J. Appl. Math. & Informatics Vol. **26***(2008), No. 5 - 6, pp. 1021 - 1035*
Website: http://www.kcam.biz

APPROXIMATE ANALYSIS OF AN N-DESIGN CALL CENTER WITH TWO TYPES OF AGENTS

CHUL GEUN PARK*, DONG HWAN HAN AND KWANG HYUN BAIK

ABSTRACT. In this paper, we analyze an N-design call center with skill-based routing, in which one pool of agents handles two types of calls and another pool of agents handles only one type of calls. The approximate analysis is motivated by a computational complexity that has been observed in the direct stochastic approach and numerical method for finding performance measures. The workforce staffing policy is very important to the successful management of call centers. So the allocation scheduling of the agents can be considered as the optimization problem of the corresponding queueing system to the call center. We use a decomposition algorithm which divides the state space of the queueing system into the subspaces for the approximate analysis of the N-design call center with two different types of agents. We also represent some numerical examples and show the impact of the system parameters on the performance measures.

AMS Mathematics Subject Classification : 60K25, 68M20
Key words and phrases : Call center, N-design, skill-based, decomposition, performance

1. Introduction

An important type of customer's contact center is the call center, which serves phone calls. The number and size of call centers as well as the number of customers and agents grow dramatically[1]. For example, in Europe, the number of call center employees in 2000 was estimated by 600,000 in the UK and 200,000 in Netherlands and 280,000 in Germany[2,3].

In the most simple design of call centers, one type of customers is served by one type of agents. The prevalent model for performance analysis of these call centers is the M/M/N queue, frequently called as Erlang-C. Though Erlang-C

Received November 14, 2007. Accepted June 4, 2008. *Corresponding author.
This work was supported by the Korea Science and Engineering Foundation(KOSEF) grant funded by the Korea government(MOST) (No. R01-2007-000-20053-0).

model has non-realistic assumption of infinite lines and customer's infinite patient times, the performance measures are calculated simply via this model. This basic queueing model can be extended to the M/M/N+M queue with customer's patient time, which is called Erlang-A model[1]. Moreover, the effect of retrials on the call center performance has already been studied[4].

The customer's requests can be routed to two different agent types and the agents can serve customers of different types, which is commonly referred to as skill based routing[3]. As examples of the call center model having skill based routing, we have the so-called N-design, X-design, W-design and M-design models[5].

In the N-design model, one of two agent types serves both types of customers and other agents are specialists for a particular customer type. An approximate analysis of the N-design model with infinite waiting queue and priority service discipline has been done[6]. In order to analyze the performance of the call center with skill based routing, if we use the stochastic approach and numerical procedure for finding performance measures, we need heavily computational burdens.

In this paper, we use an approximate analysis method of the so-called decomposition algorithm. Our N-design model with finite waiting queues is different to the previously studied model[6]. As we know well, the approximation provides sufficient accuracy reducing the necessary completion time. The paper is organized as follows. In Section 2, we describe the system model of our N-design call center. In Section 3, we explain state spaces of the corresponding queueing model and present the approximate analysis procedure by using the decomposition algorithm. In Section 4, we give some numerical results on the system performance. Finally, in Section 5, we present some conclusions and some further studies.

2. System model and subsystems

In this section, we describe the system configuration and routing procedure of our N-design model. As we show Figure 1, we have an N-design model with two types of A and B customers and two different types of agents, the specialists and generalists. Both A and B customers arrive at the respective waiting queues A and B according to Poisson processes with respective rates λ_1 and λ_2.

Both types of agents are assume to have different skills. The first type of N_1 agents serves only type A customers(the specialist). The other type of N_2 agents serves both types of A and B customers(the generalist). Service time is exponentially distributed with mean $\mu_1^{-1}(\mu_2^{-1})$ for specialists(generalists), regardless of the customer type. We assume that the number K_1 of A customers waiting or being served in the system is finite. The number K_2 of B customers in the system is finite as well. These limitations of two waiting rooms reflect the given numbers of telephone lines for two types of customers. Therefore, the number of customers in the system does not exceed the respective limitation.

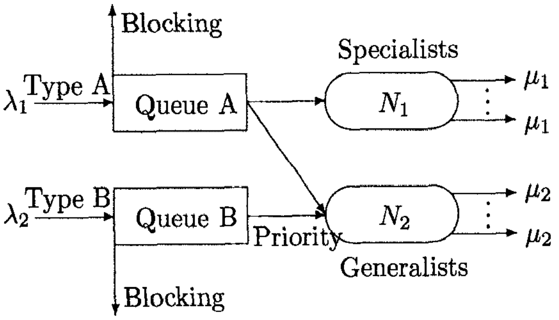

FIGURE 1. N-design model with two groups of agents

If possible, an arriving A customer will be served immediately by the specialist. Otherwise, if all specialists are busy, when a generalist is available, the generalist serves the arriving A customer. If all specialists and generalists are busy, the arriving customers join their corresponding waiting queues. The customer selection rule of generalists depends on the type of the customer. The specialists serve A customers according to FCFS(First Come First Service) rule within its own customer type. The generalist looks at B queue first and serves a waiting B customer, if possible. Otherwise, the generalist looks at A queue and serves an A customer. If there is no customer in the two queues, the generalist becomes idle.

