

Queueing Models for Analyzing Customer Contact Center Operations

Mohan Raj Chinnaswamy and Manjunath Kamath
School of Industrial Engineering and Management
Oklahoma State University
Stillwater, OK 74078, USA

Robert A. Greve
Management Science and Information Systems
Oklahoma State University-Tulsa
Tulsa, OK 74106, USA

Abstract

Customer call centers, which represent a multi-billion dollar industry, are evolving into customer contact centers, in which customer contact happens through additional types of media – e-mail, fax, and the Web. We model the operations of contact centers, where the customer contact is via e-mail. We present ongoing research on the development of queueing network models aimed at analyzing the processing of e-mails by a network of customer service agents. Problem resolution time becomes an important performance measure in analyzing such systems. Unlike response time, resolution time is not normally addressed in call center models. The analytical modeling work presented here is largely exploratory, and is expected to help identify modeling issues that need to be studied further.

Keywords: Customer contact centers, e-mail, call centers, two-moment queueing approximations.

1. Introduction

The service sector has become a dominant part of the US economy, due in part to the e-commerce revolution of the late nineties. Better quality of service delivered virtually with very little or no waiting time is what customers frequently expect today. One notable facet of the service industry is the call center industry. Simulation models, stochastic models including queueing, and other quantitative models are frequently used in designing and improving call center operations. Customer call centers, which represent a multi-billion dollar industry, are evolving into customer contact centers. “A contact center is a collection of resources providing an interface between the service provider and its remote customers [10]”. The interface can be through any type of media – telephone, e-mail, fax, paper, chat sessions and the Web.

The contact center industry is vast and growing rapidly. According to a survey by callcenternews.com [4], customers prefer e-mail as the mode of interface with service providers. The response time in an e-mail governed environment is flexible, i.e., customers do not expect an e-mail to be answered within minutes, but often get frustrated waiting on the telephone even for only a few minutes. This flexibility may allow for the possibility of postponing a response. In contact centers, the traffic can be inbound, outbound or both. In an inbound contact center, agents only receive e-mail from customers. Examples of inbound contact centers are technical product support services and travel reservation centers. In outbound contact centers, agents initiate e-mail that is sent to customers. Examples of outbound contact centers are companies conducting surveys and market research. Operational planning, workforce population, routing policies, performance measures, and organizational behavior are important issues to be addressed in any e-mail based contact center. We focus on the performance analysis of a typical e-mail based, inbound customer contact center.

2. Literature Review

Call center research has focused on analyzing waiting times and customer impatience, and finding optimal staffing and routing policies. The mathematics of call-centers [6] is to a large extent based on the standard Erlang-loss formula. When the call center’s waiting time objectives are known for a day, the formula can be used to calculate the number of servers required for the day. Koole et al. [8] present an approximation method for analyzing the performance of call centers with skill-based routing. This method is used to determine the optimal skill sets for the call center employees.

Koole and Talim [9] study a multi-skill call center as a network of queues. Calls are considered as customers and agents are considered as servers. When an arriving call finds all agents busy, the call is routed to another queue, if one exists, or the call is lost. Therefore, to study the losses of a network, they approximate each queue as an M/M/r system. The inter-overflow time process at each node in the network is approximated by an exponential distribution. The efficiency of the exponential distribution and its application to the call center design is illustrated by simulation. For a good review of the queueing models of call centers, we refer the reader to [7].

There are only a few papers on customer contact center modeling. Recent papers by Armony and Maglaras [1, 2] focus on a customer contact center that offers two modes of service. The arriving customers are informed about the delay, and the contact center is modeled as a two-class M/M/r queueing system with state-dependent arrival rates. Armony and Maglaras [1] proposed an estimation scheme for the anticipated delay time based on the heavy traffic regime and approximated the system performance. They presented a staffing rule that picks the minimum number of agents in [2]. Whitt [10] addresses the many challenging research issues surrounding customer contact center modeling.

