Top

Volume 12, Number 1, 91-110
June 2004

REPRINT

R. Srinivasan, J. Talim and J. Wang

Performance Analysis of a Call Center with Interactive Voice Response Units

Published by
Sociedad de Estadística e Investigación Operativa
Madrid, Spain

Top

Volume 12, Number 1
June 2004

Published by
Sociedad de Estadística e Investigación Operativa
Madrid, Spain

Sociedad de Estadística e Investigación Operativa
Top (2004) Vol. 12, No. 1, pp. 91–110

Performance Analysis of a Call Center with Interactive Voice Response Units

Raj Srinivasan, Jérôme Talim
Mathematical Sciences Group, University of Saskatchewan
106 Wiggins Road, Saskatoon SK S7K 5E6 Canada
E-mail: raj@poisson.usask.ca

Jinting Wang
Department of Mathematics, Beijing Jiaotong University
Beijing 100044 P.R. China
E-mail: wjting@yahoo.com

Abstract

A Call center may be defined as a service unit where a group of agents handle a large volume of incoming telephone calls for the purpose of sales, service, or other specialized transactions. Typically a call center consists of telephone trunk lines, a switching machine known as the automatic call distributor (ACD) together with a voice response unit (VRU), and telephone sales agents. Customers usually dial a special number provided by the call center; if a trunk line is free, the customer seizes it, otherwise the call is lost. Once the trunk line is seized, the caller is instructed to choose among several options provided by the call center via VRU. After completing the instructions at the VRU, the call is routed to an available agent. If all agents are busy, the call is queued at the ACD until one is free. One of the challenging issues in the design of a call center is the determination of the number of trunk lines and agents required for a given call load and a given service level. Call center industries use the Erlang-C and the Erlang-B formulae in isolation to determine the number of agents and the number of trunk lines needed respectively.

In this paper we propose and analyze a flow controlled network model to capture the role of the VRU as well as the agents. Initially, we assume Poisson arrivals, exponential processing time at the VRU and exponential talk time. This model provides a way to determine the number of trunk lines and agents required simultaneously. An alternative simplified model (that ignores the role of the VRU) will be to use an $M|M|S|N$ queueing model (where S is the number of agents and N is the number of trunk lines) to determine the optimal S and N subject to service level constraints. We will compare the effectiveness of this simplified model and other approximate methods with our model. We will also point out the drawbacks of using Erlang-C and Erlang-B formulae in isolation.

Key Words: Call center, flow controlled Jackson networks, Erlang formulae.

AMS subject classification: 60K25.

Contributed paper for the First Madrid Conference on Queueing Theory, held in Complutense University, Madrid, Spain, July 2-5, 2002

Manuscript received: September 2001. Final version accepted: October 2003.

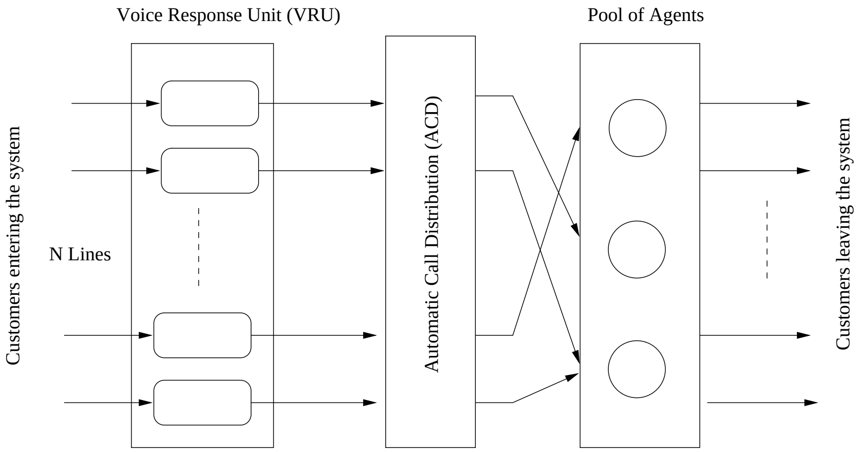
Voice Response Unit (VRU)
Pool of Agents
Customers entering the system
N Lines
Automatic Call Distribution (ACD)
Customers leaving the system

Figure 1: *Call Center with Voice Response Unit and Pool of Agents*

1 Introduction

A call center is a place where a group of agents handle large volume of incoming and outgoing calls for the purpose of sales, service, technical support, marketing or other specialized transactions. Call centers are usually classified as either inbound or outbound call centers. Inbound call centers are typically associated with businesses such as airline reservation companies, companies that sell their products through catalogue, companies that provide product information, invoice tracking, sales and service help and technical assistance. In an inbound call center customers either through a 1-800 number or a special number place a call to an agent to receive service, whereas in an outbound call center agents initiate calls to customers for tele-marketing or to provide information on sale of a new product. But some of the inbound call centers may initiate outgoing calls or even provide response via email or respond to web based calling. In this paper we will only concentrate on some key issues related to inbound call centers.