The system can be represented by a two-dimensional Markov process. Since K_1 and K_2 are finite, the state space of the process is finite. So we have the stationary probabilities of the Markov process. Let X_1 be the number of A customers in the subsystem A (customers in A queue and in service with specialists) in steady state. Let X_2 be the number of B customers in the subsystem B (customers in B queue and customers of either type in service with generalists). After all, we have to obtain the probabilities $P(X_1 = i)$ and $P(X_2 = j)$ in order to calculate the performance measures of the N-design call center.

3. The steady state probabilities of the subsystems

3.1. The probability $q_j \equiv P(X_2 = j)$

When all specialists are busy $(X_1 \geq N_1)$ and there is an available generalist $(X_2 < N_2)$, A customers are routed to a generalist. That is, A customers overflow to the generalist from the $M/M/N_1/N_1$ queue. We can easily model this overflow traffic as an IPP(Interrupted Poisson Process)[7]. Let γ_A^{-1} and γ_S^{-1} be the mean durations of the Active and Silent periods, respectively and let λ be the customer's arrival rate during Active period[8]. Note that the traffic intensity offered to the $M/M/N_1/N_1$ queue is $\rho_1 = \lambda_1/\mu_1$. Then the overflow traffic is modeled as the IPP with the following parameters[7]

$$\lambda = \rho_1 \frac{\delta_2(\delta_1 - \delta_0) - \delta_0(\delta_2 - \delta_1)}{(\delta_1 - \delta_0) - (\delta_2 - \delta_1)},$$
$$\gamma_S = \frac{\delta_0}{\lambda} \frac{\lambda - \rho_1\delta_1}{\delta_1 - \delta_0}, \gamma_A = \frac{\gamma_S}{\rho_1} \frac{\lambda - \rho_1\delta_1}{\delta_0}, \tag{1}$$

where $\delta_n(N_1) = \sigma_n(N_1)/\sigma_{n+1}(N_1)$, $n = 0, 1, 2, \cdots$ and $\sigma_n(N_1)$, $n = 0, 1, 2, \cdots$ is given by

$$\sigma_0(N_1) = \frac{\rho_1^{N_1}}{N_1!}, \sigma_n(N_1) = \sum_{k=0}^{N_1} \binom{n+k-1}{k} \frac{\rho_1^{N_1-k}}{(N_1-k)!}.$$

Now consider the $MMPP/M/N_2/K_2^*$ model in order to find the probability q_j, where K_2^* is a random variable, which varies from the minimum K_2 to the maximum $K_2 + N_2$. In numerical examples, we will take the mean $E[K_2^*]$. Thus, for simplicity, we think of the random variable as the number. Given that the overflow process is modeled as an IPP process, we have two independent input processes of customer's arrivals that want to receive their service by the generalist. One of these process is the overflow IPP process of A customers and the other is a Poisson process of B customers with service priority against A customers. It is well known that the superposition of IPP and Poisson processes makes an MMPP(Markov Modulated Poisson Process)[9]. Let Q be the infinitesimal generator of the MMPP process and let Λ be the arrival rate matrix of the MMPP. Then the superposed process MMPP is completely represented by Q and Λ as follows

$$Q = \begin{pmatrix} -\gamma_A & \gamma_A \\ \gamma_S & -\gamma_S \end{pmatrix}, \Lambda = \begin{pmatrix} \lambda + \lambda_2 & 0 \\ 0 & \lambda_2 \end{pmatrix}, \tag{2}$$

where γ_A, γ_S and λ are given in (1) and λ_2 is the Poisson arrival rate of B customers.

