A Performance Analysis of Call Centers Based on a Multi-server Queue with Retrials, Feedbacks, and Impatience

Yi-Jun Zhu, Ren-Xiang Zhu, Zhe George Zhang, Peter Haug

Abstract— **We consider a finite buffer queueing model with several key features of call centers, such as retrials, feedbacks, and impatience. In addition, because we do not completely understand the customer impatience behavior, we use a general distribution for the maximum waiting time before abandoning the call. We develop a QBD process with infinite state space for the queue in a call center situation. To solve for the stationary performance measures, we introduce an effective approximation method, and numerical examples have been presented to show the effectiveness of our method.**

Index Terms— **Multi-server queues, call centers, retrials, feedbacks, QBD process.**

I. Introduction

Queueing models are the main quantitative technique in evaluating the operating performance of call centers. There are three common characteristics in the customer's (or caller's) behavior: (1) a customer may try to call again if he or she gets a busy signal; (2) for a customer on hold, if his or her waiting time reaches a limit, he or she will hang up and leave; and (3) a customer may call again if his or her problems are not solved completely after a service (see [1], [14]). Therefore, we present a queueing model with customers' retrials, feedbacks, and impatience. In addition, to realistically model call centers, we assume a finite buffer to hold the waiting customers. There are many works on queueing models for call centers due to the recent and rapid growth of this industry. For retrial queueing models, most existing studies are on queues with one or no waiting spot (see [10]). Another class of queueing models is the multi-server queue with both customer retrials and impatience. Most studies in this class focus on models with only one or two servers (see [2], [3], and [4]). Since the retrial models with many waiting spots and/or multiple servers usually require the infinite state space Quasi-Birth-and-Death (QBD) processes, it is extremely difficult or even impossible to obtain the stationary performance measures of the systems. However, to quantitatively evaluate the performance of practical call centers, we need to consider these difficult models.

There are mainly three types of methods to solve the QBDs for call center models. Type 1 is to formulate a QBD process with a special transition probability matrix structure where the entries become the same after a certain level (see [5]). For this kind of QBDs, a matrix geometric solution can be obtained in terms of a rate matrix which can be evaluated using a numerical method. Type 2 is to use the state space truncation to convert infinite state models to finite state ones which can be solved (see [6], [7]). Type 3 is to approximate the original infinite QBD model by another infinite one which is solvable (see [8], [9], and [10]). The model of this paper has not been studied via QBD approach in the past. We formulate the QBD process for a call center system with all its main features. To solve the QBD process, we proposed a method of Type 3 in which the original model is approximated by a simpler and solvable QBD process. Then, the stationary performance measures of our original model can be obtained via this easier-to-solve QBD process. Numerical examples have been used to show the effectiveness and efficiency of our method.

The rest of the paper is organized as follows. In section 2, we formulate a QBD process model for the queueing system with the main features of call centers. In section 3, we present the approximation method to solve the QBD process model and give some useful performance measures. In section 4, we provide some computational results to discuss the effectiveness and the efficiency of the approximation method and conclude the paper with a summary.

II. Model Formulation – A QBD Process

Consider a queueing system with a waiting and service area and a retrial area (see Figure 1). In the waiting and service area, there are s servers and $k-s<\infty$ waiting spots (or the system can hold a maximum of k waiting and in-service customers.) We assume that customers arrive to the system according to a Poisson process with rate λ, and the service time is exponentially distributed with rate μ. The service discipline is a "first-come-first-served" (FCFS) sequence. An after-service customer may enter the retrial area and call again for further service. This behavior is called the feedback. The feedback probability is β <1 and the probability of leaving the system then is $\bar{\beta}\,(\bar{\beta}=1-\beta)$.

Manuscript received..

Zhe George Zhang and Peter Haug are with the Western Washington University, Bellingham, WA 98225 USA: 360-650-2867; fax: 360-650-4844; (e-mail: george.zhang@ wwu.edu).

Yi-Jun Zhu and Ren-Xiang Zhu are with [1]Science Faculty, Jiangsu University Zhenjiang Jiangsu 212013, China (e-mail: yjzhu@ujs.edu.cn).

If an arriving customer finds the waiting room is full, he or she may join the retrial group to call again. The retrial probability is α (0< α < 1) and the probably of leaving the system without retrial is $\overline{\alpha}$ ($\overline{\alpha} = 1 - \alpha$). The retrial times are exponentially i.i.d. random variables with rate λ_1, (0< λ_1 < ∞). We also assume the retrial area is infinite. In addition, some customers are impatient and may abandon the calls. We assume that if the waiting time of a customer is beyond a threshold θ, he or she will abandon the call and will not call again. θ is a random variable with a general distribution function $F_\theta(\tau)$ and $F_\theta(0) = 0$. Finally, the interarrival times, the service times, the retrial times, and the impatient limit θ are mutually independent. Therefore, the model is a multiple server queue with retrials, feedbacks, customer impatience, and a finite buffer, denoted by $M/M/s/k+G$. Note that this model has captured most of the operating features of practical call centers.