An inbound call center consists of telecommunications resources (phone lines, faxes, e-mail servers and possibly web applications) and telephone service representatives possibly with multiple skills. Even though more and more call centers provide services via faxes, email and web applications, services based on phone calls form an integral part of their operation since they involve real time interaction with their customers. The design and engineering of a call center involves the determination of appropriate telecommunication and human resources to operate it efficiently. In particular, in designing an inbound call center one should determine the number of trunk lines and the number of agents required to meet some pre-specified quality of service.

Typically a call center is made up of telephone trunk lines, a switching machine known as the automatic call distributor (ACD) together with a voice response unit (VRU) and agents to handle the volume of the incoming calls. See Figure 1. Customers usually dial a special number provided by the call center; if a trunk line is free, the customer seizes it, otherwise the call is lost. Once the trunk line is seized, the caller is instructed to choose among several options provided by the call center via VRU. After completing the instructions at the VRU, it is routed to an available agent. If all agents are busy, the call is queued at the ACD until one is free. The VRU acts like a filter in the sense that it determines the specific needs of a customer before it can be routed to an appropriate agent. The sophisticated version of the VRU is an interactive voice response unit (IVRU) which can provide a range of automatic services. The customers serviced by the IVRU may leave the system without requiring the service of the agents. Telephone banking and airline reservation systems are two examples where most of the calls are handled entirely by the IVRU.

One of the challenging issues in the design of a call center is the determination of the number of trunk lines and agents required for a given call load and a given service level. Service level is a key performance metric of a call center. The call center industry uses several forms of service level in their operation. One of the most commonly used service level is defined as "X percent of the calls answered in Y seconds", such as 80% of the calls answered in 20 seconds. This metric is related to the amount of time a customer spends at the call center. It also uses the percentage of busy signals as an another metric to determine the availability of the trunk lines. Ignoring the role of VRU, call center industries use the Erlang-C and the Erlang-B formulae in isolation to determine the number of agents and the number of trunk lines needed respectively Cleveland and Mayben (1997). In this standard approach, firstly, based on the historical call load over a typical 30-minute or a 15-minute period and the estimates on average talk time, one uses Erlang-C table to determine the number of sales agents required to meet the desired waiting time probability. Note that the use of

Erlang-C tables assumes that there is an infinite waiting room for the calls to wait whereas in a call center, the maximum number of calls that can wait in the system is limited by the number of trunk lines. Once the number of agents required to meet the desired waiting time probability is found, to determine the number of trunk lines needed to satisfy the limit on the amount of lost calls, one uses the Erlang-B formula with the service time as the sum of average talk time and the mean waiting time obtained from the Erlang-C formula. To use the Erlang-B formula correctly, one has to know the total holding time of the call, which is comprised of the amount of time spent waiting for an agent and the talk time. The amount of time waiting for an agent depends on the number of sales agents available. Using the average waiting time from the Erlang-C formula will only lead an approximation. Apart from calculating the number of agents and the number of trunk lines required separately by using Erlang-C and Erlang-B formulae, this traditional approach does not take into account the role played by the VRU or IVRU, and call abandonment. We refer to abandonment as the fact that delayed customers may judge that the service they seek is not "worth" the wait, become impatient and hang up before they are served. These customers are called "impatient customers" , see, for example, Garnett et al. (1999) for further discussion. Call center industries use some modification of the Erlang-C formula to include call abandonment rate, and one such formula is referred to as Merlang formula for modified Erlang, Cleveland and Mayben (1997). Other work related to this topic includes Borst and Serri (2000) where the call center design is based on quality of service (in terms of waiting time) provided to customers. Garnett et al. (1999) introduce the concept of impatient customers in their model. Koole and van der Sluis (1998) focus on the manpower scheduling aspect. Talim and Koole (2000) propose an approximation of multi-skill call centers with skill based routing.

One may suggest to model the call center as finite multi-server queues where arriving calls are immediately assigned to an available server. Such a model may not capture the role of the VRU. For example one could model the call center as an $M|M|S|N$ finite queue where S and N represent the number of agents and trunk lines respectively. For a pre-specified loss probability and a bound on the waiting time one could numerically calculate the number of agents and the trunk lines needed. It will be of interest to compare by how much the traditional approach over estimates the number of trunk lines and the agents required. Since the VRU plays a significant role in determining the selection of the appropriate agent for a each call, and provides part of the service, we present and analyze a model of a call center that includes the role of the VRU. In the presence of VRU, we propose a scheme to determine the number of trunk lines and agents needed to meet some specified service level.

This paper is organized as follows. The flow controlled serial network model that captures the role of the VRU explicitly is presented in Section 2. In Section 3, we provide the main results of this paper where a formula for the loss probability and the distribution of the waiting time are provided. In Section 4 we propose several approximations that may be efficient under some conditions. Section 5 is devoted to numerical examples where we illustrate the efficiency of our flow controlled serial network model and compare its performance to other approximation models. Finally, in Section 7 we provide the proofs of our main results.