To find the steady state probability q_j, let $\{(X_2, Z)\} = \{(j, k) | j \leq K_2^*, k = 1, 2\}$ be the Markov chain which represents the state of the $MMPP/M/N_2/K_2^*$ queue, where Z indicates the state of the underlying Markov process of the MMPP and X_2 indicates the number of both A and B customers in the subsystem B. Let Q^* be the infinitesimal generator of the chain $\{(X_2, Z)\}$, then we have

$$Q^* = \begin{pmatrix} Q_1 & Q_2 \\ O & Q_3 \end{pmatrix}, \tag{3}$$

where

$$Q_1 = \begin{pmatrix} Q_1(1) & \Lambda & 0 & 0 & 0 & 0 \\ \mu_2 I & Q_1(2) & \Lambda & 0 & 0 & 0 \\ 0 & 2\mu_2 I & Q_1(3) & \Lambda & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & (N_2-1)\mu_2 I & Q_1(N_2) & \Lambda \\ 0 & 0 & 0 & 0 & N_2\mu_2 I & Q_3(1) \\ 0 & 0 & 0 & 0 & 0 & N_2\mu_2 I \end{pmatrix},$$

$$Q_2 = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \Lambda_2 & 0 & 0 & \cdots & 0 & 0 & 0 \\ Q_3(1) & \Lambda_2 & 0 & \cdots & 0 & 0 & 0 \end{pmatrix},$$

$$Q_3 = \begin{pmatrix} Q_3(1) & \Lambda_2 & 0 & \cdots & 0 & 0 & 0 \\ N_2\mu_2 I & Q_3(1) & \Lambda_2 & \cdots & 0 & 0 & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & N_2\mu_2 I & Q_3(1) & \Lambda_2 \\ 0 & 0 & 0 & \cdots & 0 & N_2\mu_2 I & Q - N_2\mu_2 I \end{pmatrix},$$

where $Q_1(i) = Q - \Lambda - (i-1)\mu_2 I$, $Q_3(1) = Q - \Lambda_2 - N_2\mu_2 I$, and Q and Λ are given in (2) and $\Lambda_2 = \text{diag}(\lambda_2, \lambda_2)$ and O is $(K_2^* - N_2 - 1) \times (N_2 + 1)$-dimensional zero matrix.

Let $\pi = (\pi_0, \pi_1, \pi_2, \cdots, \pi_{K_2^*})$ with $\pi_j = (\pi_{j1}, \pi_{j2})$ be the stationary distribution of Q^* satisfying $\pi Q^* = 0$ with $\pi e = 1$[3]. Then the stationary probability of the number of customers in the subsystem B is given by

$$q_j \equiv P(X_2 = j) = \pi_{j1} + \pi_{j2}, \; j = 0, 1, 2, \cdots, K_2^*. \tag{4}$$

By the Little's law, the mean number N_A of A customers overflowed from A queue and served by the generalists is as follows

$$N_A = \sum_{j=0}^{N_2-1} \pi_{j1}\lambda/\mu_2. \tag{5}$$

3.2. The probability $p_i \equiv P(X_1 = i)$

We introduce a decomposition algorithm for approximate analysis[6]. We first divide the state space of the random variable X_2 into two disjoint regions

$S_1 = \{X_2 < N_2\}$ and $S_2 = \{N_2 \le X_2 \le K_2^*\}$. The core of the approximation algorithm is to find the following probabilities.

$$P(X_1 = i\,|X_2 = j) \approx P(X_1 = i\,|\,X_2 < N_2) \equiv p_{1,i},$$
$$P(X_1 = i\,|X_2 = j) \approx P(X_1 = i\,|\,N_2 \le X_2 \le K_2^*) \equiv p_{2,i},$$

Then the stationary probability of the number of customers in the subsystem A is given by, for $i = 0, 1, 2, \cdots, K_1^*$,

$$p_i \equiv P(X_1 = i) = p_{1,i} \sum_{j=0}^{N_2-1} q_j + p_{2,i} \sum_{j=N_2}^{K_2^*} q_j, \tag{6}$$

where q_j, $j = 0, 1, 2, \cdots, K_2^*$ are given in (4) and K_1^* is a random variable, which varies from the minimum $K_1 - N_2$ to the maximum K_1.

In the region S_1, A queue is empty and all arriving A customers are served by specialist until $\{X_1 \le N_1\}$. Thus we can model this case as the $M/M/N_1/N_1$ (Erlang-B) queue, a multi-server queue with no waiting room. The number of busy servers forms a Markov Birth-and-Death process[10]. The steady state probabilities are given by

$$p_{1,i} = (\rho_1^i/i!) \Big/ \sum_{k=0}^{N_1} \rho_1^k/k!, \; 0 \le i \le N_1, \tag{7}$$

where $\rho_1 = \lambda_1/\mu_1$ is the traffic intensity of the $M/M/N_1/N_1$ queue.

In the region S_2, all generalists are busy, the generalists are either serving B customers or are serving A customers only when B queue is empty. A customers are served by the N_1 specialists with each service rate μ_1. When all specialists are busy, A customers only see a generalist if there is no B customer in B queue($X_2 = N_2$). In this case, a service completion by a generalist diverts a customer from A queue to the generalist.