Let $N(t)$ be the number of customers in the service and waiting area at time t and $M(t)$ be the number of customers in the retrial area at time t . Thus, we can define $X(t) = (N(t), M(t))$ as the state of the system at time t with the state space $\mathbf{E} \in (0,...,k) \times (0,...,\infty)$. To model the customer impatience, we use the parameters introduced by Barrer[11] and Movaghar[12] as follows: For $t, \varepsilon \in R^+$ and $n \in N$, let

$\psi_n(t,\varepsilon) \equiv$ the probability that a customer misses its deadline during $[t, t+\varepsilon)$, given there are n customers in the system at time t .

Define $$\gamma_n(t) = \lim_{\varepsilon \to 0} \frac{\psi_n(t,\varepsilon)}{\varepsilon}. \qquad (1)$$

Note that because of the independence of all random variables and the memoryless property of exponential distributions, the customers will miss their deadlines at rate $\gamma_n(t)$ and for $X(t)$, the future state is only dependent on the present state. Thus, $X(t)$ is a two dimensional Markov process with state space **E**.

Throughout the paper, we assume that the statistical equilibrium or the steady-state of the system has been reached (the stability condition will be given in Section 3.). Therefore, $X = \lim_{t\to\infty} X(t)$ is the steady- state of the system and $\gamma_i = \lim_{t\to\infty} \gamma_i(t)$. We also define

$U_i \equiv$ the time an arriving customer with an infinite (or no) deadline must wait before its service commences in the long run, given it finds i customers in the system.

From Lemma 3.1 and the equations (3.19), (3.16), (3.15) in [12], we have

$$P(U_i \le \theta) = \frac{s\mu}{s\mu + \gamma_{i+1}} \qquad (2)$$

$$\gamma_i = \begin{cases} 0 & i \le s \\ (i-s)\dfrac{\Phi_{i-s-1}(s\mu)}{\Phi_{i-s}(s\mu)} - s\mu & i > s \end{cases} \qquad (3)$$

where $$\Phi_i(\zeta) = \int_0^\infty g_i(\tau) e^{-\zeta\tau} d\tau \qquad (4)$$

$$g_i(\tau) = \left[\int_0^v (1 - F_\theta(x)) dx \right]^i \qquad (5)$$

The details of obtaining (2) and (3) can be found in [12]. Based on the model description and the definition of parameter γ_i , the state transition diagram is shown in Figure 2. Clearly, this is an infinite quasi-birth and death (QBD) process. Due to the difficulty of solving such an infinite QBD process, we utilize an approximation method similar to the one in [8]. By letting v_{ij} be the retrial rate for state (i, j) , we construct a new Markov process X^A from X in the following way: X^A has the retrial rates in the states with $j > r$ $(0 \le r < \infty)$ as

$$v_{ij} = \begin{cases} j\lambda_1 & \begin{cases} i = k & 0 \le j \le \infty \\ 0 \le i \le k-1 & 0 \le j \le r \end{cases} \\ \infty & 0 \le i \le k-1 \quad j > r \end{cases} \cdot \qquad (6)$$

Based on (6), we can transform the state transition network of X process in Figure 2 to that of X^A process in Figure 3 by linking the dotted line in Figure 2. Then, for $j \ge r+1$, the behavior of X^A is like a $M/M/1$ queue. Obviously, X^A has the state space $\mathbf{E}_1 \in (0,...,k) \times (0,...,r) \cup k \times (r+1,...,\infty)$. Since X^A is transformed from X by changing some v_{ij} 's of X into ∞ , the traffic load of X^A must be greater than that of X . It follows that if under certain condition X^A reaches the steady-state (i.e. has the stationary distribution), X must also reach the steady state.

From Figure 3, using the lexicographical sequence for the states, the infinitesimal generator of X^A can be written as

$$\mathbf{Q} = \begin{pmatrix} \mathbf{A}_0 & \mathbf{C} & & & & \\ \mathbf{B}_1 & \mathbf{A}_1 & \mathbf{C} & & & \\ & \mathbf{B}_2 & \mathbf{A}_2 & \mathbf{C} & & \\ & & \ddots & \ddots & \ddots & \\ & & & \mathbf{B}_{k-1} & \mathbf{A}_{k-1} & \tilde{\mathbf{C}} \\ & & & & \tilde{\mathbf{B}}_k & \tilde{\mathbf{A}}_k \end{pmatrix}. \qquad (7)$$

where $\mathbf{A}_0,...,\mathbf{A}_{k-1}$, $\mathbf{B}_1,...,\mathbf{B}_{k-1}$, $\mathbf{C}$ are the $(r+1)\times(r+1)$ matrices, $\tilde{\mathbf{A}}_k$ is the $\infty \times \infty$ matrix, $\tilde{\mathbf{B}}_k$ is the $\infty \times k$ matrix, and $\tilde{\mathbf{C}}$ is the $k \times \infty$ matrix. Specifically, these matrices are as follows:

$$\mathbf{A}_i = \begin{pmatrix} a_{i0} & & & & \\ & \ddots & & & \\ & & a_{ij} & & \\ & & & \ddots & \\ & & & & a_{ir} \end{pmatrix}$$

$$a_{ij} = \begin{cases} -(\lambda + j\lambda_1 + i\mu) & 0 \le i \le s, 0 \le j < r \\ -(\lambda + j\lambda_1 + s\mu + \gamma_i) & s < i < k, 0 \le j < r \end{cases} \qquad (8)$$