2 The Model

In this paper, we consider a Markovian model of a call center in which calls arrivals are modelled by a Poisson process with constant rate λ. The call center consists of N trunk lines and S agents ($N \geq S$). Upon arrival (when a trunk line is free), a call spends some time at the VRU before it reaches an available agent. We assume that the VRU processing times are independent and identically distributed exponential random variables with rate θ. After finishing the VRU process, a call may leave the system and no longer interferes with the center; or it may request an assistance from an agent. In this case, it is either assigned to an available agent or held at ACD until an agent becomes free. Talk times with an agent are independent and identically distributed exponential random variables with rate μ, and are independent of the arrival times and the VRU processing times. Once a call completes its transaction with an agent, it releases both the trunk line and the agent simultaneously. We assume that calls that find the system full, i.e. all the trunk lines are busy are lost. We do not consider call abandonment in this paper. Our objective in this paper is determine the number of trunk lines(N) and the agents(S) required to meet service levels specified in terms of the probability of a wait for an agent and the probability of blocking are met.

The call center model we consider in this paper consists of two multi-

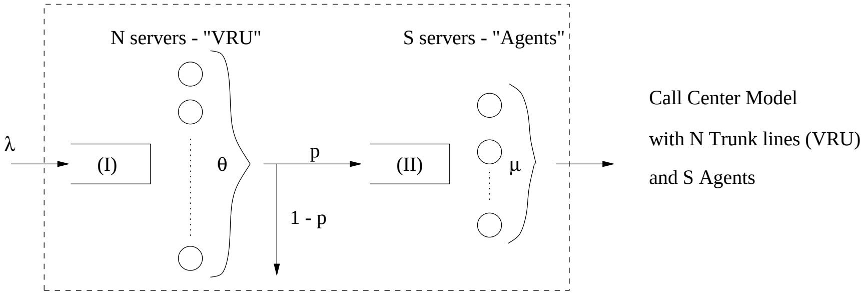

Figure 2: *Call Center Model*

server queues connected in series. See Figure 2. The first one represents the VRU processor. This VRU processor can handle at most N jobs at a time, where N represents the total number of trunk lines available. The second queue represents the agents pool where there are S agents to handle the incoming calls. Typically call center employs less agents than the number of trunk lines available, therefore $S \leq N$. Arriving calls can enter the VRU processor only when the total number of jobs at the VRU processor and the agents pool is strictly less than the number of trunk lines, N; otherwise the call is lost. Once a call is processed by the VRU, it leaves the system with probability $1-p$ and releases a trunk line, or it proceeds to the agents pool with probability p. If an agent is free, the call is served by the one of the agents, otherwise it waits for an agent and once the call is processed by an agent it releases both the agent and a trunk line. Note that in this model there is no queue at the VRU processor, whereas some calls may have to wait for an agent at the agents pool if all the agents are busy. Let $\{Q_1(t), Q_2(t)\}$ represent the number of calls at the VRU processor and the agents pool respectively. Note that $Q_1(t) + Q_2(t) \leq N$ for all $t \geq 0$. It is easy to see that $\{Q_1(t), Q_2(t)\}$ is a finite state Markov chain. In fact the model considered here can be thought of as a flow controlled Jackson network (FJN) which has a product form solution and we will use this fact to derive the stationary distribution and other related quantities of interest in Section 7. It is also interesting to point out that when there are infinitely many trunk lines, the network is simply an infinite server queue that feeds into an $M|M|S$ queue. This simplified assumption leads to a reversible network with a product-form stationary distribution and the call center model considered here is merely its conditioning on having no more than

N jobs in the system.

3 Results

In this section, we summarize the results of the analysis of our call center model. The first result provides the stationary probabilities of having a total of k calls in the center; the calls may be dealing with the VRU process, waiting for an agent or being served. From these probabilities, one can easily derive the loss probability. The second result provides the distribution of the waiting time, the time spent by a call in the trunk line, while waiting for an agent to become available.

Theorem 3.1. *The stationary probability P_k for $k \leq N$ that there are exactly k calls in the system (processed by the VRU, waiting for service or being serviced by an agent) is given by*

$$\begin{aligned} P_k = \kappa \Bigg(& \frac{\lambda^k}{k!}\left(\frac{1}{\theta}+\frac{p}{\mu}\right)^k \\ & + \sum_{j=S+1}^{k} \frac{1}{(k-j)!}\left(\frac{1}{S!S^{j-S}}-\frac{1}{j!}\right)\left(\frac{\lambda}{\theta}\right)^{k-j}\left(\frac{p\lambda}{\mu}\right)^j \mathbb{I}\{k>S\}\Bigg), \end{aligned} \tag{3.1}$$

where $\mathbb{I}\{k>S\}$ is the indicator function, and

$$\begin{aligned} \kappa := \Bigg(& \sum_{k=0}^{N} \frac{\lambda^k}{k!}\left(\frac{1}{\theta}+\frac{p}{\mu}\right)^k \\ & + \sum_{k=S+1}^{N}\sum_{j=S+1}^{k} \frac{1}{(k-j)!}\left(\frac{\lambda}{\theta}\right)^{k-j}\left(\frac{p\lambda}{\mu}\right)^j\left(\frac{1}{S!S^{j-S}}-\frac{1}{j!}\right)\Bigg)^{-1}. \end{aligned}$$

One can apply these formulae to derive P_N, the probability of having all trunk lines busy. This is also the loss probability due to the PASTA (Poisson Arrivals See Time Averages) property. See Wolff (1982).

