APPROXIMATE ANALYSIS OF A CALL CENTER WITH SKILL-BASED ROUTING

Chul Geun Park
Department of Information and Communications Engineering, Sunmoon University, Kalsan 100, Asan-si, Korea

Dong Hwan Han
Department of Mathematics, Sunmoon University, Kalsan 100, Asan-si, Korea

Keywords: Call center, N-design, Skill-based, Queueing analysis, Performance.

Abstract: Call centers have become the prevalent contact points between many companies and their customers. By virtue of recent advances in information and communication technology, the number and size of call centers has grown dramatically. As a large portion of the operating costs are related to the labor costs, efficient design and workforce staffing are crucial for the economic success of call centers. In this context, the workforce staffing level can be modeled as mathematical optimization problem using queueing theory. In this paper, we deal with an approximate analysis of the so-called N-design call center with two types of customers, two different finite queues and two different exponential patient times. We also represent some numerical examples and show the impact of the system parameters on the performance measures.

1 INTRODUCTION

Contact centers are service organizations for customers who need service via the phone, facsimile, e-mail or other telecommunication channels. A particularly important type of contact center is the call center. By virtue of recent advances in information and communication technology, the number and size of call centers as well as the number of customers and agents grow explosively(Mand,2005). 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. Indeed, some call center statistics assess that 70% of all customer-business information in the U.S. occur in call centers which employ about 3% of the U.S. workforce and 1.5 million agents(Bors,2004; Stol,2004).

In the most simple design of call centers, only 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 referred as Erlang-C. Though Erlang-C model has non-realistic assumption of infinite lines and customer's infinite patient times, the performance measures are easily calculated. Customer's patient times have a considerable effect on the performance of the system(Shim,2004; Mand,2004). This basic queueing model can be extended to the M/M/N+M queue(Erlang-A model) and the M/M/N+G queue with patient times(Mand,2005; Mand,2004).

The skill set of agents describes for which kind of service the agent is skilled and how well he provides service. The customer's requests can be routed to different agent groups and the agents can serve customers of different types, which is commonly referred to as skill based routing(Stol,2004). As examples of skill-based routing, we have the so-called N-design, X-design, W-design and M-design models(Gans,2003; Stol,2003). In the N-design model, one of two agent groups serves both types of customers and other agents are specialists for a particular customer type. Approximate analysis of the N-design model with infinite waiting queue and priority service discipline has been done(Shum,2004).

In this paper, we use an approximate analysis method of the so-called decomposition algorithm to reduce computational burdens. The considered N-design model with finite waiting queues and exponential patient times is different to the previously studied model(Shum,2004). As we know well, the approximation provides sufficient accuracy reducing the necessary completion time.

Geun Park C. and Hwan Han D. (2007).
APPROXIMATE ANALYSIS OF A CALL CENTER WITH SKILL-BASED ROUTING.
In *Proceedings of the Second International Conference on e-Business*, pages 45-50
DOI: 10.5220/0002111000450050

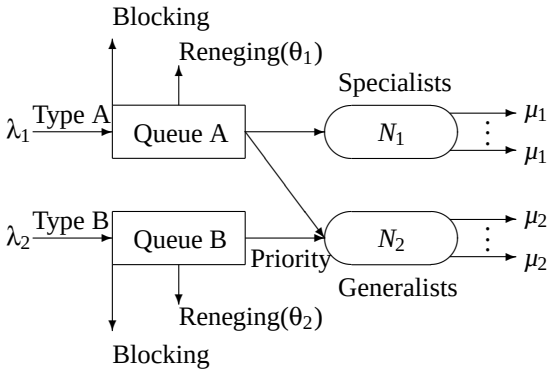

Figure 1: N-design model with two groups of agents.

2 SYSTEM MODEL

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 customers A and B and two different groups 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 customers are patient. The type A(B) customer reneges his waiting in his own queue after an exponentially distributed patient time with mean $\theta_1^{-1}(\theta_2^{-1})$, if the service has not begun. We assume that the reneging customers are lost and so there are no retrials.