On the other hand, An arriving B customer is first served by the N_2 generalists with each service rate μ_2. When all generalists are busy B customers are waiting in B queue. In addition, when A customers are waiting in A queue, a single server is sometimes available with service rate $N_2\mu_2$. This server experiences random periods of unavailability and these breakdowns correspond to the busy periods of the $M/M/1/K_2^*$ queue. Here K_2^* is a random variable, which depends on the number of generalists occupied by A customers. The busy periods of the $M/M/1/K_2^*$ queue are approximated by a hyperexponential distribution with parameters that match the first three moments of the busy periods[6].

To find the busy period of the $M/M/1/K_2^*$ queue, we have the arrival rate λ_2, the service rate $N_2\mu_2$. Let $L(t)$ be the number of customers at time t and let τ be the length of the busy period. Then we have to find the following Laplace transform

$$\phi_n(s) = E\Big[e^{-st}\,|\,L(0) = n\Big], \; n = 1, 2, \cdots, K, \; |x| \le 1, \; s > 0, \tag{8}$$

where the boundary conditions are $\phi_{K_2^*+1}(s) = \phi_{K_2^*}(s)$ and $\phi_0(s) = 1$. Thus, $\phi_1(s)$ is the required Laplace transform. We can refer to Appendix A for finding $\phi_1(s)$.

Now to match the busy period with a hyperexponential distribution, define $h(\tau)$ as follows

$$h_2(\tau) = \alpha_2\gamma_1 e^{-\gamma_1\tau} + (1-\alpha_2)\gamma_2 e^{-\gamma_2\tau}, \tag{9}$$

where τ, α_2, γ_1 and γ_2 are non-negative. The following parameters match the first three moments of the hyperexponential distribution with the three moments m_1, m_2 and m_3 of $\phi_1(s)$[6].

$$\gamma_1 = \frac{v_1 + \sqrt{v_1^2 - 4v_2}}{2},\ \gamma_2 = \frac{v_1 - \sqrt{v_1^2 - 4v_2}}{2},\ \alpha_2 = \frac{\gamma_1(1-\gamma_2 m_1)}{\gamma_1 - \gamma_2}, \tag{10}$$

where v_1 and v_2 are given by

$$v_2 = \frac{6m_1^2 - 3m_2}{(3/2)m_2^2 - m_1 m_3},\ v_1 = \frac{1}{m_1} + \frac{m_2 v_2}{2m_1}.$$

Let's return to finding $p_{2,i}$. 1) First, when there are A customers in A queue($X_1 > N_1$), only if $X_2 = N_2$, the corresponding queueing system is governed by both $M/M/1/(K_2^* - N_1 + 1)$ queue and $M/G_1/1/(K_2^* - N_1 + 1)$ queue with the arrival rate λ_1 and the hyperexponential service time given in (9). 2) Secondly, given that $\{X_1 \le N_1\}$, A customers are served by the N_1 specialists. The corresponding queueing system is modeled as the $M/M/N_1/N_1$ queue with rates λ_1 and μ_1.

At first, we consider the case 1). Given that $X_1 = i > N_1$, the specialist serves A customers with exponential service time with mean $(N_1\mu_1)^{-1}$. Hence we can conceive that the resulting service time distribution is $B_2 \sim \text{Exp}(N_1\mu_1)$. Furthermore, the generalist serves A customers with the hyperexponential service time distribution($H_2 \sim h_2(x)$) given in (9). Consequently, A customers complete their service with the minimum time of B_2 and H_2. So the corresponding queueing system is modeled as the $M/G_2/1/(K_2^* - N_1 + 1)$ queue.

Note that by PASTA, the number of customers in the system at an arbitrary time is equal to the number of customers at an arrival epoch[11]. Then we have, for $0 \le i \le K_2^* - N_1 - 1$,

$$p_{2,i+N_1} = \frac{\pi_{i+1}}{\pi_0 + \lambda_1/(\gamma + N_1\mu_1)},$$
$$p_{2,K_2^*} = 1 - \frac{1}{\pi_0 + \lambda_1/(\gamma + N_1\mu_1)},$$

where $\gamma = \alpha_2\gamma_1 + (1-\alpha_2)\gamma_2$ and γ_1, γ_2 and α_2 are given in (10).

For the case 2), the corresponding queueing system is the $M/M/N_1/N_1$ queue. By the normalization condition including the above equations, we have

$$p_{2,i} = \frac{\rho_1^i}{i!}\left(1 - \sum_{i=N_1+1}^{K_2^*} p_{2,i}\right) \Big/ \sum_{j=0}^{N_1} \frac{\rho_1^j}{j!}. \tag{11}$$

Finally, by substituting (7) and (11) into (6), we can easily get the stationary probability of the number of customers in the subsystem A, $p_i = P(X_1 = i)$, $i = 0, 1, \cdots, K_1^*$.