Lemma 3.1. *The probability that an arriving call finds all trunk lines busy* *is given by:*

$$P_N = \kappa \Bigg(\frac{\lambda^N}{N!} \left(\frac{1}{\theta} + \frac{p}{\mu} \right)^N + \sum_{j=S+1}^{N} \frac{1}{(N-j)!} \left(\frac{1}{S! S^{j-S}} - \frac{1}{j!} \right) \left(\frac{\lambda}{\theta} \right)^{N-j} \left(\frac{p\lambda}{\mu} \right)^j \mathbb{I}\{N > S\} \Bigg). \quad (3.2)$$

The second result of our analysis deals with the distribution function of this waiting time.

Theorem 3.2. *Let T denote the waiting time of a call defined as the time spent by a call waiting for an agent to become available after finishing the VRU process. Let $W(t) := P(T \leq t)$ be the distribution function of the waiting time. Then*

$$W(0) \;=\; \sum_{k=1}^{N} \sum_{j=0}^{\min(k,S)-1} \chi(k,j), \quad (3.3)$$

and for all $t > 0$,

$$W(t) \;=\; 1 - \sum_{k=S+1}^{N} \sum_{j=S}^{k-1} \chi(k,j) \sum_{l=0}^{j-S} \frac{(\mu S t)^l e^{-\mu S t}}{l!} \quad (3.4)$$

where

$$\chi(k,j) \;:=\; \frac{(k-j) \cdot \pi(k-j,j)}{\sum_{l=1}^{N} \sum_{m=0}^{l-1} (l-m) \cdot \pi(l-m,m)},$$

and

$$\pi(k,j) = \kappa \begin{cases} \frac{1}{(k-j)!j!} \left(\frac{\lambda}{\theta}\right)^{k-j} \left(\frac{p\lambda}{\mu}\right)^j & \text{if } 0 \leq j \leq S \\ \\ \frac{1}{(k-j)!S!S^{j-S}} \left(\frac{\lambda}{\theta}\right)^{k-j} \left(\frac{p\lambda}{\mu}\right)^j & \text{if } j > S. \end{cases}$$

3.1 Remarks

Once we know the loss probability P_N and the distribution of the waiting time, $W(t)$, for a fixed values of λ, θ, p and μ, the problem is to find the

the number of trunk lines (N) and the number of agents (S) such that the following constraints are satisfied:

$$P_N = \epsilon_1 \tag{3.5}$$

$$P(T \leq \tau) \geq \epsilon_2 \tag{3.6}$$

for a pre-specified ϵ_1, ϵ_2 and τ.

Before we proceed with the approximate methods and the proofs, we provide some remarks on some special cases of our model, when $p = 1$, i.e., when all calls require assistance from the agents pool after having processed through the VRU.

1. When the number of agents is equal to the number of trunk lines ($N = S$), the call center model is a $M/G/N/N$ loss system where the service time is the sum of two independent exponential random variables with rates θ and μ respectively representing the VRU processing time and the talk time. The loss probability P_N (3.2) reduces to

   $$P_N = \frac{\frac{\lambda^N}{N!}\left(\frac{1}{\theta}+\frac{1}{\mu}\right)^N}{\sum_{i=0}^{N} \frac{\lambda^i}{i!}\left(\frac{1}{\theta}+\frac{1}{\mu}\right)^i}$$

   which is exactly the Erlang's loss formula. More generally, the probabilities P_i in (3.1) in this particular case also reduces to the corresponding formula for the $M/G/S/S$ queueing system. When $N \neq S$, our system is a finite capacity $M|G|N|N$ queue.

2. If the VRU processing time is negligible when compared to the talk time of the agents, then letting $\theta \to \infty$, our model simplifies to a $M/M/S/N$ queue. The stationary probabilities P_i in (3.1) reduces to the well known formula for the $M/M/S/N$ queue.

3. Similarly, when the service time of the agents goes to 0, or μ goes to infinity, the system is equivalent to the $M/M/S/S$ loss system. Only the action of the VRU is taken into account. This might be a very unrealistic case since we are assuming that the agents are infinitely faster. By letting $\mu \to \infty$, P_i's in (3.1) agrees with the well known Erlang-B formula.

4. The expression of the distribution of the waiting time (3.4) is similar to the one for the $M/M/S/N$ queue. See Gross and Harris (1985).

5. Also note that when the number of trunk lines is much larger than the number of agents, then modeling the VRU processor as infinite server queue leads to an open Jackson network model. This model may be appropriate for a large call center where telecommunication costs are less significant than the cost towards agents.

4 Approximate Methods

Recall that our objective in the efficient design of a call center is to determine the number of trunk lines and the number of agents needed to satisfy some pre-specified service levels. Determining N and S given all other parameters can be only carried out numerically. For large call centers, the calculation of these two quantities is computationally challenging. In this section, we introduce several approximate methods of our call center that require less compuatational effort. Let $MMSN(\lambda, \mu, S, N)$ denote the $M|M|S|N$ queue. See Gross and Harris (1985) for the stationary distribution of the number in the system and the waiting time distribution function.

4.1 Simplifying the role of the VRU (SVRU)

If the VRU processing time negligible when compared to the talk time, one could ignore the role of the VRU completely and model the call center as an $M/M/S/N$ queue. This first approximation $SVRU_1$ is a $MMSN(\lambda, \mu, S, N)$ queue. When the processing time is significant this approximation will not be very good.