$$a_{ir} = \begin{cases} -(\lambda + r\lambda_1 + i\bar{\beta}\mu) & 0 \le i \le s \\ -(\lambda + r\lambda_1 + s\bar{\beta}\mu + \gamma_i) & s < i < k \end{cases}$$

$$\tilde{\mathbf{A}}_k = \begin{pmatrix} a_{k0} & \alpha\lambda & & & & & & & \\ d_1 & a_{k1} & \alpha\lambda & & & & & & \\ & \ddots & \ddots & \ddots & & & & & \\ & & d_j & a_{kj} & \alpha\lambda & & & & \\ & & & \ddots & \ddots & \ddots & & & \\ & & & & d_{r-1} & a_{k(r-1)} & \alpha\lambda & & \\ & & & & & d_r & a_{kr} & q_c & \\ & & & & & & d_{r+1} & a_{k(r+1)} & q_c \\ & & & & & & & \ddots & \ddots & \ddots \end{pmatrix}$$

$$a_{kj} = -(s\mu + \alpha\lambda + j\bar{\alpha}\lambda_1 + \gamma_k) \quad 0 \le j \le \infty$$

$$d_j = \begin{cases} j\bar{\alpha}\lambda_1 & 1 \le j \le r \\ j\bar{\alpha}\lambda_1 + s\bar{\beta}\mu + \gamma_k & j > r \end{cases} \qquad (9)$$

$$q_c = \alpha\lambda + s\beta\mu$$

$$\mathbf{B}_i = \begin{pmatrix} b_i & e_i & & \\ & \ddots & \ddots & \\ & & b_i & e_i \\ & & & b_i \end{pmatrix}$$

$$b_i = \begin{cases} i\bar{\beta}\mu & 1 \le i \le s \\ s\bar{\beta}\mu + \gamma_i & s < i \le k \end{cases} \quad e_i = \begin{cases} i\beta\mu & 1 \le i \le s \\ s\beta\mu & s < i \le k \end{cases} \qquad (10)$$

$$\tilde{\mathbf{B}}_k = \begin{pmatrix} b_i & e_i & & \\ & \ddots & \ddots & \\ & & b_i & e_i \\ & & 0 & b_i \\ & & 0 & 0 \\ & & \vdots & \vdots \end{pmatrix} \quad \begin{matrix} b_k = s\bar{\beta}\mu + \gamma_k \\ e_k = s\beta\mu \end{matrix} \qquad (11)$$

$$\mathbf{C} = \begin{pmatrix} \lambda & & & & \\ c_1 & \lambda & & & \\ & \ddots & \ddots & & \\ & & c_j & \lambda & \\ & & & \ddots & \ddots \\ & & & & c_r & \lambda \end{pmatrix} \quad c_j = j\lambda_1 \qquad (12)$$

$$\tilde{\mathbf{C}} = \begin{pmatrix} \lambda & & & & & & \\ c_1 & \lambda & & & & & \\ & \ddots & \ddots & & & & \\ & & c_j & \lambda & & & \\ & & & \ddots & \ddots & & \\ & & & & c_r & \lambda & 0 & \cdots \end{pmatrix} \quad c_j = j\lambda_1 \qquad (13)$$

III. Computation of Stationary Performance Measures

We first prove the existence of the stationary distribution.

Lemma 3.1 If $\rho = \lambda / s\mu$ is finite, X^A has the stationary distribution.

Proof. It follows from Figure 3 that X^A is an irreducible and aperiodic Markov chain. Note that when the number of customers in the retrial area is greater than r, X^A behaves like a $M/M/1$ queue with a variable service rate. From Theorem 2 in [13], we know that if

$$\sum_{j=r+1}^{\infty} \prod_{m=r+1}^{j} \left(\frac{s\beta\mu + \alpha\lambda}{m\bar{\alpha}\lambda_1 + s\bar{\beta}\mu + \gamma_k(t)} \right) < \infty, \qquad (14)$$

then the birth and death process has the stationary distribution. Let $\pi_{kj}(t) \equiv$ the probability that the system is in state (k, j) at time t and its limiting distribution $\pi_{kj} = \underset{t\to\infty}{\pi_{kj}(t)}$. Let $\rho_m = \dfrac{s\beta\mu + \alpha\lambda}{m\bar{\alpha}\lambda_1 + s\bar{\beta}\mu + \gamma_k} (r < m \le \infty)$. Thus, under the stability condition, we have $\pi_{kj} = \pi_{k(j-1)} \cdot \prod_{m=r+1}^{j} \rho_m$ (15)

Because $\gamma_k(t) \ge 0$, we get

$$\sum_{j=r+1}^{\infty} \prod_{m=r+1}^{j} \left(\frac{s\beta\mu + \alpha\lambda}{m\bar{\alpha}\lambda_1 + s\bar{\beta}\mu + \gamma_k(t)} \right) \le \sum_{j=r+1}^{\infty} \prod_{m=r+1}^{j} \left(\frac{s\beta\mu + \alpha\lambda}{m\bar{\alpha}\lambda_1 + s\bar{\beta}\mu} \right). \qquad (16)$$

Let $\rho'_m = \dfrac{s\beta\mu + \alpha\lambda}{m\bar{\alpha}\lambda_1 + s\bar{\beta}\mu}$. Since ρ'_m decreases in m, there must be a positive integer M such that $\rho'_M < 1$. Note that the condition $\rho = \lambda / s\mu$ is finite ensures that ρ'_m is finite.

