers, offering a broad range of

available services, ranging from infoline through operational customer services (e.g. telephone banking, ticket reservation, contract management) up to telemarketing of new services to existing and prospective customers.

The organizations providing such services are commonly refered to as call centers, although nowadays they can handle service requests not only in the form of telephone calls, but also make use of other communication channels, like e-mail, messaging systems, social networks or interactive internet based communications. For such multichanneled communication an alternative term 'contact center' is also used sometimes. If a call center serves incoming customer requests, the term inbound is used. Such operations, known as inbound call centers, are the primary topic of this paper.

According to a recent survey, "European Contact Survey 2012" conducted on behalf on ECCCO, the European Confederation of Contact Centre Organisations, the call center sector employs 3.2 mln people – more than 1% of the active population in Europe – and grows at about 4.5% rate, almost independently from the overall economic conditions. In Poland, the call centers offer more than 200k jobs. Regarding the type of their activities – 75,8% are devoted to serving incoming calls.

A call center comprizes a set of resources, mainly personel, computers and telecommunication equipment. Inbound call centers are highly labor-intensive operations, with the cost of staff members who handle phone calls (also known as "agents") typically consisting 60–80% of the overall operating budget (Aksin et al., 2007). The main goal of the management of the operations of an inbound call center is to optimize its costs (e.g. utilization of equipment, agents' time and knowledge) while simultanously delivering high service quality, measured both as the quality of customer interaction (competence, politeness) and quantitative quality of the operations, measured via stadardized benchmarks (e.g. service level, average time of answer). To achieve such potentially conflicting objectives, the managers are challenged with deploying an appropriate number of staff members to the right schedules in order to meet a time-varying demand for service.

Due to the size of the industry and its complexity, call centers have always attracted intense interest in the area of operational research, which is represented by current studies in various disciplines (see e.g. Aksin, Armony and Merhotra (2007) or Gans, Koole and Mandelaum (2003) for extensive overviews). As early as from the first decades of the 20th century, queuing theory, starting with the fundamental works of Andrey Markov and Agner Krarup Erlang, has provided service managers with the corresponding mathematical framework. From the modeling point of view, call centers can be viewed as queuing systems. Such models can be described by corresponding continuous time Markov chains (CTMC), whose steady-state distributions can be easily calculated using Erlang-C formula for the simplest M/M/n model or with the Erlang-A formula for models extended with phenomena of customer abandonment. However, their underlying assumptions do not allow for the modeling of more realistic systems, dealing with e.g. time varying arrivals or overflow.

The main objective of this work is to present a novel method of performance analysis of a realistic model of a call center with balking and abandonment, i.e. the customer may not stay in the queue once realizing he is put on hold, or abandon the waiting queue if the waiting time is too long. Moreover, the model can handle arbitrary varying arrival rates including the possibility of the system load exceeding 1 (overflow).

The paper is structured as follows. The next section summarizes current research on the area. Section 3 reviews the model and proposed modeling method. Section 4 presents results of numerical experiments based on an already published example of a real call center. The paper ends with a summary of results, conclusions and proposals for future research.

Managing Call Center operations – the quantitative view

The common approach to call center resources management is to build an agent schedule that minimizes costs while achieving some predefined quality measurements (e.g. service level). As such, staffing levels for each period of the scheduling horizon are used to perform the scheduling and rostering of agents. These levels (number of necessary agents) depend both on how many calls are arriving into the call center a given time (estimated by the call volume forecasts) and on the quantitative quality targets (service level). Once the forecasts and waiting time goals have been established, queuing models are used to determine the targeted number of service resources to be deployed. The simplest queuing model for a call center is the M/M/n model, which assumes e.g. both unlimited queue space and customer patience. Its steady-steady state distribution can be determined analytically using the Erlang-C formula. Another well-known model augments the M/M/n model with abandonment (with exponential waiting time) – with the Erlang-A formula (as proposed e.g. in Brown, et al. (2005)) to calculate the corresponding steady-state distribution. For more complicated models including e.g. blocking and balking, the steady state distribution can be obtained numerically as proposed by Deslauriers et al. (2007) or more recently by Phung-Duc and Kawanishi (2014).