If the processing time is significant when compared to the talk time, then one can suggest the $M/G/S/N$ queueing model where the service time now is the sum of the processing time and the talk time. This $M/G/S/N$ queue will be further approximated by $MMSN(\lambda, \hat{\mu}, N, S)$ where

$$\frac{1}{\hat{\mu}} := \frac{1}{\theta} + \frac{1}{\mu}$$

We will denote this approximation by $SVRU_2$. Note that in this approximation we are only including the time experienced by the calls at the VRU

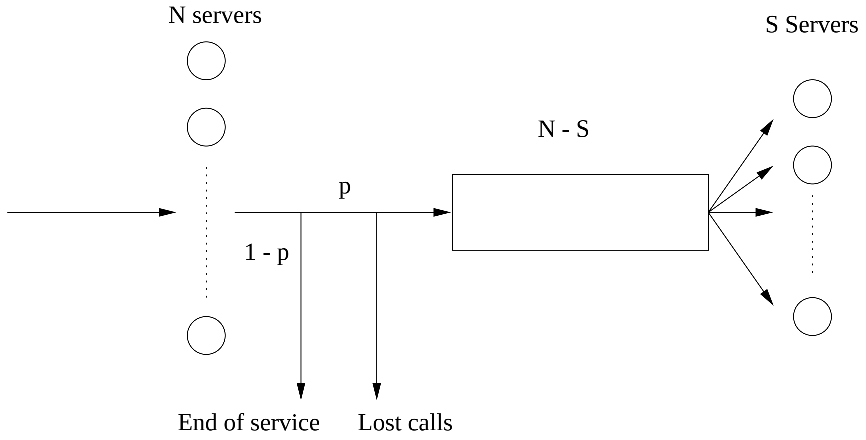

Figure 3: *MSQT Approximation of the model*

but not the actual functioning of the VRU. As we pointed out earlier multi server queues can not capture the role of the VRU.

4.2 Multi Server queues in tandem (MSQT)

Previous approximations may simplify too much the role of the VRU process. We propose in this section a more accurate model which is depicted in the Figure 3. This model deals with two multi server queues in tandem. The first block is an $M/M/N/N$ loss system or equivalently an $MMSN(\lambda, \theta, N, N)$ queue. This block aims to represent the VRU device. A portion of the departure process from it is routed to an $M/M/S/N$. This second block models the talk time with an agent and the possible wait of a call until an agent becomes available. The arrival rate to the $M/M/S/N$ queue is the departure rate from the $MMSN(\lambda, \theta, N, N)$ loss queue. It can be easily calculated as $p\hat{\lambda}$, where

$$\hat{\lambda} := \sum_{i=0}^{N} i\theta \ \hat{P}_i = \lambda \left(1 - \hat{P}_N\right).$$

Hence the second block is models as $MMSN(p\hat{\lambda}, \mu, S, N)$ queue. Note that in this approximation the role played by the trunk lines are explicitly modeled. Due to this fact, some calls proceeding to the pool of agents after processed by the VRU may be lost. This does not happen in our flow controlled serial network model.

5 Numerical Results

In order to illustrate the advantages of our model over the traditional method which uses the Erlang-B and Erlang-C formulae in isolation, we consider the example given in Cleveland and Mayben (1997), pages 91-104 with a slight modification. This example of a call center deals with a call load of 250 calls per half an hour period. The average talk time is estimated to be 180 seconds. This example adds 30 seconds to the service time as the average after call work time. Instead we add 100 seconds or 0.01 seconds to service time as the average VRU processing time. The former represents the case when the VRU processing time is significant and the later represents the case when the VRU processing time is insignificant. The probability of requesting service from an agent after a call is being processed by the VRU is varied from 0.1 to 1. When $p = 0.1$, calls rarely require the assistant of an agent, which might be typical in banking or credit card industries. When p is close to one, almost every call require the assistance of an agent. The service levels used here are defined in terms of the call blocking probability and the probability of waiting for an agent. The call blocking probability is set to 0.01, i.e., 1% blocking, and the probability of waiting for an agent less than 20 seconds must be as high as 0.8 which leads to *80% of the calls answered in* 20 *seconds*. Now the problem is to find the number of trunk lines, N and the number of agents, S such that the following service levels are met:

$$P(\text{loss}) = 0.01,$$

$$P(W < 20.0) = 0.8.$$

In Table 1, we compare the results of our method with that of other approximate methods. When one uses the traditional approach (EBC method), i.e. using the Erlang-B and Erlang-C formulae separately, the number of trunk lines needed is 38 and the number of agents required is 45. In this case the VRU processor is very slow. With our method (FJN), when $p = 1$ the number of trunk lines needed is 60 and the number of

$\theta = 0.01$			$\theta = 100$		
Method	N	S	Method	N	S
EBC	38	45	EBC	37	30
MMSN	60	43	MMSN	39	30
FJN			FJN		
$p = 0.1$	30	4	$p = 0.1$	37	4
$p = 0.5$	60	15	$p = 0.5$	37	15
$p = 0.9$	60	26	$p = 0.9$	38	26
$p = 1$	60	28	$p = 1$	39	30