3.3. Performance measures

By the equation (5), we have the mean queue capacities of queues A and B as follows

$$K_A = K_1 - N_1 - N_A, \; K_B = K_2 - N_2 + N_A.$$

The blocking probabilities for the mean queue capacities K_A and K_B are given by

$$P_A = P(X_1 = K_A + N_1), \; P_B = P(X_2 = K_B + N_2). \tag{12}$$

Given that the mean queue capacities are K_A and K_B, the mean waiting times are given by

$$W_{qA} = \frac{1}{\lambda_1(1 - P_A)} \sum_{i=N_1+1}^{K_A+N_1} (i - N_1)P(X_1 = i), \tag{13}$$

$$W_{qB} = \frac{1}{\lambda_2(1 - P_B)} \sum_{j=N_2+1}^{K_B+N_2} (j - N_2)P(X_2 = j). \tag{14}$$

4. Numerical results

In this section, we present some numerical results to show the effect of the system parameters in our N-design call center on the performance measures such as the mean waiting time and the blocking probability. We let $K_1 = 40$ and $K_2 = 30$ be two fixed numbers of telephone lines for two types of customers respectively. We choose $N_1 = 20$ agents and $N_2 = 20$ agents as the fixed numbers of the specialists and generalists, respectively. We vary customers' arrival rates per minute λ_1 and λ_2 in order to derive the proper utilizations(traffic intensities).

We assume that the generalist needs more time to serve a particular customer than the specialist. In general, customer's mean service time varies between 60 and 180 seconds[1,3]. Hence agents's service rates μ_1 and μ_2 vary between 1 and 1/3. We take the fixed values $\mu_1^{-1} = 2$ and $\mu_2^{-1} = 3$ minutes as some system parameters from Figure 2 to Figure 5.

Figure 2 shows the mean number of A customer's calls holding generalists when A customer's arrival rate varies from 2 to 20 per minute. The legend

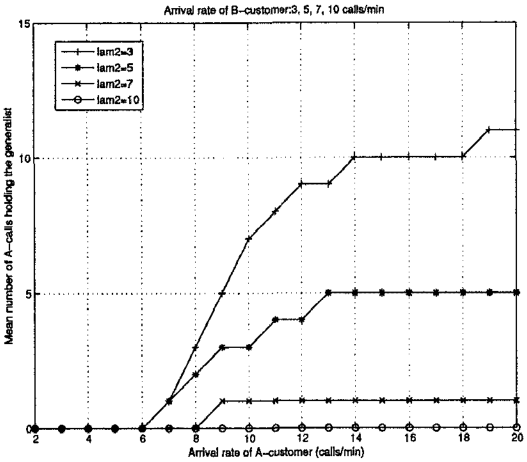

FIGURE 2. Mean number of A calls holding Generalists vs. arrival rate of A customer

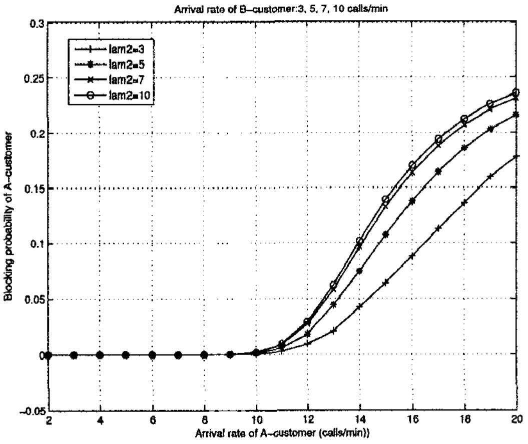

FIGURE 3. A customer's P_A vs. A customer's arrival rate

'lam2' means B customer's arrival rate to B queue and each arrival rate gives its own corresponding curve. For example, the curve of 'lam2=5' shows that A customer's arrival rate 15 calls/mimute(horizontal axis) corresponds to the mean number 5 of A customer's calls holding the generalists. Since the number of the generalists is $N_2 = 20$, we can conceive that the second agent group of the generalists serves 5 A customers and 15 B customers in average. Therefore, since the number of telephone lines for A customers is $K_1 = 40$ and the number of the specialists is $N_1 = 20$, the mean capacity of A queue becomes $K_A = 15(= 40 - 20 - 5)$. Similarly, since the number of telephone lines for B customers is $K_2 = 30$, the mean capacity of B queue becomes $K_B = 15(= 30 - 20 + 5)$. In

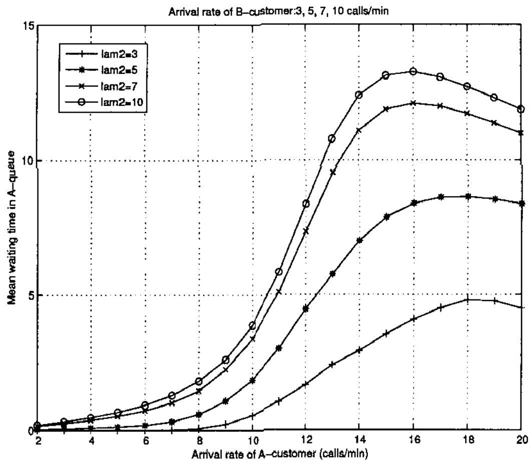

FIGURE 4. W_{qA} in A queue vs. A customer's arrival rate

Figure 2, we see that when A customer's arrival rate increases, then the mean number of A customer's calls holding the generalists.