Clearly, we have $1 > \rho'_M > \rho'_{M+1} > \rho'_{M+2} \ldots$.Based on（16）, we have

$$\sum_{j=r+1}^{\infty} \prod_{m=r+1}^{j} \left(\frac{s\beta\mu + \alpha\lambda}{m\bar{\alpha}\lambda_1 + s\bar{\beta}\mu} \right) = \sum_{j=r+1}^{\infty} \prod_{m=r+1}^{j} \rho'_m < \sum_{j=r+1}^{M} (\rho'_{r+1})^{j-r} + \sum_{j=M+1}^{\infty} \prod_{m=r+1}^{j} \rho'_m < \sum_{j=r+1}^{M} (\rho'_{r+1})^{j-r} + (\rho'_{r+1})^{M-r} \sum_{j=M+1}^{\infty} \prod_{m=M+1}^{j} \rho'_m$$

$$< \sum_{j=r+1}^{M} (\rho'_{r+1})^{j-r} + (\rho'_{r+1})^{M-r} \sum_{j=M+1}^{\infty} (\rho'_M)^{j-M} = \sum_{j=r+1}^{M} (\rho'_{r+1})^{j-r} + (\rho'_{r+1})^{M-r} \cdot \frac{\rho'_M}{1 - \rho'_M} < \infty$$

.

Thus, if $\rho = {}^{\lambda}\!/\!_{s\mu}$ is finite, the birth and death process is ergodic, and X^A has the stationary distribution. That completes the proof.

Under the stability condition, we define the stationary probability distribution of X^A as $\boldsymbol{\pi}_i (\boldsymbol{\pi}_i = (\pi_{i0},\ldots,\pi_{ir}), 0 \le i \le k), \pi_{k(r+1)}, \pi_{k(r+2)}, \ldots$

Now, we present the main result.

Theorem 3.1 If $\rho = {}^{\lambda}\!/\!_{s\mu}$ is finite, the stationary distribution X^A is given by

$$\boldsymbol{\pi}_0 = \mathbf{T}_0 \times \frac{1}{W},\ \boldsymbol{\pi}_i = \boldsymbol{\pi}_0 \mathbf{R}_i (1 \le i \le k),$$

$$\pi_{kj} = \{\boldsymbol{\pi}_0 \mathbf{R}_k \mathbf{e}_1\} \times \prod_{m=r+1}^{j} \frac{\alpha\lambda + s\beta\mu}{m\overline{\alpha}\lambda_1 + s\overline{\beta}\mu + \gamma_k} (j \ge r+1),$$

where $\mathbf{R}_0 = \mathbf{I}, \mathbf{I}$ is the identity matrix,

$$\mathbf{R}_1 = -\mathbf{A}_0 \mathbf{B}_1^{-1}, \mathbf{R}_2 = -(\mathbf{R}_0 \mathbf{C} + \mathbf{R}_1 \mathbf{A}_1)\mathbf{B}_2^{-1},$$

and $\mathbf{R}_i = -(\mathbf{R}_{i-2}\mathbf{C} + \mathbf{R}_{i-1}\mathbf{A}_{i-1})\mathbf{B}_i^{-1} (3 \le i \le k)$. $\mathbf{T}_0$ is a basic solution of equation $\mathbf{T}_0 (\mathbf{R}_k \mathbf{A}_k + \mathbf{R}_{k-1}\mathbf{C}) = \mathbf{0}$,

$$W = \mathbf{T}_0 \left\{ \sum_{i=0}^{k} \mathbf{R}_i \right\} \mathbf{e} + \mathbf{T}_0 \mathbf{R}_k \mathbf{e}_1 \cdot \sum_{j=r+1}^{\infty} \left\{ \prod_{m=r+1}^{j} \frac{\alpha\lambda + s\beta\mu}{m\overline{\alpha}\lambda_1 + s\overline{\beta}\mu + \gamma_k} \right\},$$

where $\mathbf{e}$ is a $(r+1)$ dimension column vector with all elements 1, and $\mathbf{e}_1$ is a $(r+1)$ dimension column vector $(0,\cdots,0,1)^T$. $\mathbf{B}_k$ is defined in（10）and

$$\mathbf{A}_k = \begin{pmatrix} a_{k0} & \alpha\lambda & & & & & \\ d_1 & a_{k1} & \alpha\lambda & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & d_j & a_{kj} & \alpha\lambda & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & d_{k-1} & a_{k-1} & \alpha\lambda \\ & & & & & d_r & a_{kr} \end{pmatrix} \qquad （17）$$

$$a_{kj} = \begin{cases} -(s\mu + \alpha\lambda + j\overline{\alpha}\lambda_1 + \gamma_k) & 0 \le j < r \\ -(s\overline{\beta}\mu + r\overline{\alpha}\lambda_1 + \gamma_k) & j = r \end{cases}$$

$$d_j = j\overline{\alpha}\lambda_1 \quad 1 \le j \le r .$$

Proof. Because X^A satisfies the stability condition, X^A has the stationary distribution. Thus the stationary distribution satisfies

$$\begin{cases} (\boldsymbol{\pi}_0, \ldots, \boldsymbol{\pi}_k, \pi_{k(r+1)}, \pi_{k(r+2)}, \cdots)\mathbf{Q} = \mathbf{0} \\ \displaystyle\sum_{i=0}^{k-1}\sum_{j=0}^{r} \pi_{ij} + \sum_{j=0}^{\infty} \pi_{kj} = 1 \end{cases} \qquad （18）$$