3. Contact Center Description

In this paper, we develop an approximate queueing model of a contact center example that was the subject of an extensive simulation study in Greve et al. [3]. We consider a contact center with two types of arriving e-mail. The e-mails received are identified by software as new or previously processed. If an e-mail has been previously processed, then the agent who previously processed the e-mail is identified, and the e-mail is routed to that agent. If e-mails are new, they are routed with equal probability to one of the agents. Once the e-mail is routed to an agent, the agent preprocesses it. Preprocessing involves determining the e-mail type and identifying the history of the e-mail. The history of the e-mail can be any one of the following; the e-mail can be brand new, processed by the same agent, or processed by a different agent. From this information the agent determines whether to process the e-mail or to forward the e-mail to another agent. The time required to process an e-mail is random and is directly influenced by both the e-mail type and history. If the e-mail response is sufficient to resolve the customer's problem, the e-mail leaves the system permanently. If the e-mail response is not sufficient to resolve the customer's problem, the e-mail returns to the system after a random delay. This random delay represents the time required for the customer to receive the agent's response and send a reply. The flowchart depicting the e-mail handling logic is shown in Figure 1. The notation used is summarized below.

Notation

c	category of external e-mail $c = 1, 2, \dots, C$
a_l	Agent l $l = 1, 2, \dots, A$
i, k, j	index for current, previous and next agent $i = 1, 2, \dots, A; k = 0, 1, \dots, A; j = 1, 2, \dots, A;$ $k=0$ represents a new e-mail
S_i	random variable that represents the preprocessing time at agent a_i
θ_i	$E(S_i)$
$f_i^s(x)$	probability density function (pdf) of S_i
$c_{pre(i)}^2$	SCV (=variance/mean ²) of S_i
$T_{i,c,k}$	Random variable denoting the processing time at agent a_i for email type c previously processed by agent a_k
$\tau_{i,c,k}$	$E(T_{i,c,k})$
$\tau_{proc(i)}$	Mean service time at a node i given e-mail type c
$c_{ser(i)}^2$	SCV of the service time at agent i
λ_c	arrival rate of e-mail belonging to category c
p_i	probability that an e-mail is routed to agent a_i , $\sum p_i = 1$
$p_{ij}(c,k)$	probability that agent a_i forwards an e-mail to agent a_j
$q_i(c,k)$	probability that an e-mail processed by agent a_i ends in a resolution. $(1-q_i(c,k))$ is the probability that an e-mail processed by agent a_i results in a repeat e-mail from the customer after a random delay.

4. Assumptions and Parameter Values

- The inter-arrival time of the e-mails follows an Exponential distribution.
- The e-mails are served according to a FCFS queueing discipline.
- For an unresolved problem, the e-mail enters the system with a delay independent of the prior e-mail.
- The pre-processing time is an independent random variable following a uniform distribution.
- The processing time is an independent random variable following a general distribution.

Arrival and Preprocessing Time Parameter Values

C = 2 e-mail types;

A = 3 agents

Arrival rates: $\lambda_1 = 4.2/\text{hr}$, $\lambda_2 = 5.25/\text{hr}$,

Preprocessing time: $\theta_i = 0.055\text{ hr}$, $i = 1, 2$ and 3 , $f_i^S(x) = \text{Uniform in the range } (0.01, 0.1)$

Table 1: Mean service time ($\tau_{i,c,k}$)

(c,k)	c = 1				c = 2			
	k=0	k=1	k=2	k=3	k=0	k=1	k=2	k=3
1	0.25	0.20	0.25	0.25	0.20	0.15	0.20	0.20
2	0.15	0.15	0.10	0.15	0.10	0.10	0.05	0.10
3	0.20	0.20	0.20	0.15	0.15	0.15	0.15	0.10

Table 2: Resolution probabilities ($q_i(c, k)$)