Figure 3 shows the blocking probability(P_A) of A customer's calls when A customer's arrival rate varies from 2 to 20 per minutes. The blocking probability is well calculated in (12). We can see that when A customer's arrival rate increases, the blocking probability of A customer's calls increases exponentially in case that B customer's arriving rate is fixed. When B customer's arrival rate is high(lam2=7, 10), the blocking probabilities of A customer's calls are almost same. The reason is as follows. While B customer's arrival rates are high, the numbers of A customers holding the generalists are small and so the corresponding mean lengthes of A queue are almost same.

Figure 4 shows the mean waiting time(W_{qA}) in A queue when A customer's arrival rate varies. The waiting time is well derived in (13). We can see that when A customer's arrival rate increases, the mean waiting time of B customer continues to increase until it hits a particular value(eg. the corresponding point to A customer's arrival rate 16 in case of 'lam2=10') and then it decreases. The reason is as follows. For the fixed arrival rate of B customers, when A customer's arrival rate increases, the number of A customers holding the generalists also increases, so the mean capacity of A queue decreases.

Figure 5 shows the mean waiting time(W_{qB}) of B customer in B queue when A customer's arrival rate varies. The waiting time is well derived in (14). We can see that when A customer's arrival rate is high, the mean waiting time of B customer is almost static.

From now on, we investigate the performance measures when B customer's arrival rate λ_2 varies but A customer's arrival rate λ_1 is fixed. The 5 cases of A customer's arrival rates are considered. We take the fixed values $\mu_1^{-1} = 2$ and $\mu_2^{-1} = 3$ minutes as some system parameters from Figure 6 to Figure 9.

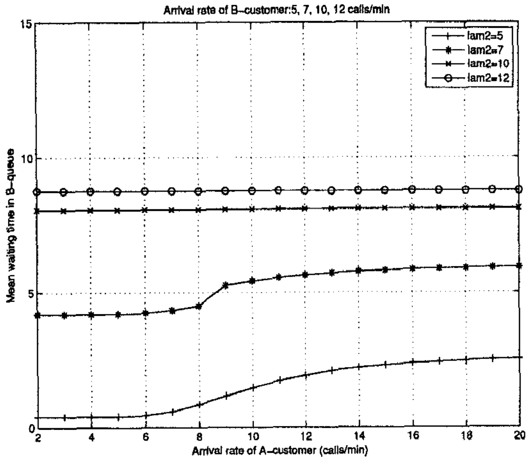

FIGURE 5. W_{qB} in B queue vs. A customer's arrival rate

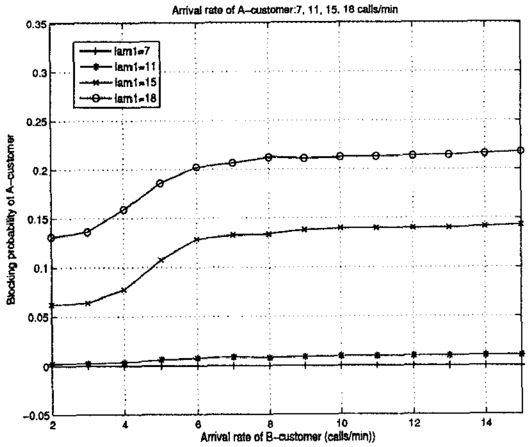

FIGURE 6. A customer's P_A vs. B customer's arrival rate

Figure 6 shows the blocking probability of A customer's calls when B customer's arrival rate varies from 2 to 15 per minute. When A customer's arrival rate is low(laml= 7, 11), it does not almost have an effect on the blocking probability of A customer's calls. When A customer's arrival rate is high(laml= 15, 18), the blocking probability of A customer continues to increase until it hits a particular value(eg. the corresponding point to B customer's arrival rate 7) and then it shows almost no change. The reason is as follows. While B customer's arrival rate is high, the number of A customers holding the generalists decreases and so the mean capacity of A queue increases. Thus the blocking probability of A customer's calls shows almost no change.