Both groups of agents are assume to have different skills. The first group of N_1 agents serves only type A customers(Specialist). The other group of N_2 agents serves both types of A and B customers(Generalist). Service times are exponentially distributed with means μ_1^{-1} and μ_2^{-1} for specialists and generalists, respectively 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 cases of given numbers of telephone lines for two types of customers respectively. In this way, when K_1 A customers are in the system, an arriving A customer receives a busy signal and lost. In the same way, the number of B customers in the system does not exceed the limitation K_2.

If possible, an arriving A customer will be served immediately by the specialist. Otherwise, if all specialists are busy, when a generalist is available, this 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. Thus the generalist has N-design routing policy with priority service discipline and gives non-preemptive priority to B customers.

Now we describe an overview of the approximation procedure. The system can be represented by a two-dimensional Markov process. Since K_1 and K_2 are finite in our N-design model, the state space of the process is finite. So the resulting two-dimensional Markov process has a stationary probability distribution. Let X_1 be the number of A customers in A queue and in service with specialists in steady state. Let X_2 be the sum of the number of B customers in B queue and the number of customers of either type in service with generalists.

We introduce a decomposition algorithm for approximate performance analysis(Shum,2004). We first divide the state space into four regions $S_1 = \{X_1 \leq N_1\} \bigcap \{X_2 < N_2\}$, $S_2 = \{N_1 < X_1 \leq K_1^*\} \bigcap \{X_2 < N_2\}$, $S_3 = \{X_1 \leq N_1\} \bigcap \{N_2 \leq X_2 \leq K_2^*\}$, and $S_4 = \{N_1 < X_1 \leq K_1^*\} \bigcap \{N_2 \leq X_2 \leq K_2^*\}$. Here K_1^* and K_2^* are random variables, which will be described in the next section. In the numerical approximation, we take the respective means $K_A = E[K_1^*]$ and $K_B = E[K_2^*]$. Thus, for simplicity, we think of these variables as the numbers. Clearly, the region S_2 is forbidden. The core of the approximation algorithm is to find the following probabilities.

$$P(X_1 = i|X_2 = j) \approx P(X_1 = i\,|\,X_2 < N_2), \quad (1)$$
$$P(X_1 = i|X_2 = j) \approx P(X_1 = i\,|\,N_2 \leq X_2 \leq K_2^*), \quad (2)$$
$$P(X_2 = j\,|X_1 = i) \approx P(X_2 = j\,|\,X_1 \leq N_1), \quad (3)$$
$$P(X_2 = j\,|X_1 = i) \approx P(X_2 = j\,|\,N_1 < X_1 \leq K_1^*). \quad (4)$$

3 THE SUBMODEL ANALYSIS

Given that the system occupies in one of the regions S_1, S_3 and S_4, we can easily find the system performance along one-dimension by using conventional queueing models.

3.1 To Get $p_{1,i} \equiv P(X_1 = i|X_2 < N_2)$

Given that $\{X_2 < N_2\}$, there is no the case $\{N_1 < X_1 \leq K_1^*\}$. Under the condition $\{X_2 < N_2\}$, A queue

is empty and all arriving A customers are served by specialist until $\{X_1 \leq N_1\}$. Thus we can model this case as the $M/M/N_1/N_1$(Erlang-B) queue. The number of busy servers forms a Markov Birth-and-Death process(Gross,1985).

3.2 To Get

$$q_{2,j} \equiv P(X_2 = j | N_1 < X_1 \leq K_1^*)$$

Given that $\{N_1 < X_1 \leq K_1^*\}$, all generalists are busy. There is no the case $\{X_2 < N_2\}$. In case of $N_2 < j < K_2^*$, the state transition from $X_2 = j$ to $j-1$ occurs with rate $N_2\mu_2$. But the state transition $X_2 = j$ to $j+1$ occurs with rate λ_2. The waiting B customers in B queue renege after an exponentially distributed patient time with mean θ_2^{-1}, if the service does not begin. Thus, in this region, the submodel is well described as the $M/M/1/K_2^* + M$ queue. Here K_2^* is a random variable, which varies from the minimum $K_2 - N_2 + 1$ (all generalists serve B customers) to the maximum $K_2 + 1$ (all generalists serve A customers). The mean B queue length can be easily calculated by the distribution of the number of A customers served by the generalists(Garn,2002).