For the $(k+1)(r+1)$ columns of the matrix (7), we have

$$\pi_{(k-1)r}\lambda + \pi_{k(r-1)}\alpha\lambda - \pi_{kr}(s\mu + \alpha\lambda + r\overline{\alpha}\lambda_1 + \gamma_k) + \pi_{k(r+1)}((r+1)\overline{\alpha}\lambda_1 + s\overline{\beta}\mu + \gamma_k) = 0 \cdot$$

(19)

From（15）, we obtain

$\pi_{k(r+1)} = \pi_{kr} \dfrac{\alpha\lambda + s\beta\mu}{(r+1)\overline{\alpha}\lambda_1 + s\overline{\beta}\mu + \gamma_k}$. Substituting $\pi_{k(r+1)}$ into（19）, we have $\pi_{(k-1)r}\lambda + \pi_{k(r-1)}\alpha\lambda - \pi_{kr}(s\overline{\beta}\mu + r\overline{\alpha}\lambda_1) = 0$. Then, we can replace $\widetilde{\mathbf{A}}_k, \widetilde{\mathbf{B}}_k, \widetilde{\mathbf{C}}$ with $\mathbf{A}_k, \mathbf{B}_k, \mathbf{C}$ to obtain $\boldsymbol{\pi}_0, \cdots, \boldsymbol{\pi}_k$.

Let

$$\mathbf{Q}^* = \begin{pmatrix} \mathbf{A}_0 & \mathbf{C} & & & & \\ \mathbf{B}_1 & \mathbf{A}_1 & \mathbf{C} & & & \\ & \mathbf{B}_2 & \mathbf{A}_2 & \mathbf{C} & & \\ & & \ddots & \ddots & \ddots & \\ & & & \mathbf{B}_{k-1} & \mathbf{A}_{k-1} & \mathbf{C} \\ & & & & \mathbf{B}_k & \mathbf{A}_k \end{pmatrix} \qquad （20）$$

Then, $(\boldsymbol{\pi}_0, \cdots, \boldsymbol{\pi}_k)\mathbf{Q}^* = \mathbf{0}$ （21）

From (21), we have

$$\boldsymbol{\pi}_0 \mathbf{A}_{.0} + \boldsymbol{\pi}_1 \mathbf{B}_1 = \mathbf{0}, \qquad （22）$$

$$\boldsymbol{\pi}_0 \mathbf{C} + \boldsymbol{\pi}_1 \mathbf{A}_1 + \boldsymbol{\pi}_2 \mathbf{B}_2 = \mathbf{0}, \qquad (23)$$

$$\vdots$$

$$\boldsymbol{\pi}_i \mathbf{C} + \boldsymbol{\pi}_{i+1} \mathbf{A}_{i+1} + \boldsymbol{\pi}_{i+2} \mathbf{B}_{i+2} = \mathbf{0}, \qquad （24）$$

$$\vdots$$

$$\boldsymbol{\pi}_{k-2} \mathbf{C} + \boldsymbol{\pi}_{k-1} \mathbf{A}_{k-1} + \boldsymbol{\pi}_k \mathbf{B}_k = \mathbf{0}, \qquad （25）$$

$$\boldsymbol{\pi}_{k-1} \mathbf{C} + \boldsymbol{\pi}_k \mathbf{A}_k = \mathbf{0}. \qquad （26）$$

Because $\mathbf{B}_i (1 \le i \le k)$ is invertible, we get

$$\boldsymbol{\pi}_1 = -\boldsymbol{\pi}_0 \mathbf{A}_0 \mathbf{B}_1^{-1} = \boldsymbol{\pi}_0 \mathbf{R}_1. \qquad （27）$$

Substituting（27）into（23), we obtain

$$\boldsymbol{\pi}_2 = -\boldsymbol{\pi}_0 (\mathbf{R}_1 \mathbf{A}_1 + \mathbf{C})\mathbf{B}_2^{-1} = \boldsymbol{\pi}_0 \mathbf{R}_2 . \qquad （28）$$

Repeating this substitution, we have $\boldsymbol{\pi}_i = \boldsymbol{\pi}_0 \mathbf{R}_i$, where $\mathbf{R}_i = -(\mathbf{R}_{i-2}\mathbf{A}_{i-2} + \mathbf{R}_{i-1}\mathbf{C})\mathbf{B}_i^{-1} (3 \le i \le k)$.