(c,k)	c = 1				c = 2			
	k=0	k=1	k=2	k=3	k=0	k=1	k=2	k=3
1	0.70	0.80	0.75	0.75	0.75	0.85	0.80	0.80
2	0.80	0.85	0.90	0.85	0.85	0.90	0.95	0.90
3	0.75	0.80	0.80	0.85	0.80	0.85	0.85	0.90

Table 3: Forwarding probabilities ($p_{ij}(c, k)$)

(i,j)	c = 1				c = 2			
	k=0	k=1	k=2	k=3	k=0	k=1	k=2	k=3
(1,2)	0.20	0.20	0.10	0.10	0.10	0.02	0.05	0.05
(1,3)	0.10	0.10	0.05	0.05	0.00	0.00	0.03	0.03
(2,1)	0.00	0.01	0.00	0.01	0.05	0.02	0.00	0.02
(2,3)	0.01	0.02	0.01	0.02	0.05	0.03	0.00	0.03
(3,1)	0.01	0.02	0.02	0.02	0.04	0.04	0.04	0.01
(3,2)	0.15	0.07	0.07	0.15	0.07	0.03	0.03	0.02

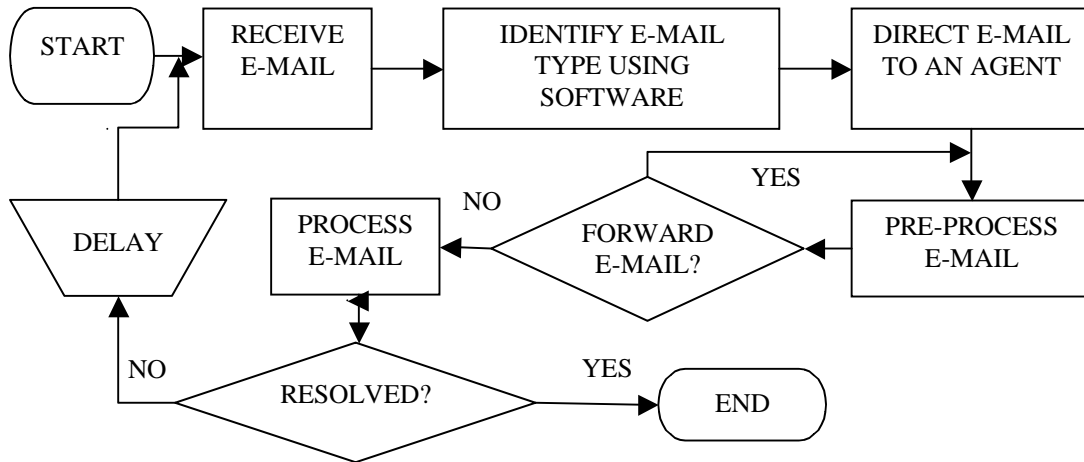


Figure 1: E-Mail handling logic

5. Modeling the E-Mail Contact Center Using an Open Queueing Network

Over the years queueing models have emerged as an important tool in stochastic modeling and are widely used in analyzing the performance of manufacturing systems, communication systems and service systems. There was a paradigm shift in mid-eighties from “exact analysis of an approximate model” to “approximate analysis of a more appropriate model” [11]. Until the eighties, queueing analysis was based on the well-known product-form solution. The product-form solution was based primarily on Poisson arrivals and Exponential service times, and was focused on obtaining an exact solution to an approximate model. The newer, parametric decomposition (PD) approach, based on two-moment queueing approximations (using mean and squared-coefficient of variation ((SCV) = Variance/Mean²)), makes it possible to model a variety of real-world problems. The PD approach aggregates the number of customer classes in the model to a single class, calculates the performance measures at each node, disaggregates the results, and reports the results for each customer class. For a detailed account of the PD approach, the reader is referred to [11].

The situation explained in Section 3 can be modeled as an open queueing network model where the nodes represent the agents and customers represent e-mails. However, there are many features that cannot be readily modeled using the available network models. For example, the routing scheme described in the previous section is not Markovian.