3.3 To Get $q_{1,j} \equiv P(X_2 = j | X_1 \leq N_1)$

Given that $\{X_1 \leq N_1\}$, B customers are served by N_2 generalists with service rate μ_2. There is no waiting A customer in A queue. A customers are routed to a generalist when all specialists are busy($X_1 = N_1$) and there is an available generalist($X_2 < N_2$). 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)(Kukz,1973).

The IPP is a Poisson process which is alternatively turn on for an exponentially distributed period(Active) and turn off for another exponentially distributed period(Silent). During Active period, the interarrival times of customers are exponentially distributed, while no customers are arrived during Silent period(Onvu,1995). 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. Let Q_I be the infinitesimal generator of the underlying Markov chain of the IPP and let Λ_I be the arrival rate matrix of the IPP. Then the IPP is completely characterized by Q_I and Λ_I as follows

$$Q_I = \begin{pmatrix} -\gamma_A & \gamma_A \\ \gamma_S & -\gamma_S \end{pmatrix}, \Lambda_I = \begin{pmatrix} \lambda & 0 \\ 0 & 0 \end{pmatrix}. \quad (5)$$

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 easily modeled as the IPP(Kukz,1973).

Given that the overflow process is modeled as an IPP process, we have two independent input processes to 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)(Heff,1986). Then the superposed process MMPP is completely represented by the infinitesimal generator Q and the arrival rate matrix Λ as follows

$$Q = Q_I, \Lambda = \Lambda_I + \Lambda_2, \quad (6)$$

where Q_I and Λ_I are given in (5) and Λ_2 =diag(λ_2, λ_2) is a diagonal matrix.

Let's return to finding the probability $q_{1,j}$ in the condition $\{X_1 \leq N_1\}$. Clearly, both A and B customers are served in FCFS order until all generalists are busy. When all generalists are busy, B customer in B queue is served by an available generalist just completing service according to the priority rule. Then the required probability $q_{1,j}$ is the steady state probability that the sum of busy generalists and B customers in B queue is j at customer's arrival epoch to the $MMPP/M/N_2/K_2^*$ queue. Here K_2^* is a random variable, which varies from the minimum K_2 to the maximum $K_2 + N_2$. Let $\{(X_2, Z)\} = \{(j,k) | j \leq K_2^*, k = 1,2\}$ be the Markov chain, 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 submodel. Let Q^* be the infinitesimal generator of the chain $\{(X_2, Z)\}$, then we have

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

where

$$Q_1 = \begin{pmatrix} Q_1(1) & \Lambda & \cdots & 0 & 0 \\ \mu_2 I & Q_1(2) & \cdots & 0 & 0 \\ 0 & 2\mu_2 I & \cdots & 0 & 0 \\ \cdots & \cdots & \ddots & \cdots & \cdots \\ 0 & 0 & \cdots & Q_1(N_2) & \Lambda \\ 0 & 0 & \cdots & N_2\mu_2 I & Q_1(N_2+1) \\ 0 & 0 & \cdots & 0 & c_1 I \end{pmatrix},$$

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

$$\begin{pmatrix} Q_3(2) & \Lambda_2 & \cdots & 0 & 0 \\ c_3 I & Q_3(3) & \cdots & 0 & 0 \\ \cdots & \cdots & \ddots & \cdots & \cdots \\ 0 & 0 & \cdots & Q_3(E) & \Lambda_2 \\ 0 & 0 & \cdots & c_{K_2^*-N_2} I & Q - c_{K_2^*-N_2} I \end{pmatrix},$$

where $Q_1(i) = Q - \Lambda - (i-1)\mu_2 I$, $c_k = N_2\mu_2 + k\theta_2$, $k = 1, 2, \cdots, K_2^* - N_2$, $Q_3(i) = Q - \Lambda_2 - c_i I$, $Q_3(E) = Q_3(K_2^* - N_2 - 1)$ and Q, Λ and Λ_2 are given in (6) and O is $(K_2^* - N_2 - 1) \times (N_2 + 1)$-dimensional zero matrix.