However, as real call centers are time inhomogenous, with varying arrival rates and changing number of servers – scheduled to meet the forecasted demand and provide break time, stationary models cannot be applied directly. It is, therefore, common to use approximations, assuming the system being pointwise stationary. Examples of such well established methods can be found e.g. in Green, Kolesar and Whitt (2007), Aksin, Armony and Merhotra (2007) or in Brown et al. (2005).

Unfortunately, stationary approximations are in many cases not adequate. For example, Deslauriers et al. (2007) compared them with simulations based on real inbound call center data, with the conclusion that due to the nonstationarity only some of the performance measures can be estimated with satisfactory accuracy. Ingolfsson et al. (2007) compared them with an inherently transient model and found their results significantly inaccurate or even entirely unreliable. Nevertheless, they are still the methods most commonly used, which is usually justified by simple implementation and low computational costs.

Many authors proposed to use simulation, which can achieve any desired accuracy. However, in order to achieve acceptable precision, very long computational times are needed, which makes it often impracticable for common applications, such as schedule planning.

An alternative approach, which is very effective in terms of the accuracy of the model, is to analyze transient CTMC using numerical methods, solving effectively their corresponding system of ordinary differential equations (ODE's), as proposed in Ingolfsson et al. (2010), Bylina et al. (2009) or by the author in Burak (2014, 2015), which will also be used in our numerical examples.

Other, less computationally intensive, analytical methods that can approximate such nonstationary systems more accurately than stationary models are closure approximations, as well as fluid and diffusion approximations, discussed e.g. in Brown et al. (2005), Green, Kolesar and Whitt (2007), Czachórski et al. (2009) and Czachórski et al. (2014) or, for the direct comparison of some examples of such methods with the numerical methods and stationary approximations, in Ingolfsson et al. (2007).

The CTMC Model of a call center

The model used for the demonstration of the proposed method comes from Deslauriers et al. (2007). The authors analyzed there five CTMC models with varying degree of complexity applied for both pure inbound and blended (inbound and outbound) operations of a call center. All models were solved numerically to obtain their steady state distributions, which were then consequently used as pointwise stationary approximations. Afterwards, the results were compared to a Monte Carlo simulation based on the real call center data. Particularly interesting in this context is the model M_1 applied to an inbound call center, where the authors detected significant amount of error in comparison to the real call center data due to the stationary approximation of an inherently non-stationary model.

The data used for comparison is the same one as used by Deslauriers et al. (2007) and can be found in the Table 1.

Table 1. Input parameters of the model

Period (30 min)	Start time (hour)	Arrival rate λ(i) per half hour	Mean patience time (sec)	Avg. service time (sec)	Inbound agents
1	08:00	32.11	400	595.6	12
2	08:30	45.96	400	595.6	18
3	09:00	58.48	400	595.6	22
4	09:30	66.5	700	595.6	25
5	10:00	73.44	700	595.6	27
6	10:30	72.87	600	595.6	26
7	11:00	74.13	600	595.6	26
8	11:30	71.4	600	595.6	24
9	12:00	68.32	600	575.1	22
10	12:30	68.04	600	575.1	23
11	13:00	71.55	500	575.1	28
12	13:30	70.11	500	575.1	25

Source: Deslauriers et al. (2007).

The CTMC model (M_1 @Deslauriers et al. (2007) and also Burak (2015)) is defined as follows: the call center consists of s identical agents with a single FIFO waiting queue. Inbound calls arrive according to an inhomogenous Poisson process with rate λ(t), the service time is i.i.d. exponentially distributed with rate μ(t). The load ρ(t) = λ(t)/s(t) μ(t) can be bigger than 1.

Service requests that are not served immediately can leave the system (hang up or balk) with probability 1 – γ set for all modeled periods to 0.995 as in all models by Deslauriers et al. (2007), otherwise, after joining the queue, they abandon after reaching their patience time. The patience times are independent and identically exponentially distributed with mean 1/η. Queued requests are FCFS served. All of this is modeled via the state transition rates of a CTMC which is described by infinitesimal generator matrix Q(t) and the initial state probability vector p(0). The transient distribution at time t – p(t), for such a given time dependent generator matrix Q(t), can be calculated using modified Kolmogorov's forward equations:

$$p'(t) = p(t)Q(t) \tag{1}$$

where the vector $p(t) = [p_0(t), ..., p_n(t)]$ gives probabilities of the system being in any of the states at time *t*.