Because the order of $\mathbf{Q}^*$ is $(k+1)(r+1)-1$, the equation system matrix (21) has only one basic solution. Letting $(\mathbf{T}_0, \cdots, \mathbf{T}_k)$ be the basic solution of（21）, we have $\mathbf{T}_{k-1}\mathbf{C} + \mathbf{T}_k \mathbf{A}_k = \mathbf{0}$, and $\mathbf{T}_0 (\mathbf{R}_k \mathbf{A}_k + \mathbf{R}_{k-1}\mathbf{C}) = \mathbf{0}$, from which $\mathbf{T}_0$ is obtained. According to Lemma 3.1, we can deduce $W < \infty$. Using

$$W = \mathbf{T}_0 \left\{ \sum_{i=0}^{k} \mathbf{R}_i \right\} \mathbf{e} + \mathbf{T}_0 \mathbf{R}_k \mathbf{e}_1 \cdot \sum_{j=r+1}^{\infty} \left\{ \prod_{m=r+1}^{j} \frac{\alpha\lambda + s\beta\mu}{m\overline{\alpha}\lambda_1 + s\overline{\beta}\mu + \gamma_k} \right\}$$

, $\boldsymbol{\pi}_0 = \mathbf{T}_0 \times \dfrac{1}{W}$ and (15), we obtain $\displaystyle\sum_{i=0}^{k-1}\sum_{j=0}^{r} \pi_{ij} + \sum_{j=0}^{\infty} \pi_{kj} = 1$ which satiates the normalization condition. This completes the proof.

Under the stationary condition, the probability of missing the deadline can be determined, and
it follows from (2) that

$$P_I=\sum_{i=s}^{k-1}\left\{P(U_i>\theta)\sum_{j=0}^{r}\boldsymbol{\pi}_{ij}\right\}=\sum_{i=s}^{k-1}\left\{\frac{\gamma_{i+1}}{\gamma_{i+1}+s\mu}\sum_{j=0}^{r}\boldsymbol{\pi}_{ij}\right\} \qquad (29)$$

According to the main theorem above, we can obtain some useful stationary performance measures of this queueing model as follows.

（i） Mean number of customers in waiting and service areas

$$E[N]=\sum_{i=1}^{k-1}i\left\{\sum_{j=0}^{r}\pi_{ij}\right\}+k\sum_{j=0}^{\infty}\pi_{kj} \qquad （30）$$

（ii） Mean number of busy servers

$$E[S_N]=\sum_{i=1}^{s}i\left\{\sum_{j=0}^{r}\pi_{ij}\right\}+\sum_{i=s+1}^{k-1}s\left\{\sum_{j=0}^{r}\pi_{ij}\right\}+s\sum_{j=0}^{\infty}\pi_{kj} \qquad (31）$$

（iii） Mean number of customers in retrial area

$$E[M]=\sum_{j=1}^{r}\left\{j\cdot\sum_{i=0}^{k-1}\pi_{ij}\right\}+\sum_{j=1}^{\infty}j\pi_{kj} \qquad (32）$$

（iv） Probability of blocking $P_B=\sum_{j=0}^{\infty}\pi_{kj}$ （33）

（v） Probability of losing customers $P_L=\overline{\alpha}P_B+P_I$ （34）

IV. A Numerical Example and Concluding Remarks

Since the stationary distribution of X is not obtainable, we cannot compare X^A with X. However, as $r\longrightarrow\infty$, it is clear that $X^A\longrightarrow X$. Another approximation method to solving the performance measures of X is to use a state space truncation. That is to use a finite buffer for the retrial area and denote this finite state space process by X^F.Clearly, we also have $X^F\longrightarrow X$ as $r\longrightarrow\infty$. We can compare our method based on X^A with the state-truncation method based on X^F in terms of some stationary performance measures such as P_L and $E[M]$. The convergence rates of P_L and $E[M]$ are shown in Figure 4 and Figure 5, respectively. It is clear that there is an advantage of X^A over X in terms of the speed of the convergence. As β increases, this advantage is more significant. This implies that as the feedback probability is going up, our method becomes more attractive for numerical analysis. Note that in Figures 4 and 5, we assume that the customers' impatience time is deterministic, and equals 1 minute and 2 minutes, respectively. A numerical example is presented below to show the computation of several stationary performance measures.

Consider a system with the following parameters: $\lambda^{-1}=\dfrac{50}{27}$（minute），$\lambda_1^{-1}=3$（minute），$\alpha=0.8$，$\beta$=0.15，$\mu^{-1}=5$(minute), $s=3$，$k=8$，$r=5$，distribution of θ is $F_\theta(\tau)=\begin{cases}0 & \tau<\theta\\ 1 & \tau\geq\theta\end{cases}$. Using Theorem 3.1 and（29）~（34），we can obtain the stationary distribution, and the useful performance measures, as follows:

$\boldsymbol{\pi}_0$ =[0.04560653828708，0.01204231115849，0.00146863657661，⋯，0.00000016526970]

$\boldsymbol{\pi}_1$ =[0.14486782750014，0.03629951129458，0.00401864390357，⋯，0.00000030257655]

$\boldsymbol{\pi}_2$ =[0.23106046457524，0.05181149560296，0.00483759547587，⋯，0.00000023391465]

$\boldsymbol{\pi}_3$ =[0.24876151520090，0.04177242352595，0.00311475585855，⋯，0.00000009618788]

$\boldsymbol{\pi}_4$ =[0.11416560738966，0.01618501339324，0.00104796846634，⋯，0.00000002460865]