To find the probability $q_{1,j}$, let π be the stationary distribution of Q^* satisfying $\pi Q^* = 0$ with $\pi e = 1$ (Stol,2004) by, for $j = 0, 1, 2, \cdots, K_2^*$,

$$\pi = (\pi_0, \pi_1, \pi_2, \cdots, \pi_{K_2^*}) \text{ with } \pi_j = (\pi_{j1}, \pi_{j2}).$$

Then the required probability $q_{1,j}$ is given by

$$q_{1,j} = \begin{cases} \pi_j \Lambda e / C, & j = 0, 1, \cdots, N_2 - 1, \\ \pi_j \Lambda_2 e / C, & j = N_2, N_2 + 1, \cdots, K_2^*, \end{cases}$$

where $C = \sum_{l=0}^{N_2-1} \pi_l \Lambda e + \sum_{l=N_2}^{K_2^*} \pi_l \Lambda_2 e$. Moreover let r_j be the ratio that the number of A customers served by generalists is j, then we have

$$r_j = \pi_j \Lambda_l e / C, \ j = 0, 1, 2, \cdots, N_2. \qquad (7)$$

3.4 To Get $p_{2,i} \equiv P(X_1 = i | N_2 \le X_2 \le K_2^*)$

Given that $\{N_2 \le X_2 \le K_2^*\}$, 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 $M/M/1/K_2^* + M$ queue. Here K_2^* is a random variable, which depends on the number of generalists occupied by A customers. The busy periods of the $M/M/1/K_2^* + M$ queue with the parameters λ_2, $N_2\mu_2$ and θ_2 are approximated by a hyperexponential distribution with parameters that match the first three moments of the busy periods(Shum,2004).

Let $L(t)$ be the number of customers at time t in the submodel and let τ be the length of the busy period. Then we have to find the following Laplace transform, on $|x| \le 1$, $s > 0$,

$$\phi_n(s) = E[e^{-st} | L(0) = n], \ n = 1, 2, \cdots, K, \qquad (8)$$

where the boundary conditions are $\phi_{K_2^*+1}(s) = \phi_{K_2^*}(s)$ and $\phi_0(s) = 1$. After all, $\phi_1(s)$ is the required Laplace transform. Conditioning on the epoch of customer's first arrival, departure or reneging(whichever occurs first), we can easily find $\phi_1(s)$.

Now to approximate the busy period distribution, define $h(\tau)$ as follows

$$h(\tau) = \alpha\gamma_1 e^{-\gamma_1 \tau} + (1 - \alpha)\gamma_2 e^{-\gamma_2 \tau}, \qquad (9)$$

where τ, α, 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 $\phi_1(s)$ (Shum,2004).

$$\gamma_1, \gamma_2 = \frac{v_1 \pm \sqrt{v_1^2 - 4v_2}}{2}, \quad \alpha = \frac{\gamma_1(1 - \gamma_2 m_1)}{\gamma_1 - \gamma_2}, \ (10)$$

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 $M/M/1/(K_2^* - N_1 + 1) + M$ queue and $M/G_1/1/(K_2^* - N_1 + 1) + M$ 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 the arrival rate λ_1 and the service rate 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 $(N_1\mu_1)^{-1}$. If we consider reneging, then we can conceive that the resulting service time distribution is $B \sim \text{Exp}(N_1\mu_1 + \theta_1)$. Furthermore, the generalist serves A customers with the hyperexponential service time ($H \sim h(x)$) given in (9). Consequently, A customers complete their service with the minimum time of B and H. So the corresponding queueing system is modeled as the $M/G_2/1/(K_2^* - N_1 + 1) + M$ 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(Taga,1993). Then we have, for $0 \le i \le K_2^* - N_1 - 1$,

$$p_{2,i+N_1} = \frac{\pi_{i+1}}{\pi_0 + \lambda_1/(\gamma + \theta_1 + N_1\mu_1)},$$

$$p_{2,K_2^*} = 1 - \frac{1}{\pi_0 + \lambda_1/(\gamma + \theta_1 + N_1\mu_1)},$$

where $\gamma = \alpha\gamma_1 + (1-\alpha)\gamma_2$ is given by (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

$$p_{2,i} = \frac{1}{i!}\left(\frac{\lambda_1}{\mu_1}\right)^i \left(1 - \sum_{i=N_1+1}^{K_2^*} p_{2,i}\right) \Big/ \sum_{j=0}^{N_1} \frac{1}{j!}\left(\frac{\lambda_1}{\mu_1}\right)^j.$$