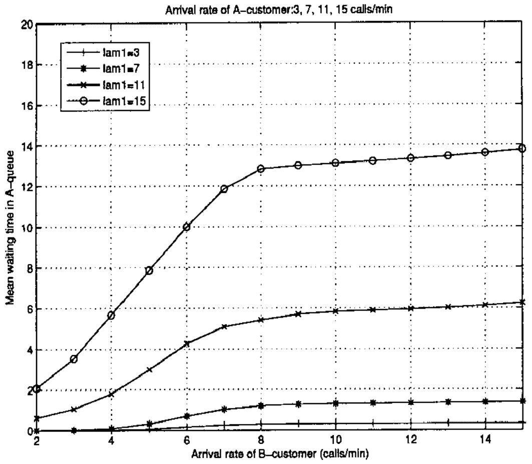

FIGURE 7. W_{qA} in A queue vs. B customer's arrival rate

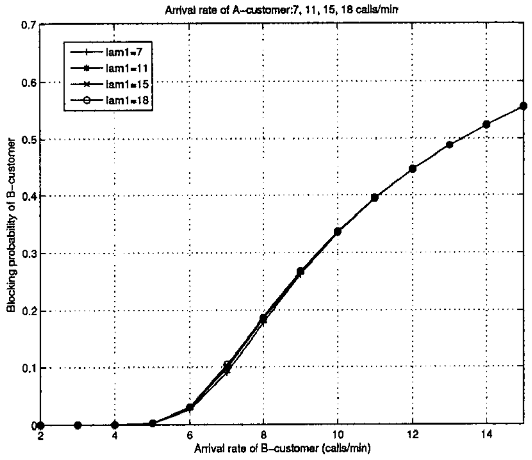

FIGURE 8. B customer's P_B vs. B customer's arrival rate

Figure 7 shows A customer's mean waiting time in A queue when B customer's arrival rate varies. We can see that when B customer's arrival rate increases, A customer's mean waiting time increases. Especially, A customer's mean waiting time continues to increase rapidly until it hits a particular value(eg. the corresponding point to B customer's arrival rate 9) and then it shows slow increase. The reason is the same as the case of Figure 6.

Figure 8 shows the blocking probability of B customer's calls when B customer's arrival rate varies. We can see that when B customer's arrival rate increases, the blocking probability of B customer's calls increases exponentially in case that A customer's arriving rate is fixed. The figure shows the similar

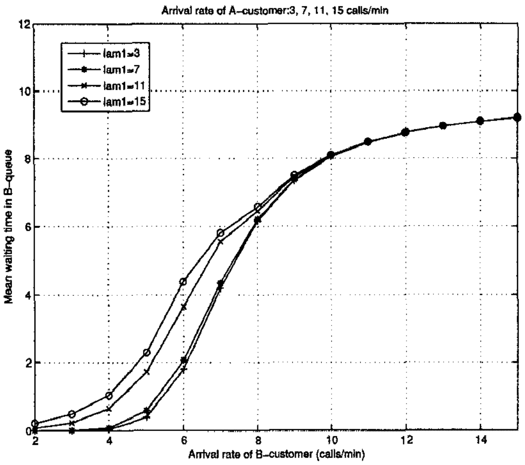

FIGURE 9. W_{qB} in B queue vs. B customer's arrival rate

behavior of the blocking probability as in the ordinary queueing systems. Especially, we show that A customer's arrival rate does not effect on the blocking probability of B customer's calls.

Figure 9 shows B customer's mean waiting time in B queue when B customer's arrival rate varies. We can see that when B customer's arrival rate increases, the mean waiting time of B customer continues to increase exponentially. The figure shows the similar behavior of the mean waiting time as in the ordinary queueing systems

5. Conclusion

In this paper, we dealt with an approximate queueing analysis of the so-called N-design call center by using the decomposition method, which divided the state space of the queueing system into small regions. Our N-design model with finite waiting queues and exponentially patent times is different to the previously studied model with infinite waiting queues. We presented some numerical examples and showed the impact of the system parameters on the performance measures of the call center. We described the behavior of the mean number of customers holding the specialists, the blocking probabilities and mean waiting times when customer's arrival rates varied. It is for further studies that we investigate the optimal workforce staffing of different agent groups under consideration for agent's different level of skills and the behavior of customer's waiting times in the queues corresponding to the customer's patient times.