Table 1: *Determination of N and S*

agents required is only 28 which is much smaller than the one calculated by the EBC method. Note that when $\theta = 100$, the VRU processor is faster, and for the case $p = 1$, our method and the EBC method provide similar results. When the VRU processor is faster, calls spend much less time at the VRU, and one can ignore the role played by the VRU. For other values of p, the number of agents required obtained by our method is much less than the number agents obtained the EBC method. With the EBC method there is no explicit way to introduce the probability of a call requiring an agent. Even if one includes the VRU processing time in this calculation as was done in this case, ignoring the behavior of the calls after they are served by the VRU processor will lead to over estimation of the number of agents required. For example, when $p = 0.5$, only half the calls require the service of an agent and this should reflect why only 15 agents are required (in both the cases) as opposed to 45 ($\theta = 0.01$) and 30 ($\theta = 100$) agents calculated by the EBC method. The $SVRU_2$ approximation denoted by MMSN also over estimates the the number of agents required when the VRU processing rate is very small. When the VRU processor is faster, this method compares very well with the EBC and the FJN ($p = 1$ case) methods. As with the EBC method, this method also can not explicitly handle the behavior of the calls after they have been processed by the VRU processor.

In Figure 4, we compare the results of the three approximation methods with our (FJN) method. The graphs provide the number of agents required for a given number of trunk lines that satisfy the waiting time service level. When $\theta = 100$, both the $SVRU_1$ and $SVRU_2$ methods over estimate the number agents required for a fixed number of trunk lines. Actually, they provide the same results since the VRU processing time is negligible. The MSQT approximation performs very well and actually it coincides with our method. When the VRU processing rate is very small ($\theta = 0.01$), the MSQT approximation compares very well to our method. The $SVRU_2$ approximation also performs reasonably well, whereas the $SVRU_1$ method overestimates the number of agents required since it completely ignores the role of the VRU processor.

6 Conclusion

In this paper, we introduced a simple queueing network model of call center which explicitly models the role of the VRU processor. The analysis of our model provide a way to calculate the number of trunk lines and the number of agents required simultaneously to meet some pre specified service levels. We also illustrated that how the traditional approach of using Erlang-B and Erlang-C formulae in isolation will lead to over estimation of the number of agents and under estimation of the number of trunk lines. Further work on extending our model to include call abandonment and retrial are in progress.

7 Appendix

This section is devoted to the calculation of the stationary probabilities and the distribution of the waiting time for the call center model.

7.1 Stationary probabilities

Since our call center model is a flow controlled Jackson network, by introducing a fictitious state dependent queue to capture the controlled arrivals, one can convert our model to three node closed Jackson network which is known to have the well known product form solution for its stationary distribution. (See Kleinrock (1975), pp. 150-152, or Gross and Harris (1985), pp. 236-243) The stationary probabilities $\pi(i, j)$ of having i calls at the VRU and j calls at the agents pool can be written in a product form as

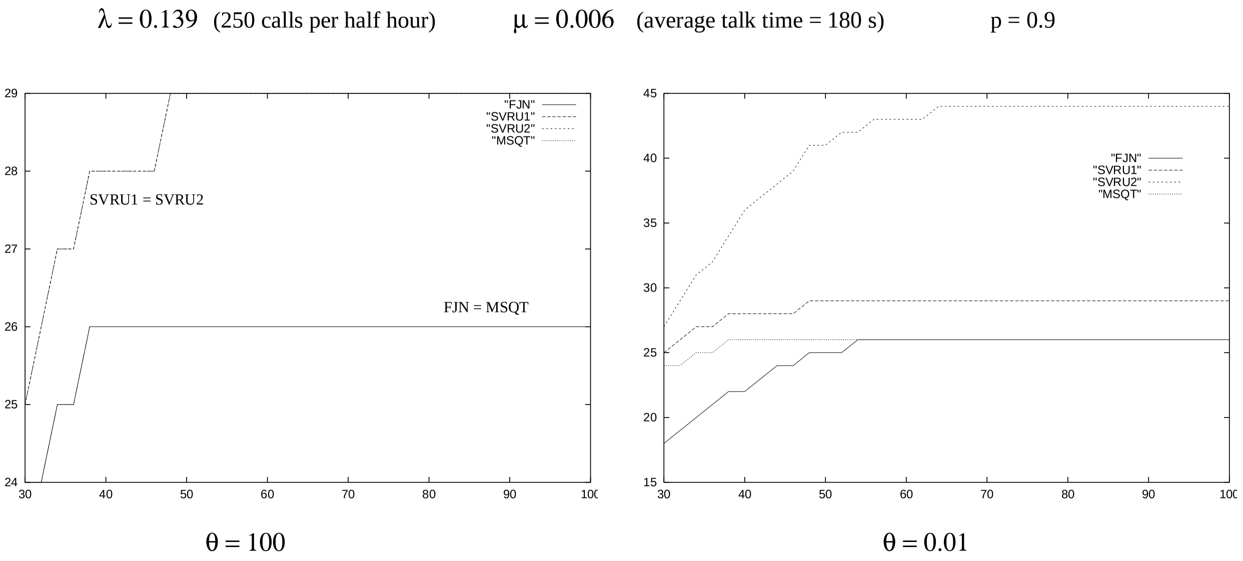

Figure 4: *Number of trunk lines vs. number of agents needed*

follows:

$$\pi(i,j) = \begin{cases} \kappa\frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i \frac{1}{\beta(j)}\left(\frac{p\lambda}{\mu}\right)^j & \text{if } 0 \le i+j \le N \\ \\ 0 & \text{otherwise,} \end{cases} \tag{7.1}$$

where

$$\beta(j) := \begin{cases} j! & \text{if } j \le S \\ \\ S!S^{j-S} & \text{if } j \ge S, \end{cases}$$

and

$$\kappa := \left(\sum_{0\le i+j \le N} \frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i \frac{1}{\beta(j)}\left(\frac{p\lambda}{\mu}\right)^j \right)^{-1}.$$

Let us rewrite the expression of the constant κ, by introducing the variable $k = i + j \in [0, N]$:

$$1/\kappa \;=\; \sum_{k=0}^{N}\sum_{j=0}^{k} \frac{1}{(k-j)!}\left(\frac{\lambda}{\theta}\right)^{k-j} \frac{1}{\beta(j)}\left(\frac{p\lambda}{\mu}\right)^j,$$

then

$$\kappa = \Bigg[\sum_{k=0}^{S}\sum_{j=0}^{k} \frac{1}{(k-j)!j!}\left(\frac{\lambda}{\theta}\right)^{k-j}\left(\frac{p\lambda}{\mu}\right)^j + \sum_{k=S+1}^{N}\left(\sum_{j=0}^{S} \frac{1}{(k-j)!j!}\left(\frac{\lambda}{\theta}\right)^{k-j}\left(\frac{p\lambda}{\mu}\right)^j + \sum_{j=S+1}^{k} \frac{1}{(k-j)!S!S^{j-S}}\left(\frac{\lambda}{\theta}\right)^{k-j}\left(\frac{p\lambda}{\mu}\right)^j \right)\Bigg]^{-1}.$$

Applying binomial formula to the first two sums in the second term of the equation yields

$$\kappa = \Bigg[\sum_{k=0}^{S} \frac{\lambda^k}{k!}\left(\frac{1}{\theta}+\frac{p}{\mu}\right)^k + \sum_{k=S+1}^{N} \frac{\lambda^k}{k!}\left(\frac{1}{\theta}+\frac{p}{\mu}\right)^k + \sum_{k=S+1}^{N}\sum_{j=S+1}^{k}\left(-\left(\frac{\lambda}{\theta}\right)^{k-j}\left(\frac{p\lambda}{\mu}\right)^j \frac{1}{(k-j)!j!}\right.$$

$$+\left(\frac{\lambda}{\theta}\right)^{k-j}\left(\frac{p\lambda}{\mu}\right)^{j}\frac{1}{(k-j)!S!S^{j-S}}\Bigg)\Bigg]^{-1}$$

$$=\Bigg(\sum_{k=0}^{N}\frac{\lambda^k}{k!}\left(\frac{1}{\theta}+\frac{p}{\mu}\right)^{k}$$

$$+\sum_{k=S+1}^{N}\sum_{j=S+1}^{k}\frac{1}{(k-j)!}\left(\frac{\lambda}{\theta}\right)^{k-j}\left(\frac{p\lambda}{\mu}\right)^{j}\left(\frac{1}{S!S^{j-S}}-\frac{1}{j!}\right)\Bigg)^{-1}.$$

From the stationary probabilities, one can deduce the probability P_k that there are exactly k calls in the system, or equivalently k occupied trunk lines using the relation

$$P_k := \sum_{j=0}^{k}\pi(k-j,j).$$

We will distinguish two cases:

1. $k \le S$:

$$P_k = \kappa\sum_{j=0}^{k}\frac{1}{(k-j)!j!}\left(\frac{\lambda}{\theta}\right)^{k-j}\left(\frac{p\lambda}{\mu}\right)^{j} = \kappa\frac{\lambda^k}{k!}\left(\frac{1}{\theta}+\frac{p}{\mu}\right)^{k}.$$

2. $k > S$:

$$P_k =\kappa\Bigg(\sum_{j=0}^{S}\frac{1}{(k-j)!j!}\left(\frac{\lambda}{\theta}\right)^{k-j}\left(\frac{p\lambda}{\mu}\right)^{j}$$

$$+\sum_{j=S+1}^{k}\frac{1}{(k-j)!S!S^{j-S}}\left(\frac{\lambda}{\theta}\right)^{k-j}\left(\frac{p\lambda}{\mu}\right)^{j}\Bigg)$$

$$=\kappa\Bigg(\frac{\lambda^k}{k!}\left(\frac{1}{\theta}+\frac{p}{\mu}\right)^{k}$$

$$+\sum_{j=S+1}^{k}\frac{1}{(k-j)!}\left(\frac{1}{S!S^{j-S}}-\frac{1}{j!}\right)\left(\frac{\lambda}{\theta}\right)^{k-j}\left(\frac{p\lambda}{\mu}\right)^{j}\Bigg).$$