3.5 Performance Measures

When we know the probabilities $\{p_{1,i}\}$, $\{p_{2,i}\}$, $\{q_{1,j}\}$ and $\{q_{2,j}\}$, by the conditional probability, we can easily get the probabilities $P(X_1 = i)$, $i = 0,1,\cdots,K_1^*$ and $P(X_2 = j)$, $j = 0,1,\cdots,K_2^*$. By the probability (7), the mean queue lengths of queues A and B are given by

$$K_A = K_1 - N_1 - \sum_{j=1}^{N_2} j r_j,\ K_B = K_2 - N_2 + \sum_{j=1}^{N_2} j r_j.$$

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

$$P_A = P(X_1 = K_A + N_1),\ P_B = P(X_2 = K_B + N_2).$$

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

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

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

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 $K_1 = 70$ and $K_2 = 50$ be two fixed numbers of telephone lines for two types of customers respectively. We choose $N_1 = 30$ agents and $N_2 = 40$ 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 get 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(Mand,2005; Stol,2004). Hence agents's service rates μ_1 and μ_2 vary between 1 and 1/3. We can usually select θ_1^{-1} and θ_2^{-1} between 120 and 240 seconds as the mean values of the exponentially distributed patient times(Mand,2004; Gans,2003). We take the fixed values $\mu_1^{-1} = 2$, $\mu_2^{-1} = 3$, $\theta_1^{-1} = 2$ and $\theta_2^{-1} = 4$ minutes as some system parameters in Figs. 2 and 3.

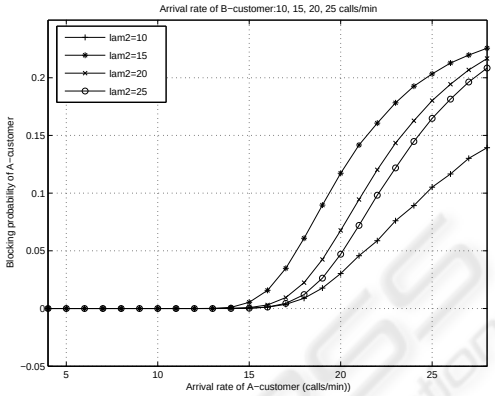

Figure 2: A customer's P_A vs. A customer's arrival rate.

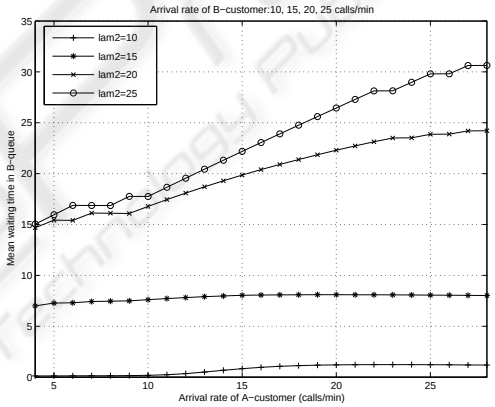

Figure 3: W_{qB} in B queue vs. A customer's arrival rate.

Fig. 2 shows the blocking probability(P_A) of A customer's calls when A customer's arrival rate varies from 4 to 28 per minutes. The blocking probability is well calculated in Section 3.5. We can see that when A customer's arrival rate increases, the blocking probability of A customer's calls increases exponentially in case that A customer's arriving rate is fixed. When B customer's arrival rate is high(lam2=20, 25), the blocking probability of A customer's calls decreases according to B customer's arrival rate. The reason is as follows. While B customer's arrival rate is high, if the number of A customers holding the generalists decreases, the mean length of A queue increases. Thus the blocking probability of A customer's calls decreases.

Fig. 3 shows B customer's mean waiting time (W_{qB}) in B queue when A customer's arrival rate varies. The waiting time is well derived in Section 3.5. We can see that when A customer's arrival rate

increases, the mean waiting time of B customer continues to increase.

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 $\mu_1^{-1} = 2$, $\mu_2^{-1} = 3$, $\theta_1^{-1} = 2$ and $\theta_2^{-1} = 4$ minutes as some system parameters in Figs. 4 and 5. Fig. 4 shows the blocking probability(P_B) 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 in case that A customer's arriving rate is fixed.