$\boldsymbol{\pi}_5$ =[0.03060603998876，0.00388122401624，0.00022882865026，⋯，0.00000000561211]

$\boldsymbol{\pi}_6$ =[0.00567674922969，0.00066989891957，0.00003787383455，⋯，0.00000000213890]

$\boldsymbol{\pi}_7$ =[0.00079371056804，0.00009526377084，0.00000616650578，⋯，0.00000000157569]

$\boldsymbol{\pi}_8$ =[0.00008330156668，0.00001637355955，0.00000190013823，⋯，0.00000000144567]

$\pi_{8,12}$ = 0.13929608080233*e-009 ，

$\pi_{8,13}$ =0.01325860883835*e-009，⋯

$E[N]$= 2.38757286440690；$E[S_N]$=2.16343283759263；$E[M]$=0.19485737706694；

P_B =1.017734284311805*e-004；P_I = 0.29155425430134；P_L = 0.29157460898702 。

We also present some performance measures of two cases in Figures 6-9. In these figures, Cases I is a system where the impatience time θ is deterministic namely $F_\theta(\tau)=\begin{cases}0 & \tau<\theta\\ 1 & \tau\geq\theta\end{cases}$, and $\overline{\theta}=\theta$；Case Ⅱ a system where θ is exponential namely $F_\theta(\tau)=1-e^{-v\tau}$, and $\overline{\theta}=v^{-1}$.

From these figures, we find that the probability of losing customers for the deterministic impatience time (case I) is lower than the stochastic impatience time (case II). However, the probability of blocking is just the opposite. This means that the variation in the impatience time increases the probability of losing the customers but decreases the probability of blocking. In Figure 8, it is noted that $\gamma_i=\lim_{t\to\infty}\gamma_i(t)$ in Case II is higher than in Case I for the entire *i* value range. Figure 3 shows that the feedback probability β affects all *i* levels while the retrial probability α affects only *i*=*k* level. Therefore, the performance measures of the system are more influenced by the feedback than by the retrial.

In this paper, we have formulated a QBD process for a multi-server queueing system with the major features of a call center and have developed a computational procedure for the stationary performance measures based on an approximate but solvable equivalent system to the original system.. The procedure provides practitioners or call center managers with a quantitative performance evaluation tool in their system design and workforce scheduling. A direction for future research is to conduct an empirical study on the distribution of the impatience time and the estimation of the feedback and retrial probabilities for a practical call center.

Acknowledgements

This work was supported by Professor Zhe George Zhang's CBE Summer Research Grant and Professor Y. Zhu's NSF of China # 70571030 and the Education Department of Jiangsu Province through fund 00KJT11003.

REFERENCES

[1] Ger Koole, Queueing models of call centers: an introduction ,Annals of Operations Research 113, 41 – 59, 2002.

[2] A. Krishnamoorthy and P. V. Ushakumari,GI/M/1/1 queue with finite retrials and finite orbits stochastic analysis and applications, 20(2), 357 – 374 (2002).

[3] B.D. Choi, Y.C. Kim and Y.W. Lee, The *M/M/c* retrial queue with geometric loss and feedback, Computers and Mathematics with Applications 36 (1998) 41–52.

[4] A. Krishnamoorthy and P. V. Ushakumari,GI/M/1/1 queue with finite retrials and finite orbits stochastic analysis and applications, 20(2),357-374(2002).

[5] J.R.Artalejo, A.Gomez-Corral, M.F.Neuts ,Analysis of multiserver queues with constant retrial rate ,European Journal of Operation Research 135(2001)569-581.

[6] S.N. Stepanov,Markov models with retrials: the calculation of stationary performance measures based on the concept of truncation, Mathematical and Computer Modelling 30 (1999) 207–228.

[7] R.I. Wilkinson, Theories for toll traffic engineering in the U.S.A., The Bell System Technical Journal 35 (1956) 421–514.

[8] G.I. Falin, Calculation of probability characteristics of a multiline system with repeat calls, Moscow University Computational Mathematics and Cybernetics 1 (1983) 43–49.

[9] M.F. Neuts and B.M. Rao, Numerical investigation of a multiserver retrial model, Queueing Systems 7 (1990) 169–190.

[10] J.R.Artalejo, M.Pozo, Numerical calculation of the stationary distribution of the main multiserver retrial queue, Annals of Operations Research 116,41-56,2002.

[11] D.Y. Barrer, Queuing with impatient customers and ordered service, Oper. Res. 5 (1957) 650–656.

[12] A. Movaghar, On queueing with customer impatience until the beginning of service, Queueing Systems, 29 (1998) 337–350.

[13] S. Karlin and J. McGregor, The classification of birth and death process, Tran. Amer. Math. Soc. 86(1957), 366-400.

[14] N. Gans, G. Koole, and A. Mandelbaum, Telephone Call Centers: Tutorial, Review, and Research Prospects, Manufacturing and Service Operations Management, 5, (2), (2003), 79-141.

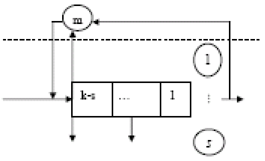


Figure 1: The system.