The routing probabilities and the service time distributions depend on the history of the e-mail. We first develop a simple aggregation scheme to convert the class and node-specific detailed routing probabilities and service time information into an approximately equivalent single-class, node-specific service time parameters and Markovian routing probabilities. This allows us to apply the PD approach to analyze the resulting open network model.

The input for the PD approach includes the number of nodes, the number of servers at each node, the SCVs of the interarrival and service time distributions at each node, and the Markovian routing probability matrix. The open network model has four nodes, one for each of the agents and an infinite-server node to model the random delay that represents the time needed for a customer to reply. Once the unresolved e-mail arrives at the delay node, the e-mails are delayed according to an Exponential distribution with a mean of four time units. The unresolved e-mail, after being held at the delay node, once again enters the system and is routed with equal probability to one of the agents (nodes) with equal probabilities. This process is continued until the e-mail gets resolved.

5.1. Calculation of the Markovian routing probabilities

We calculate the Markovian routing probability by using a two-level weighted average scheme. For each class, we first compute the forwarding probability for each agent by simply averaging the forwarding probabilities for new and previously processed e-mails. The routing probability from agent i to agent j for type c e-mail, $P_{ij}(c)$, is given by

$$P_{ij}(c) = \frac{\sum_{k=0}^A P_{ij}(c, k)}{A+1} \quad i = 1, 2, \dots, A; j = 1, 2, \dots, A; i \neq j \quad (1)$$

Now we can aggregate the class-specific routing probabilities by giving weights that are equal to the class-specific arrival rates. Hence, the routing probability from agent i to agent j , P_{ij} , is given by

$$P_{ij} = \frac{\sum_{c=1}^C \lambda_c P_{ij}(c)}{\sum_{c=1}^C \lambda_c} \quad i = 1, 2, \dots, A; j = 1, 2, \dots, A; i \neq j \quad (2)$$

The routing probability from agent i , ($i = 1, 2, \dots, A$) to the delay node, numbered $A+1$, is obtained in similar manner. The probability that an e-mail of type c processed by agent i , ($i = 1, 2, \dots, A$) goes to the delay node is given by

$$P_{i,A+1}(c) = \frac{\sum_{k=0}^A \left[\left(1 - \sum_{\substack{j=1 \\ j \neq i}}^A P_{ij}(c, k) \right) \cdot (1 - q_i(c, k)) \right]}{A+1} \quad i = 1, 2, \dots, A \quad (3)$$

$$P_{i,A+1} = \frac{\sum_{c=1}^C \lambda_c \cdot P_{i,A+1}(c)}{\sum_{c=1}^C \lambda_c} \quad i = 1, 2, \dots, A \quad (4)$$

Our assumption is that a reply or follow-up e-mail from the customer is equally likely to go to any one of the agents. Hence, the routing probability from the delay node to agent i is given by

$$P_{A+1,i} = \left(\frac{1}{A} \right) \quad i = 1, 2, \dots, A \quad (5)$$

5.2. Calculation of the mean service time at node i

The overall service time is the preprocessing time plus the actual processing time needed by the agent if the agent decides to process the e-mail himself/herself. Once again, we calculate the mean service time by using a two-level weighted average scheme. For each class, we first compute the mean processing time for each class by averaging the mean processing times for new and previously processed e-mails. The mean service time for a class c e-mail at node/agent i , $\tau_i(c)$, is given by

$$\tau_i(c) = \theta_i + \left[\left(1 - \sum_{\substack{j=1 \\ j \neq i}}^A (P_{ij}) \right) \cdot \left(\frac{\sum_{k=0}^A \tau_{i,c,k}}{A+1} \right) \right] \quad i = 1, 2, \dots, A \quad (6)$$

Now we can aggregate the class-specific mean service times by giving weights that are equal to the class specific arrival rates. Hence, the mean (overall) service time at node/agent i is given by

$$\tau_i = \frac{\sum_{c=1}^C \lambda_c \tau_i(c)}{\sum_{c=1}^C \lambda_c} \quad i = 1, 2, \dots, A \quad (7)$$