Because the process representing the system state k, equal to the number of busy servers plus number of requests waiting in the queue, is a birth-and-death process, it can be described by the following state dependent birth $q_{k;k+1}(t) = \lambda_k(t)$ and death $q_{k;k-1}(t) = \mu_k(t)$ rates:

$$\lambda_k(t) = \begin{cases} \lambda(t), & \text{if } 0 \le k \le s(t) - 1 \\ \gamma\lambda(t), & \text{if } s(t) \le k \le n - 1 \end{cases} \tag{2}$$

$$\mu_k(t) = \begin{cases} k\mu(t), & \text{if } 1 \le k \le s(t) - 1 \\ s(t)\mu(t) + (k - s(t))\eta, & \text{if } s(t) \le k \le n \end{cases} \tag{3}$$

The calculation of the state probability vectors p(t): t = i × 30 min representing the state of the system according to the input parameters in function of time has been calculated using the modified uniformization algorithm with steady-state detection as described in Burak (2014, 2015). As a result, the expected state of the system calculated from the state probability vector p(t) as:

$$ES(t) = \sum_i i\pi_i(t), \quad p(t) = [\pi_0 ... \pi_n] \tag{4}$$

has been used to calculate the expected utilization of agents time and compared with the same result derived from the stationary distribution probability vector for period i as proposed in Deslauriers et al. (2007) and is presented in Table 2.

Table 2. Comparison of the transient and (pointwise) stationary model

Period (30 min)	Inbound agents	Service level (%)		Utilization (%)	
		transient	stationary	transient	stationary
1	12	70.08	65.58	82.95	86.11
2	18	76.91	75.34	82.14	83.19
3	22	73.61	72.41	85.85	86.54
4	25	71.79	70.51	87.69	88.38
5	27	68.47	67.48	89.91	90.41
6	26	62.55	62.62	92.63	92.60
7	26	59.52	59.36	94.11	94.19
8	24	50.36	50.71	98.43	98.25
9	22	47.90	48.58	99.67	99.30
10	23	57.78	57.85	94.59	94.55
11	28	84.26	84.05	81.24	81.39
12	25	69.21	69.35	89.06	88.98

Source: own calculation.

The service level has been calculated for immediate service (no waiting) i.e. the percentage of requests that are served immediately. As we can see, the comparison between the pointwise stationary and the transient model shows greater discrepancies in cases of higher system variability (i.e. periods 1–4) where particularly the

difference in the calculated service level can diverge up to 5 percent points, which is already significant. This effect would, probably, be more visible if the traffic data was not averaged by whole model periods (30 min), justified by the pointwise steady-state assumption in Deslauriers et al. (2007) and considered there only as a rough-cut approximation, but, instead, something corresponding better with reality, e.g. by 5 min periods, as proposed in Burak (2014, 2015), as in actuality the arrival rate is not piecewise constant over fixed-length periods and the system is never in steady-state.

Conclusions and directions for future work

We have studied a realistic CTMC model of a call center both in a pointwise stationary and a transient way. The outcome shows that calculating the state probabilities in an inherently transient way provides results much closer to reality than the numerically calculated stationary approximation, which is already much more accurate than the commonly used Erlang-C and Erlang-A formulas. Although it is more computationally intensive than stationary models, it delivers results comparable with the simulations. The proposed solution could be refined further, using more accurate traffic data (i.e. aggregated by much shorter periods in order to preserve its non-stationary characteristics). It would be, in this context, particularly interesting, to compare the results with broader set of real call center data examples, which is currently the subject of our investigation.

References

Aksin, Z., Armony, M. & Mehrotra, V. (2007). The modern call center: A multi-disciplinary perspective on operations management research. Production and Operations Management, 16: 665–688. DOI: 10.1111/j.1937-5956.2007.tb00288.

Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S. & Zhao, L. (2005). Statistical analysis of a telephone call center. Journal of the American Statistical Association 100 (469): 36–50.

Bylina, J., Bylina, B., Zoła, A. & Skaraczynski, T. (2009). A Markovian model of a call center with time varying arrival rate and skill based routing. Computer Networks 2009. Springer Science Business Media: 26–33.

Burak, M. (2014). Multi-step uniformization with steady-state detection in nonstationary m/m/s queuing systems. arXiv preprint arXiv:1410.0804.

Burak, M. (2015). Inhomogeneous CTMC Model of a Call Center with Balking and Abandonment. Studia Informatica, 36, 2 (120): 23–34.

Czachórski, T., Fourneau, J.M., Nycz, T. & Pekergin, F. (2009). Diffusion approximation model of multiserver stations with losses. Electronic Notes in Theoretical Computer Science, 232: 125–143.

Czachórski, T., Nycz, T., Nycz, M. & Pekergin, F. (2014). Traffic engineering: Erlang and engset models revisited with diffusion approximation. Information Sciences and Systems 2014. Springer Science – Business Media: 249–256.

Deslauriers, A., L'Ecuyer, P., Pichitlamken, J., Ingolfsson, A. & Avramidis, A.N. (2007). Markov chain models of a telephone call center with call blending. Computers & Operations Research, 34 (6): 1616–1645.

Gans, N., Koole, G. & Mandelbaum, A. (2003). Telephone call centers: Tutorial, review, and research prospects. Manufacturing & Service Operations Management, 5 (2): 79–141.

Green, L.V., Kolesar, P.J. & Whitt, W. (2007). Coping with time-varying demand when setting staffing requirements for a service system. Production and Operations Management, 16: 13-39. DOI: 10.1111/j.1937-5956.2007.tb00164.x.

Green, L.V. & Soares, J. (2007). Computing time-dependent waiting time probabilities in m(t)/m/s(t) queuing systems. Manufacturing & Service Operations Management, 9: 54–61. DOI: 10.1287/msom.1060.0127.

Gross, D. & Miller, D.R. (1984). The randomization technique as a modeling tool and solution procedure for transient Markov processes. Operations Research, 32: 343–361. DOI: 10.1287/opre.32.2.343.

Haverkort, B.R. (2001). Markovian models for performance and dependability evaluation. Lecture Notes in Computer Science. Springer Science – Business Media: 38–83. DOI: 10.1007/3-540-44667-2_2.

Ingolfsson, A., Akhmetshina, E., Budge, S., Li, Y. & Wu, X. (2007). A survey and experimental comparison of service-level-approximation methods for nonstationary m(t)/m/s(t) queueing systems with exhaustive discipline. INFORMS Journal on Computing, 19: 201–214. DOI: 10.1287/ijoc.1050.0157.

Ingolfsson, A., Campello, F., Wu, X. & Cabral, E. (2010). Combining integer programming and the randomization method to schedule employees. European Journal of Operational Research, 202: 153–163. DOI: 10.1016/j.ejor.2009.04.026.

Malhotra, M., Muppala, J.K. & Trivedi, K.S. (1994). Stiffness-tolerant methods for transient analysis of stiff Markov chains. Microelectronics Reliability, 34: 1825–1841. DOI: 10.1016/0026-2714(94)90137-6.

Phung-Duc, T. & Kawanishi, K. (2014). Performance analysis of call centers with abandonment, retrial and after-call work. Performance Evaluation 80: 43–62.

Reibman, A. & Trivedi, K. (1988). Numerical transient analysis of Markov models. Computers & Operations Research, 15: 19–36. DOI: 10.1016/ 0305-0548(88)90026-3.

Stewart, W.J. (2009). Probability, Markov chains, queues, and simulation: the mathematical basis of performance modeling. Princeton University Press.

Van Moorsel, A.P. & Wolter, K. (1998). Numerical solution of nonhomogeneous Markov processes through uniformization. ESM: 710–717.

Cite this article as: Burak, M.R. (2015). Performance analysis of an inbound call center with time varying arrivals. Szczecin University Scientific Journal, No. 872. Service Management, 15 (1): 5–11.