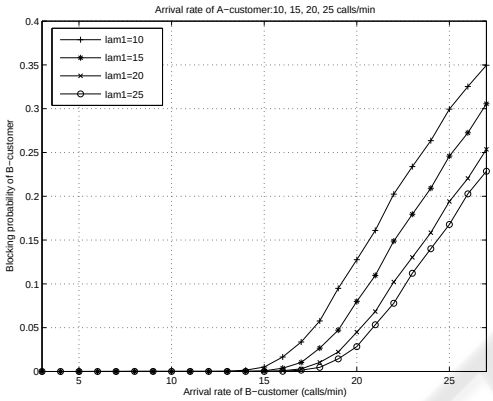

Figure 4: B customer's P_B vs. B customer's arrival rate.

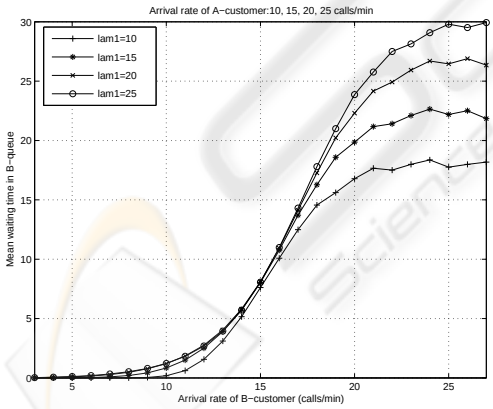

Figure 5: W_{qB} in B queue vs. B customer's arrival rate.

Fig. 5 shows B customer's mean waiting time (W_{qB}) 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. We also see that the behavior of the mean waiting time is similar to that of the ordinary queueing systems, when the arrival rate is low.

ACKNOWLEDGEMENTS

This research was supported by the MIC, Korea, under the ITRC support program supervised by the IITA.

REFERENCES

Mandelbaum, A., & Zeltyn, S. (2005). Service Engineering in Action: The Palm/ Erlang-A Queue, with Applications to Call Centers. *Israeli Science Foundation Research Report*.

Borst, S., Mandelbaum, A., & Reiman, M.I. (2004). Dimensioning Large Call Centers. *Operations Research*, 52, 17-34.

Stolletz, R., & Helber, S. (2004). Performance analysis of an inbound call center with skills-based routing. *OR Spectrum*, 26, 331-352.

Shimkin, M., & Mandelbaum, A. (2004). Rational Abandonment from Tele-Queues: Nonlinear Waiting Costs with Heterogeneous Preferences. *Queueing Systems*, 47, 117-146.

Mandelbaum, A., & Zeltyn, S. (2004). The impact of customer's patience on delay and abandon ment: some empirically-driven experiments with the M/M/n+G queue. *OR Spectrum*, 26, 377-411.

Gans, N., Koole, G., & Mandelbaum, A. (2003). Commissioned Paper, Telephone Call Centers: Tutorial, Review, and Research Prospect. *Manufacturing & Science Operations Management*, 5, 79-141.

Stolletz, R. (2003). Performance analysis and optimization of inbound call centers. *Lecture Notes in Economics and Mathematical systems*, 528.

Shumsky, R.A. (2004). Appoximation and analysis of a call center with flexible and specialized servers. *OR Spectrum*, 26, 307-330.

Gross, D., & Harris, C.H. (1985). *Fundamentals of Queueing Theory*. John Wiley & Sons, Inc.

Garnett, O., Mandelbaum, A., & Reiman, M. (2002). Designing a Call Center with Impatient Customers. *Manufacturing and Science Operations Research*, 4, 208-227.

Kukzura, A. (1973). The interrupted poisson process as an overflow process. *Bell System Technical Journal*, 52, 437-448.

Onvural, R.O. (1975). *Asychronous Transfer Mode Networks: Performance Issues.* Second edition, Artech House.

Heffes, H., & Lucantony, D.M. (1986). A Markov modulated characterization of packetized voice, data traffic and related statistical multiplexer performance. *IEEE J. Selected Areas Comm.*, 4, 856-868.

Tagaki, H. (1993). *Queueing Analysis, Vol. 2: Finite Systems.* IBM Japan, Ltd., North-Holland.