Appendix A. Derivation of Laplace transform $\phi_1(s)$

Conditioning on the epoch of customer's first arrival, departure or reneging(whichever occurs first), the recursive relation of the definition (8) with given boundary condition gives

$$\phi_1(s) = \frac{N_2\mu_2}{\lambda_2 + N_2\mu_2 + s} + \frac{\lambda_2\phi_2(s)}{\lambda_2 + N_2\mu_2 + s},$$
$$\phi_n(s) = \frac{N_2\mu_2\phi_{n-1}(s)}{\lambda_2 + N_2\mu_2 + s} + \frac{\lambda_2\phi_{n+1}(s)}{\lambda_2 + N_2\mu_2 + s}, \; 2 \le n \le K.$$

To find $\{\phi_n(s)\}$ from the above equations, define $\Phi(z, s)$ as the z-transform of $\{\phi_n(s)\}$ by

$$\Phi(z, s) = \sum_{n=1}^{K} \phi_n(s) z^{n-1}.$$

By multiplying the both sides of the above iteration equations by z^n, $n = 1, 2, \cdots, K$ and summing up all terms, we have

$$\Phi(z, s) = \frac{\left[N_2\mu_2 z^{K+1} - \lambda_2 z^K\right]\phi_K(s) + \lambda_2\phi_1(s) - N_2\mu_2 z}{N_2\mu_2 z^2 - (\lambda_2 + N_2\mu_2 + s)z + \lambda_2}.$$

Let z_1 and z_2 be the roots of the above denominator, we have

$$z_1, z_2 = \frac{(\lambda_2 + N_2\mu_2 + s) \pm \sqrt{(\lambda_2 + N_2\mu_2 + s)^2 - 4\lambda_2 N_2\mu_2}}{2N_2\mu_2}.$$

By the property of the analytic function, the roots z_1 and z_2 also satisfy the numerator of $\Phi(z, s)$. Thus we have the following linear equations

$$\left[N_2\mu_2 z_1^{K+1} - \lambda_2 z_1^K\right]\phi_K(s) + \lambda_2\phi_1(s) = N_2\mu_2 z_1,$$
$$\left[N_2\mu_2 z_2^{K+1} - \lambda_2 z_2^K\right]\phi_K(s) + \lambda_2\phi_1(s) = N_2\mu_2 z_2.$$

By solving the above linear equations, we have

$$\phi_1(s) = \frac{\left(z_1^K - z_2^K\right) - \frac{\lambda_2}{N_2\mu_2}\left(z_1^{K-1} - z_2^{K-1}\right)}{\left(z_1^{K+1} - z_2^{K+1}\right) - \frac{\lambda_2}{N_2\mu_2}\left(z_1^K - z_2^K\right)},$$

which is the required Laplace transform.

References

1. A. Mandelbaum and S. Zeltyn, *Service Engineering in Action: The Palm/ Erlang-A Queue, with Applications to Call Centers*, Israeli Science Foundation Research Report, (2005)
2. S. Borst, A. Mandelbaum and M. I. Reiman, *Dimensioning Large Call Centers*, Operations Research 52(2004), No. 1, 17-34.
3. R. Stolletz and S. Helber, *Performance analysis of an inbound call center with skills-based routing*, Queueing Systems 47(2004), 117-146.
4. S. Aguir, F. Karaesman, O.Z. Aksin and F. Chauvet, *The impact of retrials on call center performance*, OR Spectrum 26(2004), 353-376.

5. N. Gans, G. Koole and A. Mandelbaum, *Commissioned Paper, Telephone Call Centers: Tutorial, Review, and Research Prospect,* Manufacturing & Science Operations Management 5(2003), No. 2, 79-141.
6. R.A. Shumsky, *Appoximation and analysis of a call center with flexible and specialized servers,* OR Spectrum 26(2004), 307-330.
7. A. Kukzura, *The interrupted poisson process as an overflow process,* Bell System Technical Journal 52(1973), No, 3, 437-448.
8. R.O. Onvural, *Asychronous Transfer Mode Networks: Performance Issues,* Second edition, Artech House, 1995 processes and its application to packetized voice multiplexers, Proceedings of ITC-12(12th International Teletraffic Congress), Torino, Italy, pp. 1399-1405, 1-8 June, (1988).
9. K. S. Meier-Hellstern, *The Analysis of a Queue Arising in Overflow Model,* IEEE Trans. on Comm. 37(1989), 367-372.
10. L. Kleinrock, *Queueing Systems, Volume I: Theory,* John Wiley & Sons, Inc. New York, (1975).
11. W. Tagaki, *Queueing Analysis, Vol. 2: Finite Systems,* IBM Japan, Ltd., North-Holland, (1993).

Chul Geun Park
Department of Information and Communications Engineering, Sunmoon University, Tangjeong-myun, Asan-si, Chungnam, 336-708, Korea
e-mail: cgpark@sunmoon.ac.kr

Dong Hwan Han
Department of Mathematics, Sunmoon University, Tangjeong-myun, Asan-si, Chungnam, 336-708, Korea
e-mail: dhhan@sunmoon.ac.kr

Kwang Hyun Baik
Department of IT Business Administration, Sunmoon University, Tangjeong-myun, Asan-si, Chungnam, 336-708, Korea
e-mail: baik@sunmoon.ac.kr