5.3. Squared coefficient of variation of the service time distribution at node i

The calculation of the service time SCV also follows a two-stage process. For each class, we first compute the variance of the overall service time by using the fact that the effective processing time distribution is a mixture of processing time distributions for new and previously processed e-mails. The SCV of overall service time is obtained by once again using the fact that it is a mixture of class-specific distributions. Because of space limitations we present only the final expression

$$\sum_{c=1}^C \lambda_c \cdot (\tau_i^2(c_s^2(i) + 1)) = \sum_{c=1}^C \lambda_c \cdot \left(c_{pre(i)}^2 \theta_i^2(c) + \left(1 - \sum_{\substack{j=1 \\ j \neq i}}^A (P_{ij}) \right) \cdot \left(\frac{\sum_{k=0}^A \tau_{i,c,k}^2 (c_{ser}^2(i) + 1)}{(A+1)} - \tau_{proc(i)}^2(c) \cdot \left(1 - \sum_{\substack{j=1 \\ j \neq i}}^A (P_{ij}) \right) \right) + \tau_i^2(c) \right) \quad (8)$$

5.4. Aggregation results for the contact center model

Using the expressions derived in sections 5.1 through 5.3, we compute the parameter values for a general open queueing network model with Markovian routing. The detailed data presented in Section 4 results in the output shown in tables 4 and 5.

Table 4: Input parameters for the open queueing network model

Nodes	No of Servers	Arrival Rate Per hour	Arrival SCV	Mean Service time in hours	Service SCV		
					Expo	Hyper Expo	Erlang
1	1	3.15	1	0.234	1.006	3.170	0.473
2	1	3.15	1	0.160	0.614	2.090	0.245
3	1	3.15	1	0.199	0.813	2.640	0.356
4		0	0	4.000	1.000	1.000	1.000

Table 5: Routing matrix

Node	1	2	3	4
1	0.000	0.097	0.042	0.189
2	0.015	0.000	0.022	0.131
3	0.026	0.070	0.000	0.155
4	0.333	0.333	0.333	0.000

6. Analytical and Simulation Results for the Contact Center Model

The open queueing network model was solved using the Rapid Analysis of Queueing Systems (RAQS) package. RAQS is a software package for analyzing general queueing network models based on a two-moment framework [5]. The contact center simulation model was developed using Arena 7.0. Simulation estimates represent averages over ten replications. Each replication simulated 9,240 hours of operation with a warm up of 840 hours. The analytical and simulation results are shown in figures 2, 3 and 4 for three processing time distributions – Erlang (SCV = 0.25), Exponential, and Hyper Exponential (SCV = 4.0).

In predicting the average resolution time and the average number of e-mails in the system, the analytical model consistently underestimates the simulation results with a significant relative error (30-35%). It is our conjecture that the major source of error in the analytical model is the simple-minded aggregation scheme that we employed for calculating the queueing network model parameters. From a queueing network point of view, the dependence of the routing and processing schemes on the history of the e-mails is a non-standard feature and cannot be easily handled. We believe that a more detailed modeling of the routing scheme (e.g., using a discrete time Markov chain model with

an appropriate state space) might lead us to better aggregation procedures. We are currently pursuing this and other ideas to improve the accuracy of the analytical results.