The probability P_k that there are exactly $0 \leq k \leq N$ customers in the system is equal to:

$$P_k = \kappa \Bigg(\frac{\lambda^k}{k!} \left(\frac{1}{\theta} + \frac{p}{\mu} \right)^k \\ + \sum_{j=S+1}^{k} \frac{1}{(k-j)!} \left(\frac{1}{S! S^{j-S}} - \frac{1}{j!} \right) \left(\frac{\lambda}{\theta} \right)^{k-j} \left(\frac{p\lambda}{\mu} \right)^j 1\!\!\mathrm{I}\{k > S\} \Bigg)$$

7.2 Distribution of the waiting time

Remember that T is defined as the time spent by a call, just after it finishes the VRU process and until it gets serviced by an agent. For the sake of simplicity, we would say that the system in the state $(k, j), \quad 0 \leq j < k \leq N$, when it contains exactly k calls, j of which have already finished the VRU process. Let $\chi(k, j), 0 \leq j < k \leq N$ be the probability that the system is the state (k, j), when a call (among the $k - j$ customers) is about to finish its VRU process. Using the Bayes theorem, we derive:

$$\chi(k,j) := P(\text{system in state } (k-j,j) \mid \text{ call is about to leave VRU})$$

$$= \frac{P(\text{call is about to leave VRU} \mid \text{system state } (k-j,j)) \cdot \pi(k-j,j)}{\sum_{l=0}^{N} \sum_{m=0}^{l} P(\text{call is about to leave VRU} \mid \text{system state } (l-m,m)) \cdot \pi(l-m,m)}$$

$$= \frac{(k-j)\theta \cdot \pi(k-j,j)}{\sum_{l=0}^{N} \sum_{m=0}^{l} (l-m)\theta \cdot \pi(l-m,m)}$$

$$= \frac{(k-j) \cdot \pi(k-j,j)}{\sum_{l=1}^{N} \sum_{m=0}^{l-1} (l-m) \cdot \pi(l-m,m)}.$$

Let $W(0)$be the probability that a call starts its service immediately after leaving the VRU. Then it is given by

$$W(0) \quad = \quad \sum_{k=1}^{N} \sum_{j=0}^{\min(k,S)-1} \chi(k,j). \qquad (7.2)$$

For any $t > 0$, let $W(t) := P(T \leq t)$ be the distribution function of the waiting time. It follows that

$$W(t) := W(0) + \sum_{k=S+1}^{N} \sum_{j=S}^{k-1} \chi(k,j) P\big(\text{there are } j - S + 1 \text{ end of service} \\ \text{in} \leq t \;\mid\; \text{next call finishing the VRU process, system} \\ \text{state is } (k-j, j)\big).$$

When there are $j \geq S$ calls at the pool of agents (who have already finished the VRU process), all the agents are busy. The rate at which calls leave the system is $S\mu$. The distribution of the time of the $j - S + 1$ end of service is Erlang of type $j - S + 1$. Using this fact, the calculation reduces to

$$\begin{aligned} W(t) &= \sum_{k=S+1}^{N} \sum_{j=S}^{k-1} \chi(k,j) \int_0^t \frac{\mu S \, (\mu S x)^{j-S}}{(j-S)!} e^{-\mu S x} dx + W(0) \\ &= \sum_{k=S+1}^{N} \sum_{j=S}^{k-1} \chi(k,j) \left(1 - \sum_{k=0}^{j-S} \frac{(\mu S t)^k e^{-\mu S t}}{k!} \right) + W(0) \\ &= \sum_{k=S+1}^{N} \sum_{j=S}^{k-1} \chi(k,j) + W(0) - \sum_{k=S+1}^{N} \sum_{j=S}^{k-1} \sum_{l=0}^{j-S} \chi(k,j) \frac{(\mu S t)^l e^{-\mu S t}}{l!} \\ &= 1 - \sum_{k=S+1}^{N} \sum_{j=S}^{k-1} \chi(k,j) \sum_{l=0}^{j-S} \frac{(\mu S t)^l e^{-\mu S t}}{l!}. \end{aligned}$$

References

Borst S. and Serri P. (2000). Robust Algorithms for Sharing Agents with Multiples Skills. Preprint.

Cleveland B. and Mayben J. (1997). *Call Center Management on Fast Forward.* Call Center Press.

Garnett O., Mandelbaum A. and Reiman M. (1999). Designing a Call Center with Impatient Customers. Preprint.

Gross D. and Harris C.MN. (1985). *Fundamentals of Queueing Theory.* John Wiley.

Kleinrock L. (1975). *Queueing Systems, Volume 1: Theory.* John Wiley.

Koole G. and van der Sluis E. (1998). An optimal Local Search Procedure for Manpower Scheduling in Call Centers, Technical Report WS-501, Vrije Universiteit Amsterdam. `http://www.cs.vu.nl/ koole/papers/WS501.html`

Talim J. and Koole G. (2000). Exponential Approximation of Multi-Skill Call Centers Architecture, Proceedings of the Queueing Networks with Finite Capacity workshop, Ilkley, West Yorkshire, United Kingdom.

Wolff R.L. (1982). Poisson Arrivals See Time Averages. *Operations Research* 30, 223-231.

Top

Volume 12, Number 1
June 2004

CONTENTS