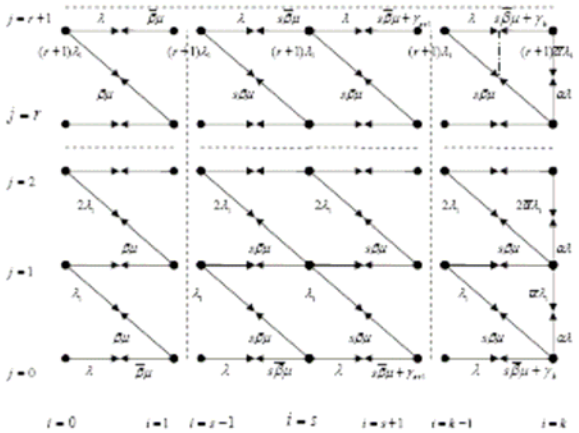


Figure 2. State transition of X

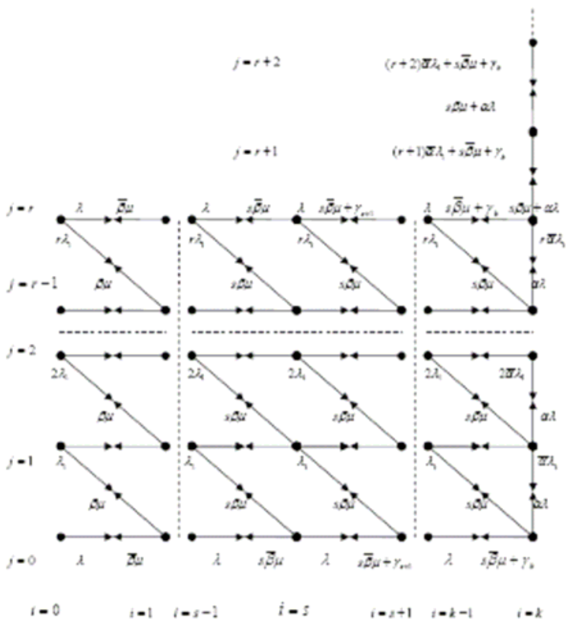
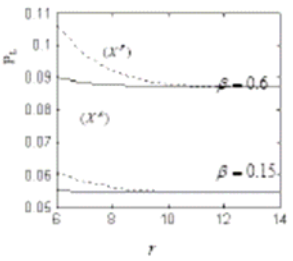
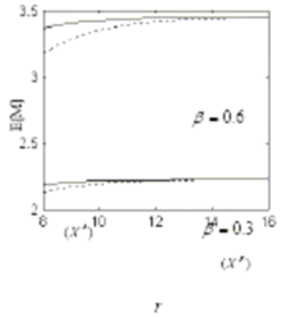


Figure 3. State transition of X^A.



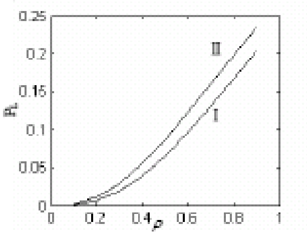
$(\lambda = 1.8, \lambda_1 = 1/3, \alpha = 0.8, \mu = 0.2, s = 3, k = 8, \bar{\sigma} = 1)$

Figure 4. Comparing X^A with X^B in terms of P_L in Case I.



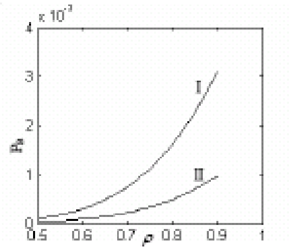
$(\lambda = 1.8, \lambda_1 = 1/3, \alpha = 0.8, \mu = 0.2, s = 3, k = 8, \bar{\sigma} = 2)$

Figure 5. Comparing X^A with X^B in terms of E[M] of Case I.



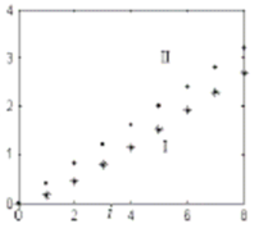
$(\lambda = \rho \cdot s\mu, \lambda_1 = 1/3, \alpha = 0.8, \mu = 0.2, s = 3, k = 8, r = 10, \bar{\sigma} = 2.5, \beta = 0.05)$

Figure 6. Probability of losing customers v.s. traffic load.



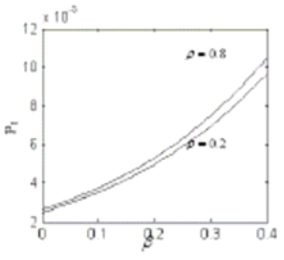
$(\lambda = \rho \cdot s\mu, \lambda_1 = 1/3, \alpha = 0.8, \mu = 0.2, s = 3, k = 8, r = 10, \bar{\sigma} = 2.5, \beta = 0.05)$

Figure7. Probability of blocking v.s. traffic load.



$(\mu = 0.2, s = 3, \bar{\sigma} = 2.5)$

Figure 8. γ_i v.s. the number of waiting and in-service customers.



$(\lambda = 1, \lambda_1 = 1/3, \mu = 0.2, s = 3, k = 8, r = 10, \bar{\sigma} = 2.5)$

Figure 9. P_I v.s. probability of feedback in Case I for two different traffic loads.