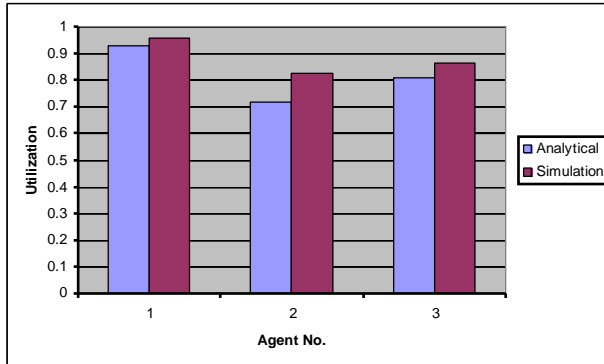


Figure 2: Agent utilization

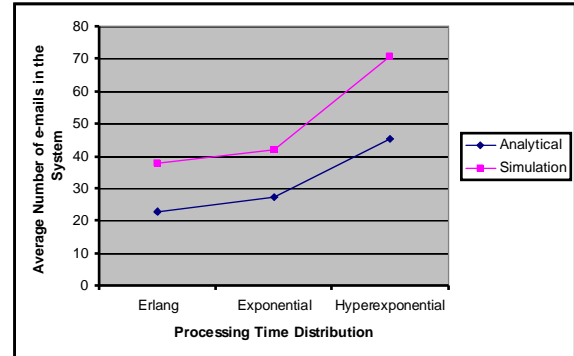


Figure 3: Average number of e-mails in the system

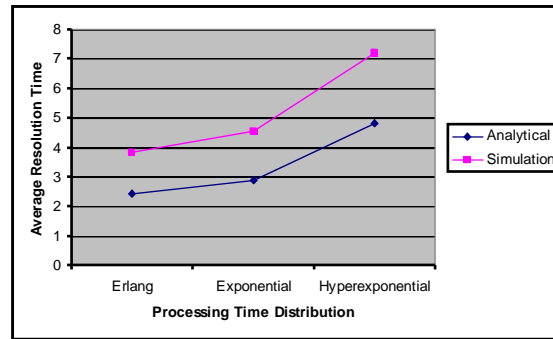


Figure 4: Average resolution time

7. Conclusions

In this paper, we presented models and results of our exploratory research in the area of queuing network models of customer contact centers. The value of analytical models is in their ability to support rapid what-if analyses that can assist the decision-maker in designing and improving real-world systems such as customer contact centers. The complexity of the system modeled in this paper clearly exceeds the capability of the currently available queuing network models. The preliminary results have helped identify certain areas of research that, if successful, could add to the current knowledge in queuing network modeling.

References

1. Armony, M. and Maglaras, C., 2003, "Contact Centers with a Call-Back Option and Real-Time Delay Information," <http://www-1.gsb.columbia.edu/faculty/cmaglaras/papers/cc-dynamic>.
2. Armony, M. and Maglaras, C., 2002, "On Customer Contact Centers with a Call-back Option: Customer Decisions, Routing Rules, & System Design," <http://www1.gsb.columbia.edu/faculty/cmaglaras/papers/static.pdf>.
3. Greve, R., Sharda, R., Kamath, M., and Kadam, A., 2004, "Modeling and Analysis of Email Management for Improved Customer Relationship Management," *International J. of Simulation and Process Modeling* (to appear).
4. <http://www.callcenternews.com/>
5. <http://www.okstate.edu/cocim/raqs/>
6. Koole, G., 2000/2001 "The Mathematics of Call Centers," Research Highlight, Annual Report, Stieltjes Institute.
7. Koole, G. and Mandelbaum, A., 2002, "Queueing Models of Call Centers – An Introduction," *Annals of Operations Research*, 113, 41-59.
8. Koole, G., Pot, A., and Talim, J., 2003, "Routing Heuristics for Multi-Skill Call Centers," *Proceedings of the Winter Simulation Conference*, December 7 -10, New Orleans, Louisiana, USA, 1813-1816.
9. Koole, G. and Talim, J., 2000, "Exponential Approximation of Multi-Skill Call Centers Architecture," *Proceedings of QNETs*, Ilkley, UK, 23/1-10.
10. Whitt, W., 2002, "Stochastic Models for the Design and Management of Customer Contact Centers: Some Research Directions," <http://www.columbia.edu/~ww2040/full.pdf>.
11. Whitt, W., 1983, "The Queueing Network Analyzer," *The Bell System Technical Journal*, 62(9), 2779-2815.